Differentially private inference via noisy optimization

Marco Avella Medina* Casey Bradshaw* Po-Ling Loh†

November 18, 2021

Abstract

We propose a general optimization-based framework for computing differentially private M-estimators and a new method for constructing differentially private confidence regions. Firstly, we show that robust statistics can be used in conjunction with noisy gradient descent or noisy Newton methods in order to obtain optimal private estimators with global linear or quadratic convergence, respectively. We establish local and global convergence guarantees, under both local strong convexity and self-concordance, showing that our private estimators converge with high probability to a small neighborhood of the non-private M-estimators. Secondly, we tackle the problem of parametric inference by constructing differentially private estimators of the asymptotic variance of our private M-estimators. This naturally leads to approximate pivotal statistics for constructing confidence regions and conducting hypothesis testing. We demonstrate the effectiveness of a bias correction that leads to enhanced small-sample empirical performance in simulations. We illustrate the benefits of our methods in several numerical examples.

1 Introduction

Over the last decade, differential privacy has evolved from a rigorous paradigm derived by theoretical computer scientists for releasing sensitive data to a technology deployed at scale in numerous applications [Ding et al., 2017, Erlingsson et al., 2014, Garfinkel et al., 2019, Tang et al., 2017]. The setting assumes the existence of a trusted curator who holds the data of individuals in a database, and the goal of privacy is to simultaneously protect individual data while allowing statistical analysis of the aggregate database. Such protection is guaranteed by differential privacy in the context of a remote access query system, where a statistician can only indirectly access the data, e.g., by obtaining noisy summary statistics or outputs of a model. Injecting noise before releasing information to the statistician is essential for preserving privacy, and the noise should be as small as possible in order to optimize statistical performance of the released statistics.

In this paper, we consider the problem of estimation and inference for M-estimators. Inspired by the work of Bassily et al. [2014], Lee and Kifer [2018], Song et al. [2013], and Feldman et al. [2020], among others, we propose noisy optimization procedures that output differentially private counterparts of standard M-estimators. The central idea of these methods is to add noise to every iterate of a gradient-based optimization routine in a way that causes each iterate to satisfy a targeted differential privacy guarantee. Even though this idea is now fairly common in the literature, our proposed methodology is novel in the following respects:

(a) Noisy gradient descent: While various versions of such algorithms have appeared in the literature, our first contribution is to provide a complete, global, finite-sample convergence analysis

*Department of Statistics, Columbia University
†Statistical Laboratory, University of Cambridge

of this algorithm under local strong convexity. We demonstrate that the resulting algorithm converges linearly to a near-optimal neighborhood of the target population parameter, which in turn shows that the resulting estimators are nearly minimax optimal and asymptotically normally distributed as their non-private counterparts. We point out some flaws in common implementations relying on clipped gradients and and show how these problems naturally disappear when one considers appropriate M-estimators from the robust statistics literature.

(b) Noisy Newton algorithm: A second main contribution of our work is to introduce a differentially private counterpart of Newton's method. Each step of this algorithm adds noise to both the gradient and the Hessian of the loss function at the current iterate, as both quantities are potential sources of privacy leakage. Our theory shows that under either local strong convexity or self-concordance, the noisy Newton algorithm converges quadratically to an optimal neighborhood of the the target population parameter. Similar to the classical convergence analysis of Newton algorithms, the convergence analysis of our algorithm has two phases—a first phase consisting of "damped Newton" steps guaranteeing improved objective values with stepsize $\eta < 1$ when the iterates are far from the solution, and a "pure Newton" phase with stepsize $\eta = 1$ when the iterates are close enough to the solution. In contrast to standard non-private algorithms, we do not rely on line backtracking, as this step is not easily implemented under privacy constraints. Instead, we take damped noisy Newton steps with a fixed stepsize until a verifiable condition is met indicating that the algorithm can safely enter the pure Newton stage. As evidenced by our simulations, our noisy Newton algorithm can lead to particularly significant improvements over noisy gradient descent when the algorithms are initialized sufficiently far from the global optimum.

(c) Noisy confidence regions: Our final contribution is to introduce a new approach for constructing confidence regions based on a noisy sandwich formula for M-estimators. The idea is fairly simple—since our noisy optimizers are asymptotically normally distributed with a known formula for the asymptotic variance, we employ differentially private estimators of the asymptotic variance. We rely on matrix-valued privacy-inducing noise in order to make the two main components of the sandwich formula differentially private. We further propose a bias correction method that significantly improves the coverage probability of our confidence regions for small sample sizes. In short, we provide a general technique for constructing asymptotically valid confidence regions, which provides substantial numerical improvements over the few existing alternatives in regression settings [Avella Medina, 2021, Barrientos et al., 2019, Sheffet, 2017, Wang et al., 2019].

1.1 Related literature

A sizable body of work is devoted to developing differentially private approaches for convex optimization in the context of empirical risk minimization. A first general optimization construction explored in the literature consists of perturbing the objective function so as to ensure that the resulting minimizer is differentially private. Representative work exploring this idea includes Chaudhuri and Monteleoni [2008], Chaudhuri et al. [2011], Kifer et al. [2012], Jain and Thakurta [2014] and more recently Slavkovic and Molinari [2021].

A related idea that is closer to the methods studied in our work considers iterative gradient-based algorithms, where the algorithm employs noisy gradients designed to ensure differential privacy at each iteration. Many such empirical risk minimization algorithms have been proposed [Balle et al., 2020, Bassily et al., 2014, 2019, Iyengar et al., 2019, Lee and Kifer, 2018, Song et al., 2013, Wang et al., 2017a], often motivated by settings such as online problems [Duchi et al., 2018, Jain

et al., 2012], multiparty classification [Rajkumar and Agarwal, 2012], Bayesian learning [Wang et al., 2015], high-dimensional regression [Cai et al., 2019, 2020, Talwar et al., 2015], and deep learning [Abadi et al., 2016, Bu et al., 2020]. While most of the literature has focused on noisy stochastic gradient descent algorithms, we utilize full gradient evaluations at every iteration, similar to standard implementations in non-private statistical software. We note that many existing results cannot be directly applied to our setting [Bassily et al., 2014, Feldman et al., 2020]. Moreover, even existing methods that use full gradient evaluations rely on different proof techniques [Cai et al., 2019, 2020], as we avoid truncation arguments by construction, allow unbounded input variables and parameter spaces, and can explicitly track the impact of potentially bad starting values. We can achieve all of the above by considering locally strongly convex or self-concordant objective functions defining Fisher-consistent bounded-influence M-estimators. We believe that all of these points are important, as they make our methods work well under standard assumptions for non-private settings.

On the optimization side, our work is related to literature on inexact oracle methods [Devolder et al., 2014, Sun et al., 2020] and stochastic optimization [d'Aspremont, 2008, Ghadimi and Lan, 2012, Wang et al., 2017b]. Indeed, our proofs for the convergence of noisy gradient descent and noisy Newton's method both rely on showing that with high probability, the noise introduced to the gradient and Hessian terms has a negligible effect on the convergence behavior of the iterates (at least up to the order of the statistical error of the non-noisy versions of the algorithms). The theory of inexact oracle methods similarly seeks to derive results for the output of iterative optimization algorithms when gradients and/or Hessians are only computed up to a certain level of accuracy. However, whereas inexact oracles necessarily rely on approximate gradients, approximate gradients alone do not constitute inexact oracles unless the domain is bounded [Devolder et al., 2014], so our results are not directly implied by existing literature in this area. The work on inexact second-order methods is more sparse: Sun et al. [2020] studied global and local convergence of an inexact oracle version of Newton's method, but their analysis only covers the class of standard self-concordant functions (denoted $(\gamma, 3)$-self-concordant in our paper), whereas our main focus is on pseudo-self-concordant (i.e., $(\gamma, 2)$-self-concordant) functions, which are appropriate for our M-estimation framework. The stochastic optimization literature is more directly applicable to the noisy gradient setting considered in our paper, and our privatized gradients can be viewed as a special instantiation of stochastic gradients, where noise is introduced not due to sampling error, but intentionally in order to preserve privacy. However, we note that our results are also not direct consequences of existing literature on first-order [Ghadimi and Lan, 2012] or second-order [Wang et al., 2017b] methods, since the objective functions we consider are at most locally strongly convex, and our noisy Newton algorithms employ a particular version of a noisy Hessian that is motivated by differential privacy. We further note that while our global convergence analysis of the noisy Newton method for self-concordant functions builds upon the work of Karimireddy et al. [2018], we show that the Hessian stability condition imposed in that paper, along with an additional uniform bound on the Hessian, are enough to ensure a much stronger guarantee of local quadratic convergence after an initial epoch of linear convergence.

The literature on hypothesis testing and confidence intervals with differential privacy is relatively limited, and has only recently received some attention in the computer science literature [Acharya et al., 2018, Awan and Slavković, 2018, Chadha et al., 2021, Covington et al., 2021, Gaboardi et al., 2016, Karwa and Vadhan, 2017, Rogers and Kifer, 2017, Uhler et al., 2013, Wang et al., 2019]. This problem has also been studied in the statistics literature in a regression setting, but in many cases, suffers from the same drawbacks encountered in estimation, e.g., the need to assume bounded data [Wang et al., 2019, Yu et al., 2014], resorting to truncation [Barrientos et al., 2019], and requiring very large sample sizes for the methods to work well [Sheffet, 2017], especially in

light of the expected $\sqrt{n}$-consistency of M-estimators [Avella Medina, 2021, Barber and Duchi, 2014, Cai et al., 2019]. Recent Bayesian inference methods include [Kulkarni et al., 2021, Peña and Barrientos, 2021, Savitsky et al., 2019]. We note that while Avella Medina [2021] provides a simple method to perform differentially private inference, the method requires a sufficiently large sample size that depends on problem parameters that are difficult to quantify in practice. Our method guarantees differential privacy for all sample sizes and gives theoretically near-optimal convergence results as well as good small-sample performance in simulated examples.

1.2 Outline

The rest of our paper is organized as follows: Section 2 introduces basic concepts of differential privacy and M-estimators. Section 3 presents our noisy gradient descent algorithm, with a complete global convergence analysis under local strong convexity, and a few examples. Section 4 introduces our noisy Newton algorithm and provides two parallel global convergence theories, one relying on local strong convexity and the other on self-concordance. We also discuss several examples that satisfy these conditions and provide some numerical illustrations. In Section 5, we present our new approach for constructing confidence regions, including a bias correction for small samples. Section 7 concludes the paper with a discussion of our results and interesting future research directions. All the proofs of our results are relegated to the Appendix.

2 Background

We begin by presenting some useful notation that will be used throughout this paper, followed by fundamentals of differential privacy and M-estimation.

2.1 Notation

For a vector $v \in \mathbb{R}^m$, we write $\|v\|_2^2 = v^\top v$. For a positive definite matrix A, we denote $\|v\|_A^2 = v^\top A v$ and $\|A\|_2 = \max_{v:\|v\|_2=1} \|Av\|_2$. We denote the smallest and largest eigenvalues of a symmetric matrix A by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively.

We write $f = O(g)$ when $f(\cdot) \le Cg(\cdot)$ for any admissible arguments of $f(\cdot)$ and $g(\cdot)$ and some positive constant C. Similarly, we write $f = \Omega(g)$ when $f(\cdot) \ge cg(\cdot)$ for any admissible arguments of $f(\cdot)$ and $g(\cdot)$ and some positive constant c. We write $f \asymp g$ when both $f = O(g)$ and $f = \Omega(g)$. For sequences of random variables $\{X_n\}$ and $\{Y_n\}$, we write $X_n = O_p(Y_n)$ to denote boundedness in probability, i.e., for every $\epsilon > 0$, there exist M and N such that $\mathbb{P}\left(\left|\frac{X_n}{Y_n}\right| < M\right) > 1-\epsilon$ for all $n \ge N$.

We write $x_{1:n} \in \mathbb{R}^{n\times m}$ to denote the data matrix with i^{th} row equal to $x_i \in \mathbb{R}^m$. For any two matrices $x_{1:n}, x'_{1:n} \in \mathbb{R}^{n\times m}$, we define their Hamming distance $\mathrm{d_H}(x_{1:n}, x'_{1:n}) := \left|\{i = 1, \dots, n : x_i \neq x'_i\}\right|$ to be the number of coordinates which differ between $x_{1:n}$ and $x'_{1:n}$.

For $\theta \in \mathbb{R}^p$ and $r > 0$, we write $\mathcal{B}_r(\theta)$ to denote the Euclidean ball of radius r around θ. For a set (event) E, let $\mathbb{1}_E$ or $\mathbb{1}\{E\}$ denote the indicator function.

2.2 Gaussian differential privacy

In what follows, we call a *random function* any function $h : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ which maps each $x_{1:n} \in \mathbb{R}^{n\times m}$ to a Borel-measurable random variable $h(x_{1:n})$. A statistic is typically a deterministic function of some observed data set $x_{1:n}$, e.g., the sample mean $\bar{x}$, whereas $h(x)$ would be a randomized estimate, i.e., the output of a randomized algorithm obtained by perturbing the deterministic output $\bar{x}$.

Definition 1 ((ϵ, δ)-differential privacy). *Let $\varepsilon, \delta > 0$. A random function h is called (ε, δ)-differentially private (DP) if and only if for every pair of data sets $x_{1:n}, x'_{1:n} \in \mathbb{R}^{n \times m}$ such that $\mathrm{d_H}(x_{1:n}, x'_{1:n}) = 1$ and for all Borel sets $B \subseteq \mathbb{R}^p$, we have*

$$\mathbb{P}[h(x_{1:n}) \in B] \leq e^{\varepsilon} \mathbb{P}[h(x'_{1:n}) \in B] + \delta.$$

It is important to note that the probabilities in the above definition are computed over the randomness of the function h, for any fixed neighboring data sets $x_{1:n}$ and $x'_{1:n}$. The most basic approach for constructing such a randomized function consists of adding random noise to a (deterministic) statistic, as we will discuss below.

The definition of (ε, δ)-DP limits the ability of a potential adversary to identify whether an individual is present or absent in a data set based on released randomized outputs, as it is difficult to distinguish between the probability distributions of $h(x_{1:n})$ and $h(x'_{1:n})$. This remark leads to a hypothesis testing interpretation of differential privacy pointed out by Wasserman and Zhou [2010]. Indeed, one can interpret differential privacy as a protection guarantee against an adversary that tests two simple hypotheses of the form

$$H_0 : x_i = s \quad \text{versus} \quad H_1 : x_i = t.$$

Privacy is ensured when this testing problem is difficult, and (ε, δ)-DP assesses the hardness of the problem via an approximate worst-case likelihood ratio of the distributions of $h(x_{1:n})$ and $h(x'_{1:n})$ over all neighboring data sets.

Dong et al. [2021] further built on this hypothesis testing interpretation of differential privacy and advocated for a new definition of Gaussian differential privacy that we will use throughout the rest of our paper. The definition involves a transparent interpretation of the privacy requirement: determining whether any individual is in a data set is at least as hard as distinguishing between two normal distributions $N(0, 1)$ and $N(\mu, 1)$ based on one random draw. The formal definition is a little more complex:

Definition 2 (Gaussian differential privacy). *Let $h : \mathbb{R}^{n \times m} \to \mathbb{R}^p$ be a randomized function.*

1. *We say that h is f-DP if any α-level test between simple hypotheses of the form $H_0 : x_i = t$ vs. $H_1 : x_i = s$ has power function $\beta(\alpha) \leq 1 - f(\alpha)$, where f is a convex, continuous, non-increasing function satisfying $f(\alpha) \leq 1 - \alpha$ for all $\alpha \in [0, 1]$.*

2. *We say that h is μ-Gaussian differentially private (GDP) if h is f-DP and*

$$f(\alpha) \geq \Phi(\Phi^{-1}(1 - \alpha) - \mu)$$

for all $\alpha \in [0, 1]$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

The following notion of sensitivity will be central in our construction of differentially private procedures. In particular, it is used in the most basic algorithms that make some output $h(x)$ private by simply releasing $h(x) + u$, where u is an independent noise term whose variance is scaled according to the sensitivity of h.

Definition 3 (Global sensitivity). *Let $g : \mathbb{R}^{n \times m} \to \mathbb{R}^p$ be a deterministic function. The global sensitivity of g is the (possibly infinite) number*

$$\mathrm{GS}_g = \sup_{x_{1:n}, x'_{1:n} \in \mathbb{R}^{n \times m}} \{\|g(x'_{1:n}) - g(x_{1:n})\|_2 : \mathrm{d_H}(x_{1:n}, x'_{1:n}) = 1\}.$$

The following theorem concerns a procedure that will be a primary building block for our private algorithm. This method is called the Gaussian mechanism, and can be easily tuned to achieve a desired (ε, δ)-DP [Dwork and Roth, 2014, Theorem A.1] or μ-GDP guarantee, as stated below:

Theorem 1 (Theorem 1 in Dong et al. [2021]). *Let $g : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ be a function with finite global sensitivity GS_g. Let Z be a standard normal p-dimensional random vector. For all $\mu > 0$ and $x \in \mathbb{R}^{n\times m}$, the random function $h(x) = g(x) + \frac{\mathrm{GS}_g}{\mu} Z$ is μ-GDP.*

Our proposed optimization methods will rely on compositions of a sequence of differentially private outputs computed using the same data set, where each step uses information from prior private computations. A question of paramount importance is to characterize the overall privacy guarantee of such an analysis. Intuitively, this guarantee should degrade as one composes more and more private outputs. Dong et al. [2021] very elegantly argued that in such scenarios, Gaussian differential privacy is very special: Its privacy guarantee under composition can be characterized as the K-fold composition of μ_k-GDP mechanisms, and is exactly μ-GDP, where $\mu = \sqrt{\mu_1^2 + \cdots + \mu_K^2}$. More fundamentally, GDP is a canonical privacy guarantee in an asymptotic sense, as a central limit theorem phenomenon shows that the composition of a large number of f-DP algorithms is approximately μ-GDP for some parameter μ [Dong et al., 2021, Corollary 2 and Theorem 6].

2.3 M-estimators for parametric models

M-estimators are a simple class of estimators, stemming from robust statistics and constituting a very general approach to parametric inference [Huber, 1964, Huber and Ronchetti, 2009]. They will be the main tool used in our private approach to estimation and inference.

Suppose we observe an i.i.d. sample $x_1, \ldots, x_n$ with common cumulative distribution F and wish to estimate a population parameter $\theta_0 = T(F)$ lying in a parameter space Θ. Our construction of differentially private estimators relies on noisy optimization techniques that will lead to private counterparts of M-estimators $\hat{\theta} = T(F_n)$ of θ_0, defined as minimizers of the form

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}_n(\theta) = \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, \theta) = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}_{F_n}[\rho(X, \theta)], \tag{1}$$

where F_n denotes the empirical distribution of $x_1, \ldots, x_n \in \mathbb{R}^m$. This class of estimators is a generalization of the class of maximum likelihood estimators.

When the loss function ρ in expression (1) is differentiable and convex, the estimator $\hat{\theta}$ can alternatively be viewed as the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(x_i, \hat{\theta}) = 0, \tag{2}$$

where $\Psi(x, \theta) = \frac{\partial}{\partial\theta}\rho(x, \theta)$. M-estimators defined by a function $\Psi(x, \theta)$ which is bounded in $x \in \mathcal{X} \subseteq \mathbb{R}^m$ are particularly appealing in robust statistics, since this guarantees that the influence function of the functional $T(F)$ is bounded and therefore ensures infinitesimal robustness to outliers [Hampel et al., 1986].

We also note that under mild general conditions [Huber, 1967], M-estimators are asymptotically normally distributed as

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V(\Psi, F)),$$

where

$$\begin{aligned} V(\Psi, F) &:= M(\Psi, F)^{-1} Q(\Psi, F) M(\Psi, F)^{-1}, \\ M(\Psi, F) &:= -\mathbb{E}_F[\dot{\Psi}(Z, \theta_0)], \\ Q(\Psi, F) &:= \mathbb{E}_F[\Psi(Z, \theta_0)\Psi(Z, \theta_0)^\top], \end{aligned} \tag{3}$$

and $\dot{\Psi}(x, \theta) := \frac{\partial}{\partial \theta^\top} \Psi(x, \theta)$. We will establish analogous results for our noisy estimators in the following sections.

3 Randomized M-estimators via noisy gradient descent

The theory of empirical risk minimization in convex optimization suggests a variety of algorithms which can be used to compute the global optimum (1) over a convex set Θ [Boyd and Vandenberghe, 2004]. One of the most elementary methods is gradient descent, which approximates $\hat{\theta}$ via the iterates

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\eta}{n} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)}). \tag{4}$$

In a classical statistical setting, one would for example consider the iterates (4) until a numerical condition is met, e.g., $\| \frac{1}{n} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)}) \|_2 < \epsilon$, for some prespecified tolerance level ϵ. Since the optimization error can be made negligible by setting ϵ arbitrarily small, classical statistical theory usually ignores the effect of carrying out inference on the final iterate $\theta^{(K)}$ as opposed to the global optimum $\hat{\theta}$, which is typically analyzed theoretically. Indeed, one can in principle take K as large as needed in order to ensure that $\theta^{(K)}$ is numerically identical to $\hat{\theta}$. This is in stark contrast to our proposed differentially private version of gradient descent, since the number of iterations K needs to be set before running the algorithm in order to ensure that the final estimator respects a desired level of differential privacy. Intuitively, the larger the number of data (gradient) queries of the algorithm, the more prone it will be to privacy leakage. Our private version of gradient descent considers the following noisy version of the iterates (4):

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\eta}{n} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)}) + \frac{2\eta B \sqrt{K}}{\mu n} Z_k, \tag{5}$$

where the final estimate is again denoted by $\theta^{(K)}$. Here, η is the stepsize, $B > 0$ is a constant, and $\{Z_k\}_{k=0}^{K-1}$ is a sequence of i.i.d. standard p-dimensional Gaussian random vectors. The number of iterations K needs to be set beforehand and critically impacts the statistical performance of this estimator, as discussed below. We note that many related variants of this noisy gradient descent procedure exist in the literature [Bassily et al., 2014, Feldman et al., 2020, Lee and Kifer, 2018, Song et al., 2013]; however, our analysis provides novel insights into the properties of this algorithm, as our convergence analysis relies on local strong convexity and provides a general consistency and asymptotic normality theory for differentially private M-estimators.

3.1 Convergence analysis

Taking $B = 0$ in the iterates (5) recovers the standard gradient descent algorithm. Classical optimization theory tells us that for the choice $K = O(\log(1/\Delta))$, gradient descent incurs an optimization error of $\|\theta^{(K)} - \hat{\theta}\|_2 = O(\Delta)$. Therefore, $\log(n)$ steps suffice if we want to make sure that the optimization error is of the same order as the parametric convergence rate $O\left(\sqrt{\frac{p}{n}}\right)$. We will require local strong convexity (LSC) of the loss function $\mathcal{L}_n$ in a ball of radius r around the true parameter

θ_0, as well as global smoothness. Strong convexity and smoothness are both standard conditions for the convergence analysis of gradient-based optimization methods [Boyd and Vandenberghe, 2004]. Several useful properties of strongly convex and smooth functions are reviewed in Appendix B.

Condition 1 (Local strong convexity/smoothness)**.** *The loss function $\mathcal{L}_n$ is locally τ_1-strongly convex and τ_2-smooth, i.e.,*

$$\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_2) \geq \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \tau_1 \|\theta_1 - \theta_2\|_2^2, \quad \forall \theta_1, \theta_2 \in \mathcal{B}_r(\theta_0),$$

and

$$\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_2) \leq \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \tau_2 \|\theta_1 - \theta_2\|_2^2, \quad \forall \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^p.$$

The following theorem shows that the iterates (5) with $B \geq \sup_{x \in \mathcal{X}, \theta \in \Theta} \|\Psi(x, \theta)\|_2$ are μ-GDP and converge to the non-private M-estimator as the sample size increases. More specifically, it shows that the private iterates lie in a neighborhood of the target non-private M-estimator whose radius is directly comparable to the privacy-inducing noise added in each noisy gradient descent step. The proof is in Appendix D.1, and hinges on the fact that the assumptions on the loss function and initial value ensure that $\theta^{(0)}$ and all successive iterates lie in the LSC ball $\mathcal{B}_r(\theta_0)$, with high probability (cf. Lemmas 17 and 18.)

Theorem 2. *Assume $\mathcal{L}_n$ satisfies Condition 1, and suppose $\mathcal{L}_n(\theta^{(0)}) - \mathcal{L}_n(\hat{\theta}) \leq \frac{r^2}{4}\tau_1$ and $\hat{\theta} \in \mathcal{B}_{r/2}(\theta_0)$. Suppose $\mathcal{L}_n$ is twice-differentiable almost everywhere in $\mathcal{B}_r(\theta_0)$. Further let $\eta \leq \frac{1}{2}\min\left\{\frac{1}{\tau_2}, 1\right\}$, $n = \Omega\left(\frac{\sqrt{K \log(K/\xi)}}{\mu}\right)$, and $K = \Omega(\log n)$. Then*

(i) $\theta^{(K)}$ is μ-GDP, and

(ii) $\|\theta^{(K)} - \hat{\theta}\|_2 \leq C\frac{B\sqrt{K}(\sqrt{p}+\sqrt{2\log(K/\xi)})}{\mu n}$, with probability at least $1 - \xi$, where C is a constant depending on τ_1, τ_2, and η.

Remark 1. *Here and in the sequel, we treat τ_1, τ_2, and r as constants, and concern ourselves primarily with the dependence of the error bounds and sample size requirements on n and K. Note, however, that Condition 1 is a statement about the empirical loss function $\mathcal{L}_n$, so the parameters (τ_1, τ_2, r) could in principle also be functions of n. However, in the M-estimation settings we consider, it can be shown that Condition 1 holds with high probability for fixed values of (τ_1, τ_2, r) when n is sufficiently large, i.e., under the prescribed minimum sample size conditions appearing in the statements of the theorems. (For additional details, see the discussion in Section 4.3 below.)*

Theorem 2 immediately allows us to derive statistical properties of the noisy gradient descent estimator. The following result is a consequence of the triangle inequality and standard M-estimation theory for $\hat{\theta}$, e.g., [Huber and Ronchetti, 2009, Corollary 6.7], which guarantees that $\|\hat{\theta} - \theta_0\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right)$.

Corollary 1. *Assume the same conditions as in Theorem 2, and suppose θ_0 is such that $\mathbb{E}[\Psi(Z, \theta_0)] = 0$ and $\mathbb{E}_F[\dot{\Psi}(x, \theta)]$ is continuous and invertible in a neighborhood of $\theta = \theta_0$. Then*

(i) $\theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K \log K}}{\mu n}\right)$, and

(ii) $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F_{\theta_0}))$.

Remark 2. *Corollary 1(i) implies that* $\theta^{(K)} = \theta_0 + O_p\left(\sqrt{\frac{p}{n}} + \frac{\sqrt{K \log K}}{\mu n}\right)$. *Thus, we see that the extra* $O_p\left(\frac{\sqrt{K \log K}}{\mu n}\right)$ *term is the "cost of privacy" of our noisy gradient descent algorithm. Note that lower bounds of* $\Omega_p\left(\frac{1}{\epsilon n}\right)$ *on the cost of privacy were derived for the regression setting in Cai et al. [2019] and Cai et al. [2020] under the framework of* (ϵ, δ)*-DP. If we were to adopt the* (ϵ, δ)*-DP framework in our analysis, the noise introduced at each iterate would be of the form* $\Theta\left(\frac{K\sqrt{\log(c/\delta)}}{\epsilon n}\right)\eta Z_k$ *rather than* $\Theta\left(\frac{\sqrt{K}}{\mu n}\right)\eta Z_k$, *so the same optimization-theoretic analysis in Theorem 2 would lead to a cost of privacy of* $O_p\left(\frac{\sqrt{\log(c/\delta)}}{\epsilon n}\right)$, *matching the known lower bounds up to* $\log n$ *factors. Indeed, the same lower bounds for* (ϵ, δ)*-DP, combined with the fact that an algorithm is* μ*-GDP if and only if it is* $(\mu, \delta(\mu))$*-DP, for* $\delta(\mu) = \Phi(-1 + \mu/2) - e^{\mu}\Phi(-1 - \mu/2)$ *[Dong et al., 2021, Corollary 2.13], can be used to derive a lower bound of* $\Omega_p\left(\frac{1}{\mu n}\right)$ *on the cost of privacy in the* μ*-GDP setting for small values of* μ, *showing that our estimation error in Corollary 1 is minimax optimal (up to logarithmic factors).*

Remark 3. *A more careful analysis would allow us to remove the twice-differentiability assumption on* $\mathcal{L}_n$ *in Theorem 2 and Corollary 1, since the gradient descent algorithm only requires one derivative, and asymptotic normality of M-estimators does not require the loss function to be twice-differentiable [Huber, 1967].*

Theorem 2 requires the suboptimality gap of the objective function between the initial value $\theta^{(0)}$ and the global optimum $\hat{\theta}$ to be bounded by $\frac{r^2}{4}\tau_1$. The following proposition, proved in Appendix D.2, shows that an initial fixed number of noisy gradient descent iterations can be used to ensure that the starting value condition is met.

Proposition 1. *Assume Condition 1 holds and* $\eta \leq \frac{1}{2\tau_2}$. *Let* $R = \|\theta^{(0)} - \hat{\theta}\|_2$. *Then there exists a constant* $c_0 > 0$ *such that with probability at least* $1 - \xi_0$, *after* $K_0 = \frac{R^2}{\eta\Delta}$ *noisy gradient descent iterations (5) and for* $n \geq C_0 \frac{(R+K_0+1)\{4\sqrt{p}+2\sqrt{2\log(K_0/\xi_0)}\}\sqrt{K_0}}{\Delta\mu}$, *we have*

$$\mathcal{L}_n(\theta^{(K_0)}) - \mathcal{L}_n(\hat{\theta}) \leq \Delta.$$

Here, C_0 *is a constant which depends on* η *and* B.

Taking $\Delta = \frac{r^2}{4}\tau_1$ in Proposition 1, we see that $K_0 = \frac{4}{\tau_1 r^2 \eta}\|\theta^{(0)} - \hat{\theta}\|_2^2$ iterations of noisy gradient descent are sufficient to ensure that $\theta^{(K_0)}$ meets the initialization condition of Theorem 2. We remark that both Theorem 2 and Proposition 1 require the minimum sample size scaling $n = \Omega\left(\frac{\sqrt{K \log K}}{\mu}\right)$. Therefore, the additional $K_0 \asymp \|\theta^{(0)} - \theta_0\|_2^2$ iterations needed to ensure the starting value assumption of Theorem 2 does not substantively affect the conclusion of the theorem, since that result already assumes that $K = \Omega(\log n)$.

Remark 4. *The expression for the minimum sample size in Proposition 1 involves* R, *which is defined in terms of* $\hat{\theta}$ *and is therefore also a function of* n. *However, recall that we are working in a scenario where* $\|\hat{\theta} - \theta_0\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right)$, *so* $R = \|\theta^{(0)} - \theta_0\|_2 + o_p(1)$.

3.2 Examples

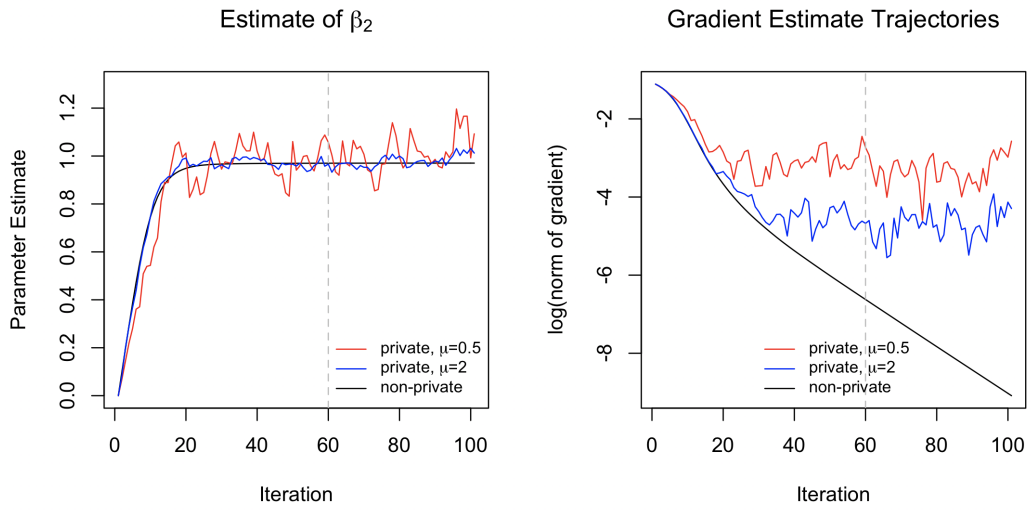
We now present two numerical simulations to illustrate the theory thus far. In particular, we demonstrate the behavior of noisy gradient trajectories and the benefits of our proposed approach based on M-estimation, in contrast to gradient clipping.

3.2.1 Linear regression

We now explore the behavior of the noisy gradient descent algorithm on simulated data from a linear regression model. The data $\{(x_i, y_i)\}_{i=1}^n$ are generated according to the model $y_i = x_i^T\beta + \epsilon_i$, where $\epsilon_1, \ldots, \epsilon_n$ are an i.i.d. sample from $N(0, \sigma^2)$ and the covariate vectors are given by $x_i = (1, z_i)^T$, where $z_i \overset{i.i.d.}{\sim} N(0, \sigma_z^2 \mathbb{I}_3)$. We take our loss function to be

$$\mathcal{L}_n(\beta, \sigma) = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma\rho_c\left(\frac{y_i - x_i^T\beta}{\sigma}\right) + \frac{1}{2}\kappa_c\sigma\right)w(x_i), \tag{6}$$

where ρ_c is the Huber loss function with tuning parameter c, the constant κ_c is chosen to ensure consistency of $\hat{\sigma}$, and $w(x_i) = \min\left(1, \frac{2}{\|x_i\|_2^2}\right)$ downweights outlying covariates. By construction, the gradient of this loss function with respect to parameter vector $\theta = (\beta, \sigma)$ has finite global sensitivity, allowing us to use the noisy gradient descent iterates (5) to estimate β and σ. Specifically, the global sensitivity of $\nabla_\beta \mathcal{L}_n(\beta, \sigma)$ equals $2\sqrt{2}c$, and the global sensitivity of $\nabla_\sigma \mathcal{L}_n(\beta, \sigma)$ is $\frac{1}{2}c^2$, resulting in global sensitivity of $\sqrt{8c^2 + \frac{1}{4}c^4} \coloneqq 2B$ for$\nabla_\theta \mathcal{L}_n(\theta)$. Setting $\beta = (1, 1, 1, 1)^T$, $\sigma = 2 = \sigma_z$, $c = 1.345$, and $n = 1000$, Figure 1 below plots sample trajectories of the noisy gradient descent iterates for the coordinate of the parameter vector corresponding to β_2.



(a) (b)

Figure 1: Noisy gradient descent trajectories for linear regression. (a) Estimates of a single coordinate of the regression vector. (b) Gradient of the loss function evaluated at the current iterate, plotted on a log scale.

Figure 1(a) shows that in the early iterations, both the private and non-private estimates move away from the initial value toward the true parameter value. While the non-private version converges to the true parameter value in later iterations, the μ-GDP version varies in a window around the true value as the iterates progress.

Since the random noise added to the gradient at each iteration (5) has the same fixed variance for a given sample size, the gradient of our loss function does not become arbitrarily small as the number of iterations increases, nor do the values at successive iterations become arbitrarily close to each other (cf. Figure 1(b)). From a practical standpoint, we can assess convergence of

our algorithm by considering whether the gradient of the loss function is still large relative to the standard deviation of the random noise term. As noted above, the maximum number of iterations K must be set beforehand. However, if the loss function gradient is already small relative to the standard deviation of the noise term at some iteration $k < K$, empirical evidence suggests no practical advantage to continuing through all K budgeted iterations to obtain $\left(\hat{\beta}^{(K)}, \hat{\sigma}^{(K)}\right)$.

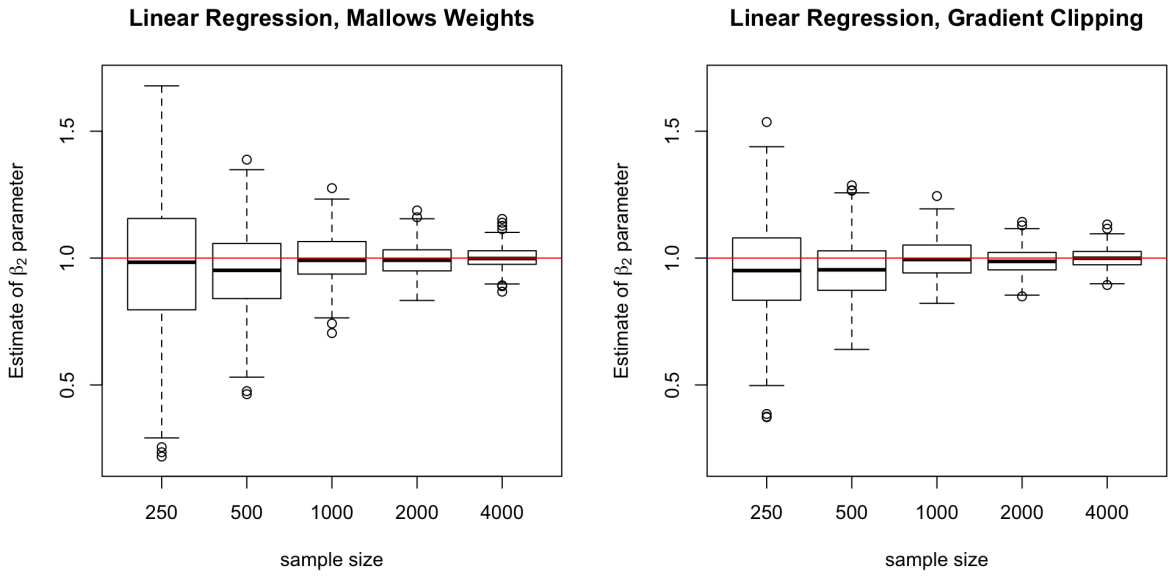
3.2.2 Clipping and logistic regression

Applying noisy gradient descent to optimize an objective function with bounded gradients is an alternative to explicitly clipping the gradient values so as to achieve finite global sensitivity. One motivation for avoiding such clipping is that the resulting estimators may fail to be consistent. To fix ideas, suppose $x_1, \ldots, x_n \in \mathcal{X} \subseteq \mathbb{R}^m$ are i.i.d. according to F_{θ_0}. If one seeks to compute a differentially private maximum likelihood estimator via clipped gradients, one will in fact be computing a differentially private counterpart of the clipped maximum likelihood estimator $\tilde{\theta} = T(F_n)$ defined as the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n} h_c\left(\nabla \log f(x_i; \tilde{\theta})\right) = 0,$$

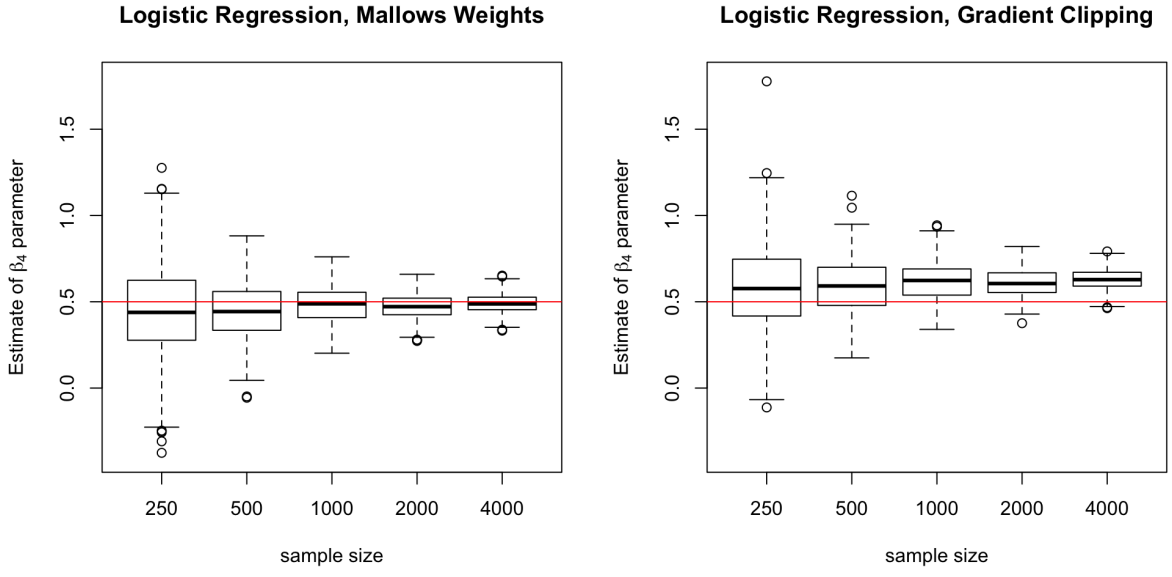
where $f(\cdot; \theta_0)$ is the density function of the model and $h_c(z) = z \min\left\{1, \frac{c}{\|z\|_2}\right\}$ is the multivariate Huber function [Hampel et al., 1986, p.239]. While clipping guarantees a bounded sensitivity of the estimating equations, the clipped maximum likelihood estimator is in general not consistent, since the estimating equations are in general not unbiased, i.e., $\mathbb{E}_{F_{\theta_0}}[h_c(\nabla \log f(x_i; T(F_{\theta_0})))] \neq 0$. Hence, even though gradient clipping is a common suggestion in the differential privacy literature, it is not the most appealing from a statistical viewpoint.

We note that in a classical linear regression setting with squared loss and symmetric errors about zero, clipping does not lead to inconsistent estimators. However, we do encounter this issue when estimating the parameters of a logistic regression model (with the cross-entropy loss). For example, we compare the performance of our proposed noisy gradient descent algorithm (insert reference) on simulated data from a linear regression model and a logistic regression model, using either Mallows weights (as in (6)) or gradient clipping.



(a) (b)

Figure 2: Gradient clipping and consistency. In the linear regression setting, both Mallows weights (plot (a)) and gradient clipping (plot (b)) approaches are consistent.



(a) (b)

Figure 3: Gradient clipping and consistency. Clipping results in a positive bias for logistic regression (plot (b)), which does not arise using Mallows weights (plot (a)).

Figure 2 compares the results of (a) a clipping procedure and (b) Mallows weights simulated linear regression data, and illustrates that both methods lead to consistent estimators. Figure 3 compares these methods on simulated data from a logistic regression model, and illustrates that the

parameter estimates obtained via gradient clipping exhibit bias which does not shrink toward zero as the sample size increases. For the linear regression simulation, data were generated according to the model $y_i = x_i^T\beta + \epsilon_i$, where $\beta = (1.5, 1, -1, 0.5)^T$, $\epsilon_1, \ldots, \epsilon_n$ are an i.i.d. sample from $N(0, 2^2)$, and the covariate vectors are given by $x_i = (1, z_i)^T$, where $z_i \overset{i.i.d.}{\sim} N(0, I_3)$. At each sample size, 400 repetitions were performed. The gradient of the loss function was clipped such that its ℓ_2-norm was no larger than 1. For the logistic regression simulation, data were generated from the model $y_i \sim \text{Bernoulli}\left(\left\{1 + \exp(-x_i^T\beta)\right\}^{-1}\right)$, with the same value of β and the same scheme for generating the covariates x_i as in the linear regression simulation. Again, 400 repetitions were performed at each sample size and the loss function gradient was clipped at a threshold of 1.

4 Randomized M-estimators via noisy Newton's method

We now present a noisy version of Newton's method, as a second-order alternative to the noisy gradient descent algorithm described in Section 3. Recall that the classical Newton-Raphson algorithm finds the global optimum $\hat{\theta}$ of the objective (1) via the iterations

$$\theta^{(k+1)} = \theta^{(k)} - \left(\sum_{i=1}^n \dot{\Psi}(x_i, \theta^{(k)})\right)^{-1} \sum_{i=1}^n \Psi(x_i, \theta^{(k)}). \tag{7}$$

The key difference between the Newton-Raphson optimization procedure and gradient descent is the use of the Hessian term $\nabla^2\mathcal{L}_n$. Our differentially private version of this algorithm will therefore add noise to both the gradient and the Hessian of the empirical loss function. Note that we will assume throughout this section that $\nabla^2\mathcal{L}_n$ exists everywhere.

4.1 Matrix-valued noise for private Hessians

It is important for the convergence of the iterations (7) that the matrix $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \dot{\Psi}(x_i, \theta)$ be positive definite. Since we will use noisy versions of this matrix, we also want to guarantee that the randomized quantities remain positive definite.

The approach that we will consider exploits the fact that in many M-estimation problems, the matrix $M_n(\theta)$ can be viewed as an empirical covariance matrix of the form $\frac{1}{n}\sum_{i=1}^n a_i a_i^\top$. An intuitive idea for outputting a differentially private matrix is to add i.i.d. noise to each individual component. The following result shows that we can indeed add a symmetric matrix with appropriately scaled element-wise i.i.d. Gaussian noise [Dwork et al., 2014, Algorithm 1]:

Lemma 1 (Matrix Gaussian mechanism)**.** *Consider a data matrix $A \in \mathbb{R}^{n\times m}$ such that each row vector a_i satisfies $\|a_i\|_2 \le 1$. Further define the function $h(A) = \frac{1}{n}A^\top A$, and let W be a symmetric random matrix whose upper-triangular elements, including the diagonal, are i.i.d. $\frac{1}{\mu}N(0,1)$. Then the random function $\tilde{h}(A) = h(A) + W$ is μ-GDP.*

Remark 5. *One obvious drawback of the matrix Gaussian mechanism described in Lemma 1 is that the noise is not positive definite. This could be problematic for the computation of differentially private positive definite Hessians, especially for small sample sizes. Note, however, that $\tilde{h}(A)$ can be projected onto a cone of positive definite matrices $\{H : H \succeq \varepsilon I\}$ defined by $\varepsilon > 0$ via the convex optimization problem*

$$\tilde{h}(A)_+ = \arg\min_{H\succeq\varepsilon I} \|H - \tilde{h}(A)\|_2.$$

By definition, $\|\tilde{h}(A)_+ - \tilde{h}(A)\|_2 \le \|h(A) - \tilde{h}(A)\|_2$, so the triangle inequality yields

$$\|\tilde{h}(A)_+ - h(A)\|_2 \le \|\tilde{h}(A)_+ - \tilde{h}(A)\|_2 + \|\tilde{h}(A) - h(A)\|_2 \le 2\|\tilde{h}(A) - h(A)\|_2.$$

Hence, the price to pay for the projection is no more than a factor of two in the spectral norm error, which does not affect the order of the convergence rate. In practice, the projection amounts to truncating the eigenvalues as $\max\{\lambda_j, \varepsilon\}$ *[Boyd and Vandenberghe, 2004, p.399]. The projected matrix is clearly also differentially private, as it results from a deterministic post-processing step applied to a differentially private output.*

4.2 Differentially private Newton's method and convergence analysis

We will require some regularity conditions on the Hessian matrix $\nabla^2 \mathcal{L}_n(\theta)$. In particular, we need the Hessian to be factorizable in a way that allow us to leverage Lemma 1. This Hessian structure is typical of loss minimization problems with linear predictors such as linear regression, robust regression, and generalized linear models. We will also assume that the spectral norm of the Hessian is uniformly bounded.

Condition 2 (Hessian assumptions)**.** *The Hessian is positive definite for all* $\theta \in \mathcal{B}_r(\theta^*)$ *and is of the form* $\nabla^2 \mathcal{L}_n(\theta) = \frac{1}{n} A^\top A = \frac{1}{n} \sum_{i=1}^n a(x_i, \theta) a(x_i, \theta)^\top$*, where* $a : \mathcal{X} \times \Theta \mapsto \mathbb{R}^m$ *and* $\sup_{x,\theta} \|a(x, \theta)\|_2^2 \leq \bar{B} < \infty$*.*

We note that the constant $\bar{B}$ introduced in Condition 2 implies τ_2-smoothness with $\tau_2 = \frac{\bar{B}}{2}$.

We are now ready to present our differentially private counterpart of the Newton iterates (7). We propose a noisy damped quasi-Newton method that follows the updates

$$\theta^{(k+1)} = \theta^{(k)} - \eta H_k^{-1} \left(\frac{1}{n} \sum_{i=1}^n \Psi(x_i, \theta^{(k)}) + \frac{2B\sqrt{2K}}{\mu n} Z_k \right), \tag{8}$$

where $\eta > 0$ is the stepsize, $B \geq \sup_{x \in \mathcal{X}, \theta \in \Theta} \|\Psi(x, \theta)\|_2$ upper-bounds the gradients as before, $\bar{B}$ is as in Condition 2, $\{Z_k\}_{k=1}^K$ is a sequence of i.i.d. standard p-dimensional Gaussian random vectors, and

$$H_k = \frac{1}{n} \sum_{i=1}^n \dot{\Psi}(x_i, \theta^{(k)}) + \frac{2\bar{B}\sqrt{2K}}{\mu n} W_k$$

is a noisy Hessian, where $\{W_k\}_{k=1}^K$ is a sequence of i.i.d. symmetric random matrices whose upper-triangular elements, including the diagonals, are i.i.d. standard normal. In practice, one can set the smallest eigenvalues of H_k to some small positive value ε, as discussed in Remark 5. If $\varepsilon \to 0$, the effect will be asymptotically negligible and will not affect the theoretical conclusions stated in our paper.

Finally, we note that the level of noise introduced to both the gradient and Hessian terms to ensure privacy is $O\left(\frac{1}{n}\right)$, which is appreciably smaller than the $\Theta\left(\frac{1}{\sqrt{n}}\right)$ fluctuations involved if we were to treat the gradient and Hessian as empirical versions of $\nabla \mathcal{L}_n(\theta^{(k)})$ and $\nabla^2 \mathcal{L}_n(\theta^{(k)})$ (or the level of noise introduced by subsampling-type stochastic optimization methods). Thus, we believe that the analysis of the iterative algorithms studied in this paper has not previously appeared in the optimization theory literature.

4.2.1 Local strong convexity theory

Assuming local strong convexity, we can show that the noisy Newton algorithm leads to the same statistical error bounds as noisy gradient descent, but requires fewer iterations to achieve convergence. In order to establish this result, we will also require the Hessian to be Lipschitz continuous, a condition which is commonly used to establish the quadratic convergence of Newton's method under strong convexity [Boyd and Vandenberghe, 2004, Ch 9.5]:

Condition 3 (Lipschitz continuity of Hessian). *The Hessian $\nabla^2 \mathcal{L}_n(\theta)$ is L-Lipschitz continuous, i.e., $\|\nabla^2 \mathcal{L}_n(\theta_1) - \nabla^2 \mathcal{L}_n(\theta_2)\|_2 \le L\|\theta_1 - \theta_2\|_2$ for all $\theta_1, \theta_2 \in \mathbb{R}^p$.*

The following theorem, proved in Appendix E.1, shows that under standard regularity conditions for robust M-estimators, $\Omega(\log\log n)$ iterations of the noisy Newton step (8) suffice to obtain a μ-GDP estimator lying in a neighborhood of $\hat{\theta}$ whose radius is proportional to the privacy-inducing noise of the algorithm.

Theorem 3. *Assume Conditions 1, 2, and 3 hold, and suppose $\hat{\theta} \in \mathcal{B}_{r/2}(\theta_0)$ and $\|\nabla \mathcal{L}_n(\theta^{(0)})\|_2 \le \min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\}$. Further let $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$ and $K = \Omega(\log\log n)$. Then the K^{th} noisy Newton iterate (8) with $\eta = 1$ satisfies*

(i) $\theta^{(K)}$ is μ-GDP, and

(ii) $\|\theta^{(K)} - \hat{\theta}\|_2 \le C\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu n}$, with probability at least $1-\xi$, where $C > 0$ is a constant depending on L, τ_1, B, and $\bar{B}$.

Remark 6. *Theorem 3 imposes a slightly different initialization condition than Theorem 2, namely that the gradient of the loss at the initial point must be small. However, this condition can similarly be guaranteed after a few initial iterates of noisy gradient descent: taking $\Delta = \min\left\{\frac{r^2}{4}\tau_1, \frac{\tau_1^3 r^2}{4\tau_2^2}, \frac{\tau_1^5}{4\tau_2^2 L^2}\right\}$ implies that $\theta^{(K_0)} \in \mathcal{B}_r(\theta_0)$ by Proposition 1 and Lemma 17, where $K_0 = c_0\|\theta^{(0)} - \hat{\theta}\|_2^2$. Hence, local strong convexity further implies that $\Delta \ge \mathcal{L}_n(\theta^{(K)}) - \mathcal{L}_n(\hat{\theta}) \ge \tau_1\|\theta^{(K)} - \hat{\theta}\|_2^2$, which combined with smoothness gives*

$$\|\nabla \mathcal{L}_n(\theta^{(K_0)})\|_2 \le 2\tau_2\|\theta^{(K)} - \hat{\theta}\|_2 \le 2\tau_2\sqrt{\frac{\Delta}{\tau_1}} \le \min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\}.$$

We note that in practice, in order to benefit from the improved iteration complexity of the noisy Newton algorithm, one should be able to assess whether the initial condition $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \le \frac{\tau_1^2}{L}$ is met. If this condition fails to hold, the algorithm can diverge, as illustrated in Section 4.4. We note that this is analogous to the well-known behavior of Newton's method. This drawback has been addressed in the literature by backtracking, but even that analysis relies on strong convexity [Boyd and Vandenberghe, 2004, Ch. 9.5]. In the context of differential privacy, there is no obvious counterpart of backtracking, since this technique involves evaluating the objective function at various candidate iterates. We propose an alternative solution in Section 4.4.

Finally, as in the case of Corollary 1, it follows directly from Theorem 3 and standard M-estimation theory that the noisy Newton algorithm leads to μ-GDP estimators that are $\sqrt{n}$-consistent and asymptotically normally distributed.

Corollary 2. *Assume the conditions of Theorem 3 and let θ_0 be such that $\mathbb{E}[\Psi(Z, \theta_0)] = 0$ and $\mathbb{E}_F[\dot{\Psi}(x, \theta)]$ is continuous and invertible in a neighborhood of $\theta = \theta_0$. Then*

(i) $\theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K\log K}}{\mu n}\right)$, and

(ii) $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F_{\theta_0}))$.

4.2.2 Self-concordance theory

We now derive results for global convergence of our noisy Newton algorithm, under an alternative assumption of self-concordance. As discussed in Section 4.2.1 above, fast convergence of the noisy Newton algorithm (Theorem 3) is only guaranteed under a suitably close initialization, which we propose to obtain via an initial batch of iterates from noisy gradient descent (cf. Remark 6). However, one practical drawback of the local strong convexity analysis is that we need a method for detecting when the gradient condition $\|\nabla \mathcal{L}_n(\theta^{(0)})\|_2 \leq \frac{\tau_1^2}{L}$ is obtained by the noisy gradient descent iterates, which requires (approximately) computing the LSC parameter τ_1. As the results of this section show, imposing a self-concordance rather than LSC assumption on $\mathcal{L}_n$ allows us to (a) obtain a more easily checked initial value condition on $\theta^{(0)}$, and (b) use noisy Newton iterates for the entire duation of the algorithm, rather than switching from noisy gradient descent to noisy Newton. As the examples in Section 4.3 will illustrate, the self-concordance property is indeed satisfied by several robust M-estimators that are useful both from a privacy and statistical error perspective.

We will use the following notion of generalized self-concordance [Sun and Tran-Dinh, 2019]:

Definition 4 (Generalized self-concordance)**.** *A univariate function $f : \mathbb{R} \to \mathbb{R}$ is (γ, ν)-self-concordant if*

$$|f'''(x)| \leq \gamma \left(f''(x)\right)^{\nu/2},$$

for all x. A multivariate function $f : \mathbb{R}^p \to \mathbb{R}$ is (γ, ν)-self-concordant if

$$\left|\langle \nabla^3 f(x)[v]u, u\rangle\right| \leq \gamma \|u\|^2_{\nabla^2 f(x)} \|v\|^{\nu-2}_{\nabla^2 f(x)} \|v\|_2^{3-\nu},$$

for all $x, u, v \in \mathbb{R}^p$.

We will primarily be interested in the cases $\nu = 2$ and $\nu = 3$. Appendix C contains several useful results about generalized self-concordant functions.

The following theorem concerns convergence guarantees for the noisy Newton iterates assuming a starting value condition, which is stated in terms of the Newton decrement function

$$\lambda(\theta) := \left((\nabla \mathcal{L}_n(\theta))^T \nabla^2 \mathcal{L}_n(\theta)^{-1} (\nabla \mathcal{L}_n(\theta))\right)^{1/2}.$$

The Newton decrement is a standard quantity which appears in the usual convergence analysis of Newton's method under self-concordance. The proof of the theorem leverages stability properties of the Hessian of generalized self-concordant functions, as outlined in Appendix C.2, allowing us to derive uniform lower bounds on the minimum eigenvalue of the Hessian at successive iterates without requiring local strong convexity. In particular, Lemma 16 introduces a quantity $\tau_{1,r}$ that bounds the minimum eigenvalues of Hessians at successive iterates, which we use with $r = \frac{2}{\gamma}$ in place of the LSC parameter.

Theorem 4. *Suppose $\mathcal{L}_n$ satisfies Condition 2. Further assume that $\mathcal{L}_n$ is $(\gamma, 2)$-self-concordant, and suppose $\hat{\theta} \in \mathcal{B}_{1/\gamma}(\theta_0)$ and $\lambda_{\min}^{-1/2}(\nabla^2 \mathcal{L}_n(\theta^{(0)}))\lambda(\theta^{(0)}) \leq \frac{1}{16\gamma}$. Let $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$ and $K = \Omega(\log\log n)$. Then the K^{th} noisy Newton iterate (8) with $\eta = 1$ satisfies*

(i) $\theta^{(K)}$ is μ-GDP, and

(ii) $\|\theta^{(K)} - \hat{\theta}\|_2 \leq C\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu n \tau^2_{1,2/\gamma}}$, with probability at least $1-\xi$, where $C > 0$ is a constant depending on γ, B, and $\bar{B}$.

The proof of Theorem 4 is provided in Appendix E.2. Note that whereas the traditional analysis of Newton's method via (canonical) self-concordance proceeds by showing that $\lambda(\theta^{(k)})$ decays quadratically with k, our proof for $(\gamma, 2)$-self-concordant functions, based on adapting the arguments of Sun and Tran-Dinh [2019], first establishes quadratic convergence of the related quantity $\lambda_{\min}^{-1/2}(\nabla^2 \mathcal{L}_n(\theta^{(k)}))\lambda(\theta^{(k)})$.

Comparing the initial value assumptions of Theorem 4 with those of Theorem 3, we see that the only condition we need to check involves evaluating the rescaled Newton decrement at $\theta^{(0)}$ with the value $\frac{1}{16\gamma}$. In practice, γ is directly computable from the form of the M-estimator used in the regression problem (cf. Section 4.3 below). Hence, the initial condition for self-concordant functions is much more practically checkable than the one stated in Theorem 3 for functions which satisfy LSC.

Remark 7. *In light of the results on second-order inexact oracle algorithms for $(\gamma, 3)$-self-concordant functions in the optimization literature, it is natural to wonder if a version of Theorem 4 could also be derived for $(\gamma, 3)$-self-concordant functions. Indeed, as the examples in Section 4.3 will show, loss functions of interest have been previously proposed which were designed to be $(\gamma, 3)$-self-concordant. In fact, the convergence results stated in Theorem 4 actually hold for (γ, ν)-self-concordant losses, for any $\nu \geq 2$: By Lemma 10, the assumption that $\mathcal{L}_n$ is (γ, ν)-self-concordant, together with Condition 2, implies that $\mathcal{L}_n$ is also $(\bar{B}^{\nu/2-1}\gamma, 2)$-self-concordant.*

The following proposition can be used to establish global convergence under the same conditions as Theorem 4. It shows that the suboptimality gap of successive noisy Newton iterates decreases geometrically, without assuming a sufficiently close initialization. The proof is contained in Appendix E.3; the main argument adapts the global convergence analysis of Karimireddy et al. [2018] for (non-noisy) Newton's method under self-concordance. In the statement of the theorem, we abuse notation slightly and use τ_{1,R_0} to denote the quantity defined in Lemma 16 with θ_0 replaced by the global optimum $\hat{\theta}$.

Proposition 2. *Suppose $\mathcal{L}_n$ is a $(\gamma, 2)$-self-concordant function which satisfies Condition 2. Let $\Delta_0 = \mathcal{L}_n(\theta^{(0)}) - \mathcal{L}_n(\hat{\theta})$, and define $R_0 = g^{-1}\left(\frac{\gamma^2 \Delta_0}{\lambda_{\min}(\nabla^2 \mathcal{L}_n(\hat{\theta}))}\right)$, where $g(t) := e^{-t} + t - 1$ for $t > 0$. Let $K \geq 1$, and suppose the stepsize satisfies $\eta \leq \min\left\{\frac{1}{2\exp(2\gamma R_0)}, \frac{4\tau_{1,R_0}^2}{\bar{B}^2}\right\}$ and the sample size satisfies $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$. Then with probability at least $1 - \xi$, the Newton iterates satisfy*

$$\mathcal{L}_n(\theta^{(K)}) - \mathcal{L}_n(\hat{\theta}) \leq \left(1 - \frac{\eta}{\exp(2\gamma R_0)}\right)^K \left(\mathcal{L}_n(\theta^{(0)}) - \mathcal{L}_n(\hat{\theta})\right) + r_{priv},$$

where $r_{priv} = C' \frac{\sqrt{Kp\log(Kp/\xi)}}{\mu n \tau_{1,R_0}^2}$, and $C' > 0$ is a constant depending on τ_{1,R_0}, B, and $\bar{B}$.

Remark 8. *As argued in Remark 6, the suboptimality bound*

$$\mathcal{L}_n(\theta) \leq \mathcal{L}_n(\hat{\theta}) + \Delta$$

implies that $\|\nabla \mathcal{L}_n(\theta)\|_2 \leq \bar{B}\sqrt{\frac{\Delta}{\tau_{1,R_0}}}$, for $\Delta \leq \Delta_0$, since $\theta \in \mathcal{B}_{R_0}(\hat{\theta})$ by Lemma 27 and τ_{1,R_0} takes the place of the local strong convexity parameter. Hence, if we take

$$\Delta \leq \min\left\{\Delta_0, \frac{\tau_{1,R_0}}{(32\gamma\bar{B})^2}\right\},$$

we can guarantee that

$$\lambda(\theta) = \|\nabla \mathcal{L}_n(\theta)\|_{\nabla^2 \mathcal{L}_n(\theta)^{-1}} \leq 2\tau_{1,R_0} \cdot \bar{B} \sqrt{\frac{\Delta}{\tau_{1,R_0}}} \leq \frac{1}{16\gamma}.$$

This means that Proposition 2 can be used to ensure that the starting condition of Theorem 4 holds after an initial batch of $O\left(\log\left(\min\left\{\Delta_0, \frac{\tau_{1,R_0}}{(32\gamma\bar{B})^2}\right\}\right)\right)$ *iterates, which is negligible in light of the* $K = \Omega(\log\log n)$ *iterates used in Theorem 4.*

Remark 9. *One can further refine the suboptimality bound of Proposition 2 in order to obtain convergence rates for some iterate* $\theta^{(K)}$ *and* K *sufficiently large. In particular, one can adapt the argument provided in the proof of Theorem 2 in order to guarantee that with high probability,* $\|\theta^{(K)} - \hat{\theta}\|_2 = O(r_{priv})$*; however, the suboptimality guarantee as stated is sufficient for our purposes, since it already implies a negligible initial number of iterations prior to applying Theorem 4.*

As in the case of Corollaries 1 and 2, Theorem 4 immediately allows us to derive the following corollary:

Corollary 3. *Assume the conditions of Theorem 4 and let* θ_0 *be such that* $\mathbb{E}[\Psi(Z, \theta_0)] = 0$ *and* $\mathbb{E}_F[\dot{\Psi}(x, \theta)]$ *is continuous and invertible in a neighborhood of* $\theta = \theta_0$*. Then*

(i) $\theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K\log(K)}}{\mu n}\right)$*, and*

(ii) $\sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F_{\theta_0}))$*.*

4.3 From univariate to multivariate self-concordant functions

Although we have proved our optimization results for general multivariate functions, our primary focus in this paper is regression M-estimators. Accordingly, we now discuss how univariate self-concordant functions can be composed to obtain M-estimators which are multivariate self-concordant. We will focus on (γ, ν)-self-concordance with $\nu = 2$ or 3. We will discuss specific examples in the following subsection.

Regression with Mallows weights: We first consider Mallows-weighted estimators of the form

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho(y_i, x_i^T\theta) w(x_i). \tag{9}$$

One can leverage recent results in Sun and Tran-Dinh [2019] to show that if ρ is univariate $(\gamma, 2)$-self-concordant, that the Mallows estimator is also (multivariate) self-concordant:

Lemma 2. *Suppose* ρ *is* $(\gamma, 2)$*-self-concordant. Then the Mallows loss (9) is* $(\gamma \max_i \|x_i\|_2, 2)$*-self-concordant.*

Proof. Note that Lemma 9 implies that $\rho(y_i, x_i^T\theta)$ is $(\gamma\|x_i\|_2, 2)$-self-concordant. Then Lemma 8 implies that $\mathcal{L}_n$ is $(\gamma \max_i \|x_i\|_2, 2)$-self-concordant. □

As Lemma 2 suggests, Mallows weighting is only desirable when we assume the Euclidean norm of the covariates is bounded (e.g., our logistic regression example below).

Note that if we instead assume that ρ is $(\gamma, 3)$-self-concordant, then Lemma 9 implies that $\rho(y_i, x_i^T\theta)$ is $(\gamma, 3)$-self-concordant. Then Lemma 8 implies that $\mathcal{L}_n$ is $\left(\gamma, \sqrt{\frac{n}{\max_i w(x_i)}}, 3\right)$-self-concordant. However, both the growth of the self-concordance parameter with $\sqrt{n}$ and the inverse dependence on the Mallows weights is problematic.

Linear regression with Schweppe weights: We now consider linear regression loss functions of the form

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho((y_i - x_i^T\theta)v(x_i)), \tag{10}$$

where $\rho : \mathbb{R} \to \mathbb{R}$. In particular, we would consider weight functions $v : \mathbb{R}^p \mapsto \mathbb{R}_{\geq 0}$ such that $\|v(x)x\|_2 \leq C < \infty$. This objective function corresponds to a robust regression loss with Schweppe weights [Hampel et al., 1986, Ch. 6.3]. It is an alternative to the Mallows loss function, and bypasses the problem of requiring covariates to be bounded in order to obtain a multivariate self-concordance parameter which agrees with the univariate self-concordance parameter up to a scale factor.

As in the case of Mallows weighting, we can use the lemmas in Appendix C to derive the following result:

Lemma 3. *Suppose the Schweppe weights satisfy* $\|v(x)x\|_2 \leq C < \infty$.

(i) Suppose ρ *is* $(\gamma, 2)$*-self-concordant. Then the Schweppe loss* (10) *is* $(\gamma C, 2)$*-self-concordant.*

(ii) Suppose ρ *is* $(\gamma, 3)$*-self-concordant. Also suppose* $\|\rho''\|_\infty < \infty$*. Then the Schweppe loss* (10) *is* $(\gamma C\|\rho''\|_\infty^{1/2}, 2)$*-self-concordant.*

Proof. For (i), note that Lemma 9 implies that $\rho((y_i - x_i^T\theta)v(x_i))$ is $(\gamma\|v(x_i)x_i\|_2, 2)$-self-concordant. Then Lemma 8 implies that $\mathcal{L}_n(\theta)$ is $(\gamma \max_i \|v(x_i)x_i\|_2, 2)$-self-concordant.

For (ii), note that Lemma 9 implies that $\rho((y_i - x_i^T\theta)v(x_i))$ is $(\gamma, 3)$-self-concordant. Note that by assumption, each function $\rho'((y_i - x_i^T\theta)v(x_i))v(x_i)x_i$ is L-Lipschitz continuous with $L = C^2\|\rho''\|_\infty$. Hence, Lemma 10 then implies that $\rho((y_i - x_i^T\theta)v(x_i))$ is also $(\gamma L^{1/2}, 2)$-self-concordant. Hence, Lemma 8 implies that $\mathcal{L}_n(\theta)$ is $(\gamma L^{1/2}, 2)$-self-concordant. □

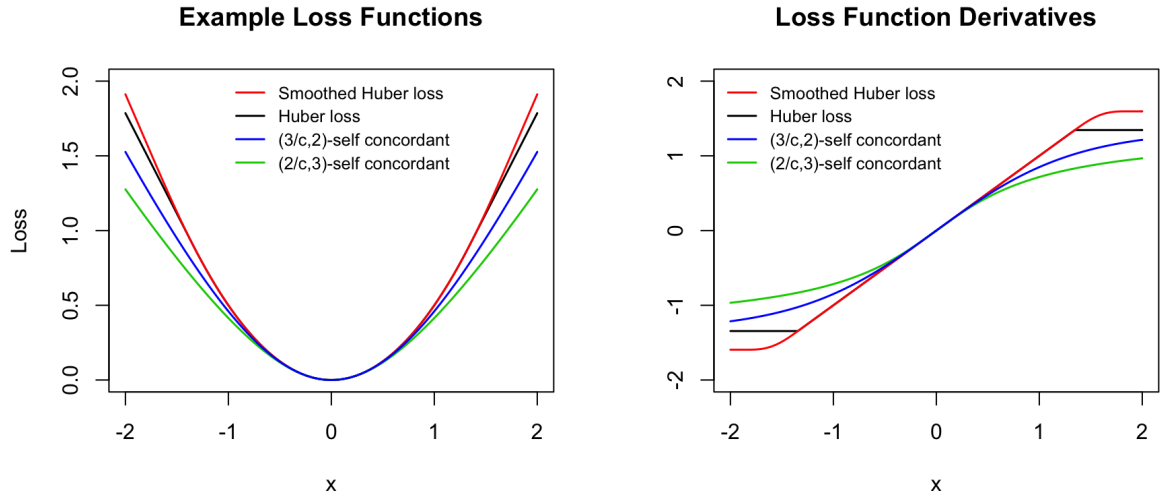


Figure 4: Univariate loss functions corresponding to variants of the Huber loss. For more details, see Examples 1 and 2.

Example 1 (Smoothed Huber loss)**.** *An intuitive way to circumvent the fact that Huber's* ψ_c*-function is not differentiable at* $\{-c, c\}$ *is to smooth out the corners where differentiability is violated.*

In particular, one could consider the following smooth approximation to Huber's score function:

$$\psi_{c,h}(t) = \begin{cases} t & \text{for } |t| \le c, \\ P_4(t) & \text{for } c < |t| < c+h, \\ c' & \text{for } |t| \ge c+h, \end{cases}$$

where $c' > c$ and $P_4(t)$ is a piecewise fourth-degree polynomial, ensuring that $\psi_{c,h(t)}$ is twice-differentiable everywhere. Hence, by construction,

$$\psi'_{c,h}(t) = \begin{cases} 1 & \text{for } |t| \le c, \\ P'_4(t) & \text{for } c < |t| < c+h, \\ 0 & \text{for } |t| \ge c+h \end{cases} \quad \text{and} \quad \psi''_{c,h}(t) = \begin{cases} 0 & \text{for } |t| \le c \text{ and } |t| \ge c+h, \\ P''_4(t) & \text{for } c < |t| < c+h. \end{cases}$$

This smoothed Huber function and related ideas have been discussed in the robust statistics literature [Fraiman et al., 2001, Hampel et al., 2011]. The proposed $\psi_{c,h}(t)$ can be used to define a Mallows loss function that meets the conditions of Theorem 3. Indeed, it is easy to see that for all $t_1, t_2 \in [-c, c]$ and some $\bar{t}$ between t_1 and t_2, we have

$$\big(\psi_{c,h}(t_1) - \psi_{c,h}(t_2)\big)(t_1 - t_2) \ge \psi'_{c,h}(\bar{t})(t_1 - t_2)^2 = (t_1 - t_2)^2,$$

which implies that $\psi_{c,h}(t)$ is $\frac{1}{2}$-locally strongly convex. This in turn can be used to establish that the objective function satisfies local strong convexity, with high probability, in a straightforward fashion. Clearly, the objective function is also $\frac{1}{2}$-smooth, since $|\psi'_{c,h}(t)| \le 1$ for all t. However, this smoothed Huber loss is not (γ, ν)-self-concordant, as we cannot find a constant γ such that $|\psi''_{c,h}(t)| \le 2\gamma\{\psi'_{c,h}(t)\}^{\nu/2}$ for all $|t| \in (c, c+h)$.

Example 2 (Self-concordant Huber regression with Schweppe weights)**.** *In the setting of Lemma 3(i), we can choose the univariate $(3/c, 2)$-self-concordant Huber functions*

$$\phi_c(t) = c^2 \log(\cosh(t/c)) \quad \text{and} \quad \phi_c(t) = c^2\big(\sqrt{1 + (t/c)^2} - 1\big).$$

Alternatively, we could consider the following $(2/c, 3)$-self-concordant Huber loss which, for $t \neq 0$, is defined as

$$\phi_c(t) = \frac{c^2}{2}\left[\sqrt{1 + 4(t/c)^2} - 1 + \log\left(\frac{\sqrt{1 + 4(t/c)^2} - 1}{2(t/c)^2}\right)\right].$$

The corresponding values at $t = 0$ are defined to be $\phi_c(0) = 0$. Figure 4 illustrates the various proposals for smoothed and self-concordant versions of the Huber loss, as well as their derivatives. We note that these univariate self-concordant Huber losses were discussed by Ostrovskii and Bach [2021] in the context of a finite-sample theory for M-estimators. These functions can also be shown to be locally strongly convex, since for all $t_1, t_2 \in [-c, c]$, we have

$$(\phi(t_1) - \phi(t_2))(t_1 - t_2) \ge \min_{-c \le t \le c} \phi'_c(t)(t_1 - t_2)^2$$

and $\min_{-c \le t \le c} \phi'_c(t) > 0$. Assuming a linear model $y_i = x_i^\top \theta_0 + u_i$ for $i = 1, \dots, n$, and assuming that $\|v(x)x\|_2 \le C$ and $|v(x)| \le 1$ for all $x \in \mathbb{R}^p$, it is also easy to check that the self-concordant Huber regression loss with Schweppe weights is also locally strongly convex with parameter $\tau_1 = \frac{1}{2}\lambda_{\min}\big(\frac{1}{n}\sum_{i=1}^n v^2(x_i) x_i x_i^\top \psi'_c(|u_i| + Cr)\big)$ within $\mathcal{B}_r(\theta_0)$. We note that this estimate holds for a fixed data set. One can give a deterministic τ_1 that holds with high probability by the argument of Proposition 2 in Loh [2017]. We also note that this loss function is τ_2-smooth with $\tau_2 = \frac{1}{2}C^2\psi'_c(0)$.

Example 3 (Logistic regression). *Assuming that* $\max_i \|x_i\|_2 \le C$, *Bach [2010] showed that the logistic regression loss is* $(C, 2)$*-self-concordant. Indeed, defining* $\phi(t) = \log(e^{-t/2} + e^{t/2})$, *we see that* $\phi'''(t) \le |\phi''(t)|$, *so* ϕ *is* $(1, 2)$*-self-concordant. Hence, by Lemma 2, the logistic regression loss is* $(C, 2)$*-self-concordant. It is also clear that* $\phi(t)$ *is locally strongly convex, since for all* $t_1, t_2 \in [-c, c]$, *we have*

$$(\phi(t_1) - \phi(t_2))(t_1 - t_2) \ge \left(\min_{-r \le t \le r} \phi'(t)\right)(t_1 - t_2)^2 = \frac{e^r}{(1+e^r)^2}(t_1 - t_2)^2.$$

Assuming again that $\max_i \|x_i\|_2 \le C$, *we can then show that the logistic regression loss is locally strongly convex with parameter* $\tau_1 = \frac{1}{2}\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\frac{\exp\{C(\|\theta_0\|_2 + r)\}}{[1+\exp\{C(\|\theta_0\|_2+r)\}]^2}$ *within* $\mathcal{B}_r(\theta_0)$. *It is also easy to see that see that the logistic regression loss is* τ_2*-smooth where* $\tau_2 = \frac{1}{4}\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)$.

It is not very obvious to construct a (γ, ν)*-self-concordant loss for binary regression such that the parameter* γ *does not depend on the data. Indeed, a Mallows-type estimator such as the one considered in Cantoni and Ronchetti [2001] runs into the problem discussed in the remark following Lemma 2. Furthermore, the Schweppe estimator for logistic regression proposed by Künsch et al. [1989] is not twice-differentiable.*

4.4 Practical considerations

Theorems 3 and 4 prove that the noisy Newton algorithm with $\eta = 1$ converges quadratically to a nearly-optimal neighborhood of the target parameter $\hat{\theta}$ when the starting value lies in a suitable neighborhood of the solution. Figure 5 illustrates this improved performance of (noisy) Newton's method relative to (noisy) gradient descent.

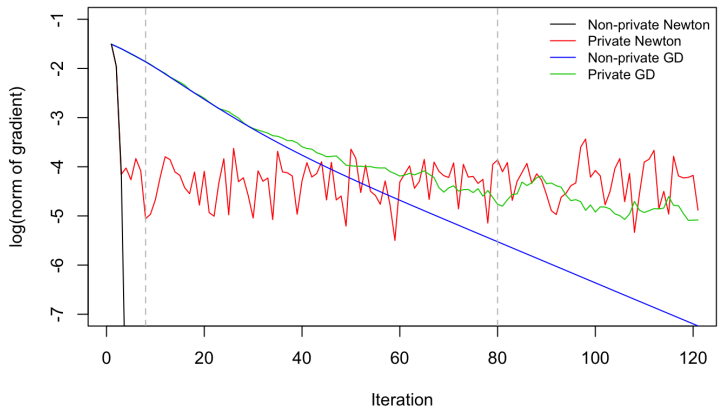


Figure 5: Noisy Newton's method vs. gradient descent. The vertical axis records the norm of the gradient trajectory, plotted on a log scale.

In this example, the noisy Newton algorithm is calibrated to achieve 2-GDP in 8 iterations, while noisy gradient descent is calibrated to achieve 2-GDP in 80 iterations. This is meant to reflect the fact that Newton's method tends to converge faster, so in practice, we would schedule fewer iterations of Newton's method compared to gradient descent. To retain the 2-GDP privacy guarantees, we would terminate the algorithms at the gray lines on the plot (iteration 8 for Newton and iteration 80 for gradient descent), but for purpose of illustration, we have forced both algorithms to continue, with the same amount of noise added in the extra steps.

4.4.1 Convergence issues

Since the starting value condition required for the quadratic convergence of Newton's method cannot typically be guaranteed a priori, one must consider two regimes: (i) damped Newton updates with $\eta < 1$, and (ii) pure noisy Newton updates with $\eta = 1$. Indeed, even in general non-private settings, Newton's method with stepsize $\eta = 1$ may or may not converge depending on the initial point chosen. Figure 6 displays the results of simulations using the damped version of Newton's method. The data in this simulation, $\{(x_i, y_i)\}_{i=1}^{1000}$, are generated according to the model $y_i = x_i^T\beta + \epsilon_i$, where $\beta = (1,1,1,1)^T$, $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, 2^2)$, and the covariate vectors are given by $x_i = (1, z_i)^T$, where $z_i \overset{i.i.d.}{\sim} N(0, 2^2 \cdot \mathbb{I}_3)$. At each stepsize, 500 repetitions were performed, each beginning at the initial point $\beta_{init} = (0,0,0,0)^T$, $\sigma_{init} = 1$, and tuned for a 2-GDP privacy guarantee.

For the same sample size, initial point, and true parameter values, we see that using damped Newton with a fixed stepsize η tends to help with the divergence problem: With all else equal, the proportion of trials leading to divergent iterates tends to decrease as the stepsize shrinks toward zero, although at very small stepsizes, the algorithm may fail to converge within the budgeted number of iterations.

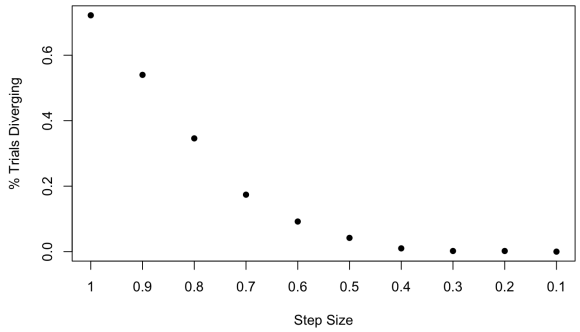


Figure 6: Divergence of noisy damped Newton algorithm. When $\eta = 1$ (pure Newton), the algorithm may not converge if the initial point does not lie close enough to the global optimum $\hat{\theta}$. Using smaller values of η remedies this problem.

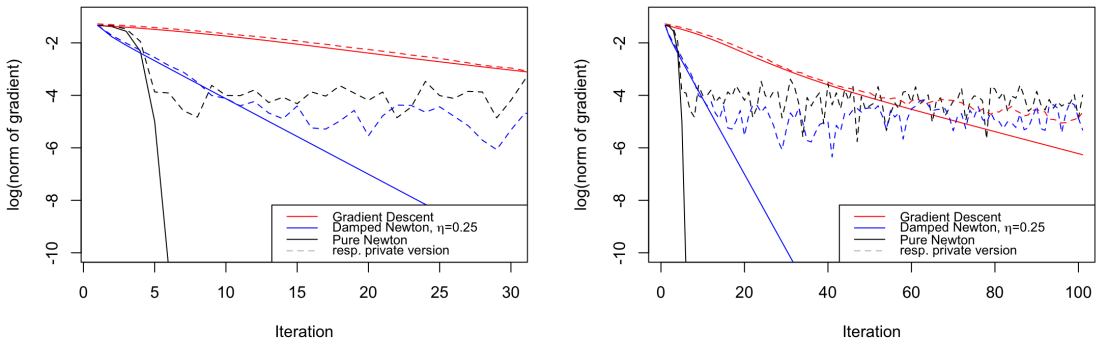


Figure 7: Trajectories of noisy gradient descent, pure Newton, and damped Newton algorithms.

Even though taking smaller Newton stepsizes guarantees the global convergence of the algorithm, it only guarantees linear convergence, as illustrated in Figure 7. A popular strategy employed to in order to obtain a provably quadratic convergent Newton algorithm is to rely on a damped version of Newton's method, in which the update is scaled by an adaptively-chosen stepsize via backtracking line search [Boyd and Vandenberghe, 2004, Ch. 9.5.2]. However, selecting this stepsize requires evaluating the loss function, and privately releasing the value of the loss function via the Gaussian mechanism from Theorem 1 would require a finite global sensitivity in our setting.

4.4.2 A private alternative to backtracking line search

As we prefer to avoid assuming explicit bounds on the data or objective function, we depart from the main ideas behind backtracking. Instead, we propose to privately estimate the magnitude of successive Newton steps in order to determine whether the starting value conditions of Theorems 3 and 4 are met. We provide a few details here.

Locally strongly convex case: This is particularly challenging if one relies on local strong convexity theory, as one needs to check whether $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \leq \frac{\tau_1^2}{L}$. In practice, one would check whether a noisy gradient meets this inequality, but one also needs to explicitly evaluate L and τ_1. We discuss this issue in the context of a Mallows estimator for linear regression. The Lipschitz constant L can be estimated as follows: The intermediate value theorem and the Cauchy-Schwarz inequality yield the upper bound

$$\begin{aligned}\|\nabla^2 \mathcal{L}_n(\theta_1) - \nabla^2 \mathcal{L}_n(\theta_2)\|_2 &= \left\| \frac{1}{n} \sum_{i=1}^n x_i x_i^\top w(x_i) \{\psi'_{c,h}(y_i - x_i^\top \theta_1) - \psi'_{c,h}(y_i - x_i^\top \theta_2)\} \right\|_2 \\ &\leq \sup_r |\psi''_{c,h}(r)| \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top w(x_i) \|x_i\|_2 \right) \|\theta_1 - \theta_2\|_2.\end{aligned}$$

One can obtain a differentially private estimate of $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top w(x_i)\|x_i\|_2$ with the matrix Gaussian mechanism with an appropriate choice of $w(x_i)$, e.g., defining $w(x_i) = \max\left\{1, \frac{1}{\|x_i\|_2^3}\right\}$. This automatically leads to a private estimate of $\lambda_{\max}(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top w(x_i)\|x_i\|_2)$, and hence of L. Estimating the LSC constant τ_1 is trickier: Since $\psi'_{c,h}(r) \geq \mathbb{1}(|r| \leq c)$, we also have

$$\lambda_{\min}\left(\nabla^2 \mathcal{L}_n(\theta)\right) \geq \lambda_{\min}\left(\frac{1}{n} \sum_{i=1}^n w(x_i) x_i x_i^\top \mathbb{1}(|y_i - x_i^\top \theta| \leq c) \right) = l(\theta).$$

In order to obtain a differentially private estimate of $\frac{\tau_1^2}{L}$, one can try to approximate τ_1 by lower-bounding $l(\theta)$ over $\mathcal{B}_r(\theta^{(k_0)})$ for some $k_0 < K$. One could also rely on the heuristic that $l(\theta^{(0)})$ is typically much smaller for a bad starting value than $l(\theta^{(k)})$ after a few iterations, provided $l(\theta^{(0)}) \neq 0$.

Self-concordant case: In the case of a $(\gamma, 2)$-self-concordant loss function such as in Examples 2 and 3 above, it may also be advantageous to begin Newton's method with a fixed stepsize η, then switch to pure Newton ($\eta = 1$) at an iteration k for which $\lambda_{\min}^{-1/2}(\nabla^2 \mathcal{L}_n(\theta^{(k)}))\lambda(\theta^{(k)}) \leq \frac{1}{16\gamma}$. To preserve differential privacy, one may use private estimates of $\lambda(\theta^{(k)})$ and $\nabla^2 \mathcal{L}_n(\theta^{(k)})$. Since our method already employs private estimates of the gradient and Hessian at each iteration, privately computing the value of the Newton decrement and the minimal value of the Hessian come at no

additional cost with respect to the privacy budget. Figure 7 illustrates the two potential main benefits of implementing an adaptive damped noisy Newton algorithm relative to noisy gradient descent. Firstly, when the algorithms are initialized at a starting value lying outside the local strong convexity region, Newton's method converges linearly to this region. This is in stark contrast with noisy gradient descent, which can only approach the local strong convexity region at a very slow sub-linear rate. This slow rate of convergence is inherited from the slow sub-linear rate of convergence of gradient descent in the absence of strong convexity [Bubeck, 2015, Nesterov, 2018]. Secondly, once the noisy Newton iterates become sufficiently close to the global optimum, one can hope to detect this transition and obtain improved quadratic convergence rates by taking private pure Newton steps in successive iterations.

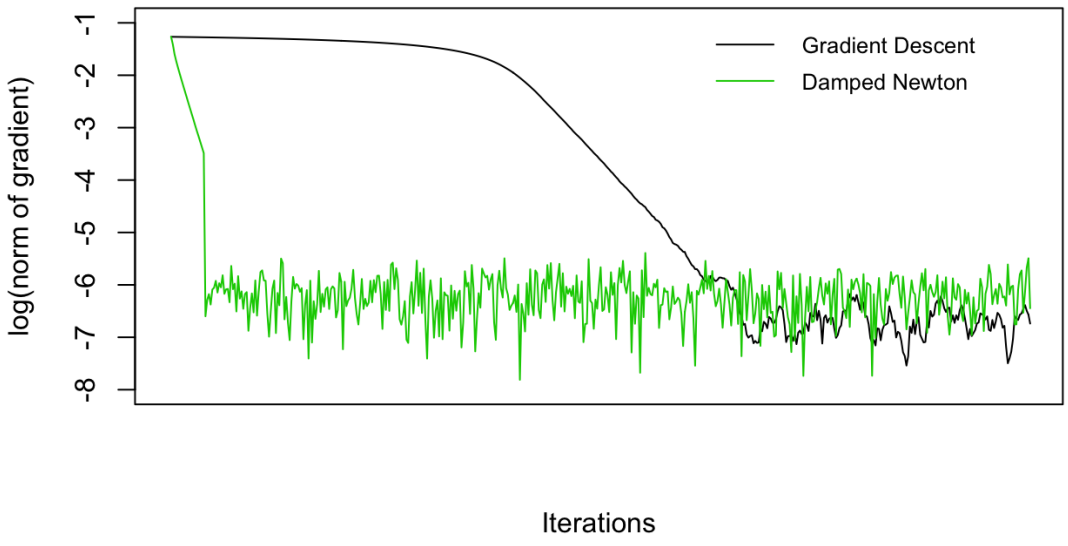


Figure 8: Trajectories of gradient descent and damped Newton algorithm. The damped Newton algorithm converges much faster in relation to gradient descent.

5 Inference via noisy variance estimators

We now show how to leverage our results on asymptotic normality of the private estimators $\theta^{(K)}$, obtained from either noisy gradient descent or noisy Newton's method, to construct confidence regions for θ_0. We will present two asymptotically valid construction based on these ideas.

5.1 Private sandwich formula

The most common approach for constructing confidence intervals for θ_0 using an M-estimator $\hat{\theta}$ is to use asymptotic theory to construct an asymptotic pivot. In particular, one typically relies on the fact that under regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, V(\theta_0)),$$

where

$$V(\theta_0) := M(\theta_0)^{-1} Q(\theta_0) M(\theta_0)^{-1},$$
$$M(\theta_0) := -\mathbb{E}_F[\dot{\Psi}(X, \theta_0)],$$
$$Q(\theta_0) := \mathbb{E}_F[\Psi(X, \theta_0)\Psi(X, \theta_0)^\top],$$

are the analogues of equations (3). The asymptotic variance can accordingly be estimated via the plug-in sandwich estimator $V_n(\hat{\theta}) = M_n(\hat{\theta})^{-1} Q_n(\hat{\theta}) M_n(\hat{\theta})^{-1}$, where $M_n(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n \dot{\Psi}(X_i, \hat{\theta})$ and $Q_n(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n \Psi(X_i, \hat{\theta})\Psi(Z_i, \hat{\theta})^\top$. Consistency of $\hat{\theta}$ and the continuous mapping theorem show that $V_n(\hat{\theta}) \to_p V(\theta_0)$, so Slutsky's theorem justifies the use of the $(1-\alpha)$-confidence intervals

$$\mathrm{CI}_{1-\alpha}(\theta_{0j}) = \left[\hat{\theta}_j - \frac{\sqrt{V_n(\hat{\theta})_{jj}}}{\sqrt{n}} z_{1-\alpha/2}, \hat{\theta}_j + \frac{\sqrt{V_n(\hat{\theta})_{jj}}}{\sqrt{n}} z_{1-\alpha/2}\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal. However, in the differential privacy setting, this plug-in construction cannot be applied directly, as neither $M_n(\theta^{(K)})$ nor $Q_n(\theta^{(K)})$ are differentially private. Instead, we use the approach discussed in Section 4.1 to construct differentially private analogues of these two matrices: Assuming Condition 2 holds, Lemma 1 suggests using the estimates

$$\tilde{M}_n(\theta^{(K)}) = M_n(\theta^{(K)}) + \frac{2\bar{B}}{\mu n} G_1 \quad \text{and} \quad \tilde{Q}_n(\theta^{(K)}) = Q_n(\theta^{(K)}) + \frac{2B^2}{\mu n} G_2,$$

where G_1 and G_2 are i.i.d. symmetric random matrices whose upper-triangular elements, including the diagonals, are i.i.d. standard normal. The differentially private matrices $\tilde{M}_n(\theta^{(K)})$ and $\tilde{Q}_n(\theta^{(K)})$ are not necessarily positive definite, but can be easily projected onto a cone of positive definite matrices $\{H : H \succeq \varepsilon I\}$ without paying a large statistical cost (cf. Remark 5). Equipped with the projected matrices

$$\tilde{M}_n(\theta^{(K)})_+ = \arg\min_{H \succeq \varepsilon I} \|H - \tilde{M}_n(\theta^{(K)})\|_2 \quad \text{and} \quad \tilde{Q}_n(\theta^{(K)})_+ = \arg\min_{H \succeq \varepsilon I} \|H - \tilde{Q}_n(\theta^{(K)})\|_2,$$

we can construct the differentially private sandwich estimator

$$\tilde{V}_n(\theta^{(K)}) = \tilde{M}_n(\theta^{(K)})_+^{-1} \tilde{Q}_n(\theta^{(K)})_+ \tilde{M}_n(\theta^{(K)})_+^{-1}. \tag{11}$$

We note that essentially the same idea was suggested by Wang et al. [2019] in the context of differentially private estimators computed via objective perturbation and output perturbation. Those techniques suffer some of the drawbacks mentioned in the introduction, and in particular need to assume bounded data.

The composition property of μ-GDP estimators along with the consistency of our noisy estimators $\theta^{(K)}$ easily imply the following result:

Proposition 3. *Assume the conditions of Theorems 2 or 3 hold. Then $\tilde{V}_n(\theta^{(K)})$ is $\sqrt{3}\mu$-GDP and $\tilde{V}_n(\theta^{(K)}) \to_p V(\theta_0)$.*

We can therefore release the following $\sqrt{3}\mu$-GDP $(1-\alpha)$-confidence intervals:

$$\mathrm{CI}^{priv}_{1-\alpha}(\theta_{0j}) = \left[\theta_j^{(K)} - \frac{\sqrt{\tilde{V}_n(\theta^{(K)})_{jj}}}{\sqrt{n}} z_{1-\alpha/2}, \theta_j^{(K)} + \frac{\sqrt{\tilde{V}_n(\theta^{(K)})_{jj}}}{\sqrt{n}} z_{1-\alpha/2}\right].$$

Although this construction is asymptotically valid, it tends to be too liberal in small samples. The correction proposed in the next subsection partially addresses this issue.

5.2 A finite-sample correction

We now discuss a small correction one can add to the noisy gradient descent and noisy Newton's method algorithms, which leads to better performance in practice.

5.2.1 Noisy gradient descent

The formula (11), by construction, underestimates the variance of $\theta^{(K)}$ in finite samples, as it fails to account for the additional variability introduced by the privacy-preserving random noise mechanism. In the case of noisy gradient descent, this can be mitigated by making the following correction to equation (11):

$$\hat{V}_n(\theta^{(K)}) = \tilde{V}_n(\theta^{(K)}) + \frac{8\eta^2 B^2 K}{n\mu^2} I. \tag{12}$$

The correction (12) is motivated by the behavior of the noisy iterates of our algorithms at convergence. Indeed, from the proof of Theorem 2, we know that with high probability, every iteration of the algorithm grows closer to the non-private solution $\hat{\theta}$, up to a privacy-preserving noisy neighborhood of radius $\frac{4\eta^2 B^2 K}{n\mu^2}$. Once this neighborhood is attained, the iterates approximately behave as a random walk around its boundary. Since we inject normally distributed noise with variance $\frac{4\eta^2 B^2 K}{n\mu^2}$, we see that $\theta^{(K)}$ will be approximately within a $\frac{8\eta^2 B^2 K}{n\mu^2}$-neighborhood of $\hat{\theta}$, which is precisely what we try to account for in equation (12). This corrected variance formula yields $(1-\alpha)$-confidence intervals of the form

$$\mathrm{CI}^{cor}_{1-\alpha}(\theta_{0j}) = \left[\theta_j^{(K)} \pm \sqrt{\frac{\tilde{V}_n(\theta^{(K)})_{jj}}{n} + 2\left(\frac{2\eta B\sqrt{K}}{n\mu}\right)^2} z_{1-\alpha/2}\right].$$

We demonstrate the practical impact of this correction in Figure 9 below, using noisy gradient descent on simulated data to estimate the parameters of a multivariate linear regression model. Data for this simulation were drawn from the same model as in Section 3.2.1. We see that the empirical coverage of the 95% confidence intervals (for one particular regression coefficient) are indeed closer to the nominal 95% level when this correction is included, and the effect is more pronounced for smaller sample sizes. Naturally, one can also construct corrected confidence regions given ellipsoids centered at $\theta^{(K)}$ and shaped according to equation (12). We illustrate this construction in Figure 10.

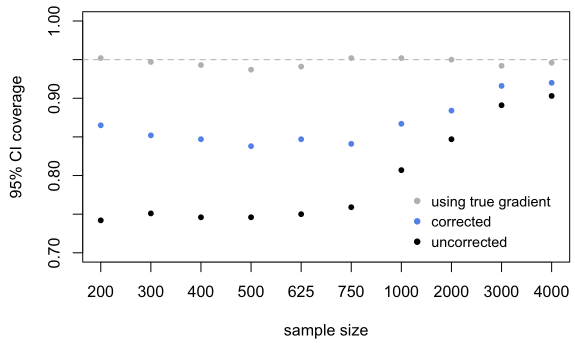


Figure 9: Empirical confidence interval coverage for confidence intervals constructed using noisy gradient descent estimators. The blue dots show the empirical coverage probabilities of the last iterate utilizing the noisy sandwich formula and the corrected formula. The gray points illustrate the coverage of the corrected formula computed using the first iterate whose gradient is smaller (in Euclidean norm) than the standard error of the noise added at each iteration. We note that this gradient uses a non-private stopping rule.

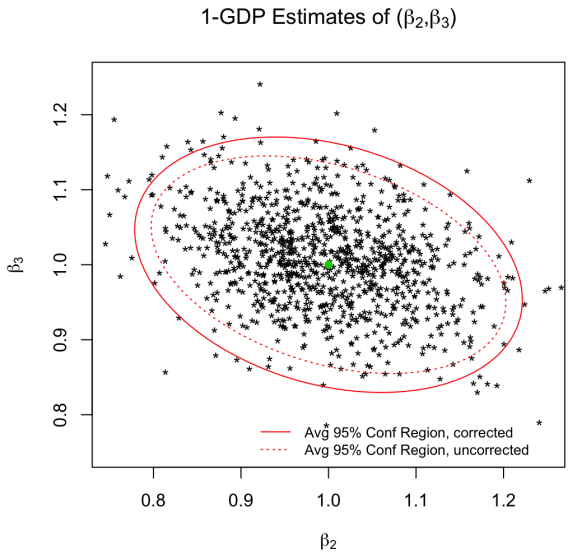


Figure 10: Two-dimensional coverage region for a pair of parameters, showing the tighter confidence region resulting from our finite-sample correction to the sandwich formula.

5.2.2 Noisy Newton's method

In Appendix E.1, we see that the noisy Newton's method update can be expressed as the sum of an ordinary Newton's method update plus a noise term. Specifically, the noise term is $\eta\tilde{N}_k$, where $\tilde{N}_k$ is an infinite sum given by equation (48). We wish to account for the additional variance introduced by this noise term. We propose to retain only the first term in $\tilde{N}_k$, as the other terms are higher-order terms. Thus, truncating the error term to $\eta\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})\right\}^{-1}\tilde{Z}_k$, the variance attributable to

this term is $\eta^2 \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \mathrm{var}(\tilde{Z}_k) \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1}$. Finally, to maintain the desired privacy guarantee, we substitute the noisy estimate of the Hessian at the current iterate for $\nabla^2 \mathcal{L}_n(\theta^{(k)})$. This gives an approximate correction to the variance of the form:

$$C_{Newton} \coloneqq \eta^2 \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)}) + \tilde{W}_k\right\}^{-1} \left(\frac{2B\sqrt{2K}}{\mu n}\right)^2 \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)}) + \tilde{W}_k\right\}^{-1}.$$

In the case of noisy Newton's method, the variance of $\theta^{(K)}$ can be approximated by making the following correction to equation (11):

$$\hat{V}_n(\theta^{(K)}) = \tilde{V}_n(\theta^{(K)}) + nC_{Newton}. \tag{13}$$

Analogously to the gradient descent case above, this corrected variance formula yields $(1-\alpha)$-confidence intervals of the form

$$\mathrm{CI}^{cor}_{1-\alpha}(\theta_{0j}) = \left[\theta_j^{(K)} \pm \sqrt{\frac{\tilde{V}_n(\theta^{(K)})_{jj}}{n} + (C_{Newton})_{jj}} z_{1-\alpha/2}\right].$$

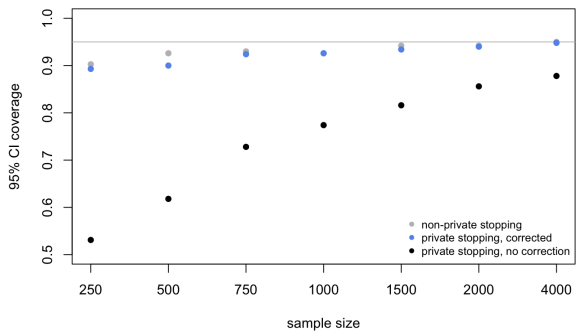


Figure 11: Empirical confidence interval coverage for confidence intervals constructed using noisy Newton's method estimators. The blue dots show the empirical coverage probabilities of the last iterate utilizing the noisy sandwich formula and the corrected formula, for an algorithm that terminates before the budgeted number of iterations if the noisy gradient is sufficiently small in magnitude. The gray dots illustrate the coverage of the corrected formula with a stopping rule based on the true gradient.

6 Numerical illustrations

In this section, we explore the performance of our algorithms on several real data sets.

6.1 Linear regression

To further explore the performance of our proposed methods, we consider fitting a linear regression model to a housing price data set utilized in Lei [2011]. The data consist of price, square footage, year of sale, and county for 348,189 homes sold in the San Francisco Bay Area from 2003 to 2006. We excluded all records with missing values, and combined several counties such that the number of county categories was reduced from 9 to 6. The remaining data set contains 286,537 records.

Additionally, price, square footage, and year were standardized by subtracting their medians and dividing by median absolute deviation. (Note that differentially private estimates of the median and MAD of these variables could also be returned. This comes at an additional cost in terms of privacy budget, but may be desirable for interpretability or prediction purposes.)

With this data set, we used noisy gradient descent with loss function as in equation (6) to estimate a linear regression model for predicting the home price from the remaining variables. The total privacy budget for estimation and corresponding inference was $\mu = 0.25$, and the number of iterations K was 100. The estimated coefficients of this regression model are in Table 1.

	Value	Std. Error	z value	p value
(Intercept)	-0.1835649	0.002174312	-84.42439	0
bsqft	0.6434390	0.001414192	454.98718	0
date	0.4578594	0.001687287	271.35829	0
Contra Costa	-0.2743499	0.003251559	-84.37488	0
MSS	0.8731473	0.004131656	211.33105	0
NS	-0.1559515	0.003765833	-41.41222	0
Santa Clara	0.2388637	0.003122972	76.48601	0
Solano	-0.6718810	0.003125814	-214.94595	0

Table 1: Coefficients estimated via noisy gradient descent ($\mu = 0.25$)

In this large data set, the inference procedures outlined in Section 5 indicate that all of the regression coefficients are statistically significant at a $\alpha = 0.05$ threshold. To obtain an approximate notion of performance on smaller data sets, we repeatedly applied noisy gradient descent to random subsamples of the original data set. In each implementation, the total privacy budget was $\mu = 0.25$, and the number of iterations K was 100.

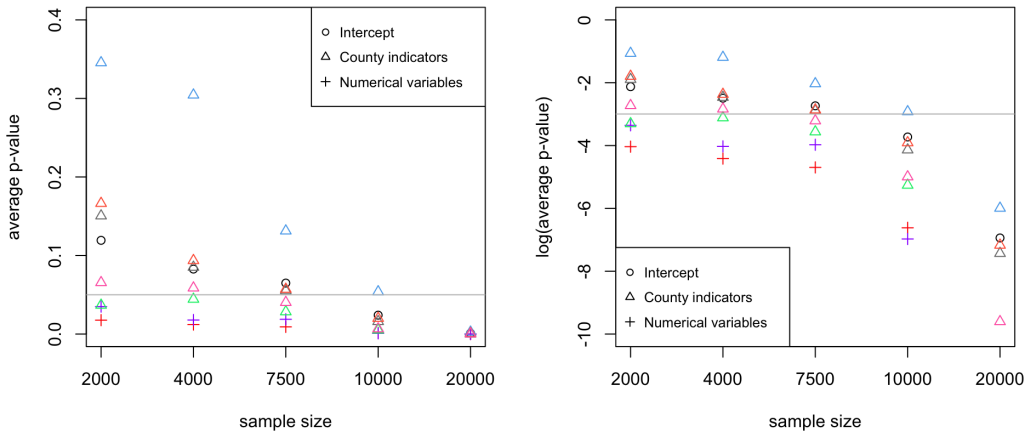


Figure 12: Noisy gradient descent applied to random subsamples of the housing data set.

In these subsamples, we see that for sample size $n = 2000$, only 3 of the 8 regression coefficients have p-values averaging less than 0.05 across 200 repetitions. As we increase the sample size to 10,000, 7 of the 8 coefficients have average p-values below that significance threshold, and by a sample size 20,000, all coefficients are detected as significant.

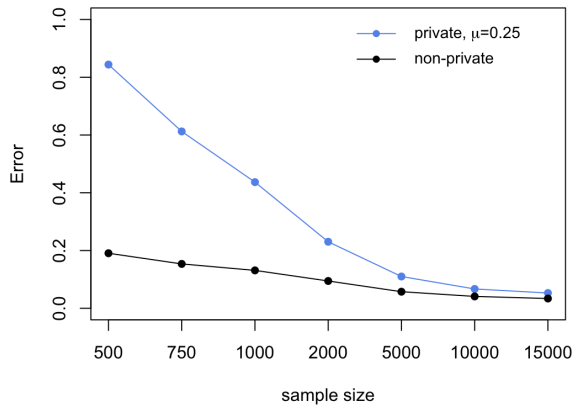


Figure 13: Subsampled estimates of $\mathbb{E}(\|\hat{\beta} - \beta\|_2)$.

Figure 13 demonstrates the error of private and non-private estimates of the parameter vector β calculated from subsamples of the housing data set. Error is calculated with respect to the non-private estimator obtained from the full data set and averaged over 400 repetitions at each sample size. The number of iterations K was set to 50 for sample size 500; 75 for sample sizes 750, 1000, and 2000; and 100 for all larger sample sizes. The privacy mechanism has a relatively large impact on the estimation error for β in small samples, but the gap between the corresponding private and non-private estimators closes with increasing sample size.

6.2 Logistic regression

In this example, we fit a logistic regression model to a data set from Moro et al. [2014], describing customer relationships with a bank in Portugal. To fit this model, we used noisy gradient descent to minimize a version of cross-entropy loss modified to include Mallows-style weights:

$$\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i \log\left(\frac{1}{1+\exp(x_i^\top\beta)}\right) + (1-y_i)\log\left(\frac{\exp(x_i^\top\beta)}{1+\exp(x_i^\top\beta)}\right)\right) w(x_i), \tag{14}$$

where $w(x_i) = \min\left(1, \frac{25}{\|x_i\|_2^2}\right)$.

This data set contains 45,211 records consisting of customer attributes such as age, job type, and types of business conducted with the bank. The response is whether the customer subscribed to a term deposit. For preprocessing, numeric covariates were standardized and categorical covariates were converted to one-hot encoding (i.e., representing a categorical variable with k levels, with $k-1$ binary indicator variables). One covariate (days since previous marketing contact with customer) was excluded as it was undefined for over 80% of observations. After this preprocessing, the data set had 41 covariates.

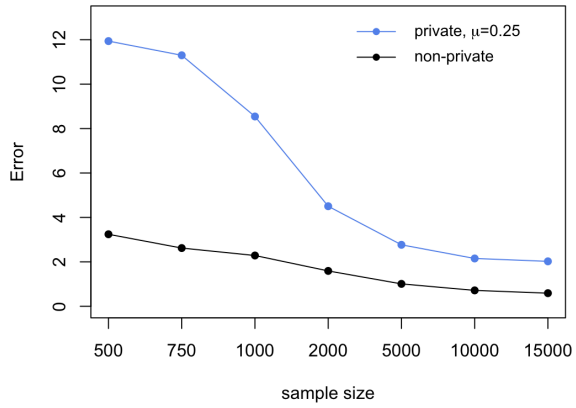


Figure 14: Subsampled estimates of $\mathbb{E}(\|\hat{\beta} - \beta\|_2)$.

Figure 14 shows the error of private and non-private estimates of the parameter vector β. As in the previous section, error is calculated with respect to the non-private estimator obtained from the full data set. Results are averaged over 200 repetitions at each sample size. The number of iterations K was set to 50 for sample size 500; 75 for sample sizes 750, 1000, and 2000; and 100 for all larger sample sizes.

7 Discussion

In this work, we have studied the theoretical properties of noisy versions of gradient descent and Newton's method for computing general differentially private M-estimators. We have analyzed the statistical properties of the iterates of these algorithms and provided a general approach for computing differentially private confidence regions based on asymptotic pivots.

Our noisy optimization theory shows that the convergence rates of our private M-estimators are nearly optimal, and many well-known convex optimization results are naturally generalized by our framework. In particular, the noisy optimizers considered in this work preserve the same rates of convergence as their standard non-noisy counterparts in the sense that they converge in the same order of iterations to a neighborhood of the solution to the non-private objective function minimization problem. Our noisy algorithms exhibit several noteworthy distinct features. By construction, the iterates cannot approach the non-private M-estimators beyond a privacy-preserving neighborhood proportional to the noise injected at each step of the algorithms. They instead bounce around this neighborhood once it is attained. Furthermore, the numbers of iterations has to be scheduled in advance and directly impacts the noise added to each step of the algorithms; the more steps, the more data queries, and hence the larger the required privacy-inducing noise.

Our analysis highlights the statistical importance of (local) strong convexity, since it justifies scheduling only $O(\log n)$ noisy iterates, which in turn leads to the nearly-optimal statistical cost of privacy of the order $O\left(\frac{p \log n}{n\mu^2}\right)$. Without strong convexity, standard gradient-based algorithms are known to require $O(\sqrt{n})$ iterations to achieve an optimization error comparable to the statistical parametric rate $O\left(\sqrt{\frac{p}{n}}\right)$. This would be quite damaging for differentially private M-estimators, as they would lead to a statistical cost of $O\left(\sqrt{\frac{p}{n}}\frac{1}{\mu^2}\right)$, which would in fact be the dominating term driving the asymptotic efficiency of the resulting estimators.

We have also introduced a general technique for computing confidence regions based on noisy estimates of the asymptotic variance of the M-estimator. Since any noisy iterate has an additional built-in noise that is not captured by the usual asymptotic variance, we have proposed simple corrections that account for an extra noise term. This approach significantly mitigates the inevitable systematic underestimation of the variance that one has to incur by only relying on the usual sandwich formula.

Two natural extensions of our work would be to study noisy stochastic gradient descent algorithms and consider high-dimensional penalized M-estimators with noisy proximal methods. While these ideas have been explored in the literature, previous analyses share many of the limitations of the existing literature of noisy gradient descent, i.e., they rely on (restricted) strong convexity, assume bounded data and/or bounded parameter spaces, or employ truncation ideas [Bassily et al., 2014, Cai et al., 2019, 2020, Dong et al., 2021, Jain and Thakurta, 2014, Talwar et al., 2015, Wang et al., 2017a]. We believe that some of the techniques used in this paper may bypass the aforementioned drawbacks and could also be used in conjunction with restricted local strong convexity and general penalty functions, as in Loh and Wainwright [2015] and Loh [2017].

We finally point out that we do not address the local differential privacy framework of Duchi et al. [2018], where one also has to filter the data when it is collected, in the absence of a trusted curator. Noisy gradient descent has been one of the main general algorithms proposed in that setting and it would be very interesting to explore whether our techniques can lead to better understanding of the optimization problems tackled there.

A Concentration inequalities

The following two lemmas will be instrumental in the analysis of our noisy algorithm.

Lemma 4. *Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with variance proxy σ^2. For any $\alpha > 0$, with probability at least $1 - \alpha$,*

$$\|X\|_2 \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\alpha)}.$$

Proof. This result can be found in [Rigollet and Hütter, 2017, Theorem 1.19]. □

Lemma 5. *Let W be a symmetric $p \times p$ random matrix whose upper triangular elements, including the diagonal, are i.i.d. $N(0,1)$. For any $\alpha > 0$, with probability at least $1 - \alpha$,*

$$\|W\|_2 \leq \sqrt{2p\log(2p/\alpha)}.$$

Proof. Letting E_{jk} denote the matrix which the value 1 in the $(j,k)^{\text{th}}$ component and 0 everywhere else, we see that

$$W = \sum_{1 \leq j \leq k \leq p} Z_{jk}(E_{jk} + E_{kj} - \mathbb{1}_{\{j=k\}}E_{jj}),$$

where $\{Z_{jk}\}_{1 \leq j,k \leq p}$ is an i.i.d. sequence of standard normal variables. It follows from [Tropp, 2015, Theorem 4.1.1] that with probability at least $1 - \alpha$,

$$\|W\|_2 \leq \sqrt{2v(Z)\log(2p/\alpha)},$$

where

$$v(Z) = \left\| \sum_{1 \leq j \leq k \leq p} (E_{jk} + E_{kj} - \mathbb{1}_{\{j=k\}}E_{jj})^2 \right\|_2 = \|pI\|_2 = p.$$

□

B Convex analysis

The following results are standard, and proofs can be found in Boyd and Vandenberghe [2004] or Bubeck [2015].

Lemma 6. *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is strongly convex with parameter τ_1, i.e.,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \tau_1 \|x - y\|_2^2, \qquad \forall x, y \in \mathbb{R}^p.$$

Then

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 2\tau_1 \|x - y\|_2^2, \qquad \forall x, y \in \mathbb{R}^p,$$

and if f is twice-differentiable, then

$$\nabla^2 f(x) \succeq 2\tau_1 I, \qquad \forall x \in \mathbb{R}^p.$$

Lemma 7. *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is convex and τ_2-smooth, i.e.,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tau_2 \|x - y\|_2^2, \qquad \forall x, y \in \mathbb{R}^p.$$

Then

$$\frac{1}{2\tau_2} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle, \qquad \forall x, y \in \mathbb{R}^p,$$

and

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq 2\tau_2 \|x - y\|_2, \qquad \forall x, y \in \mathbb{R}^p.$$

If f is twice-differentiable, then

$$\nabla^2 f(x) \preceq 2\tau_2 I, \qquad \forall x \in \mathbb{R}^p.$$

C Results about self-concordant functions

In this appendix, we state several useful results about generalized self-concordant functions.

C.1 Generalized self-concordant functions

The following results are taken from Sun and Tran-Dinh [2019].

Lemma 8 (Proposition 1 from Sun and Tran-Dinh [2019])**.** *Suppose f_i are (γ_i, ν)-self-concordant functions for $1 \leq i \leq n$, where $\gamma_i \geq 0$ and $\nu \geq 2$. The function $f(x) := \sum_{i=1}^n \beta_i f_i(x)$, for $\beta_i > 0$, is (γ, ν)-self-concordant with*

$$\gamma = \max_{1 \leq i \leq n} \left\{ \beta_i^{1-\nu/2} \gamma_i \right\}.$$

Lemma 9 (Proposition 2 from Sun and Tran-Dinh [2019])**.** *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is (γ, ν)-self-concordant with $\nu \in (0, 3]$, and consider any affine function $x \mapsto Ax + b$. Then $g(x) = f(Ax + b)$ is $(\gamma \|A\|_2^{3-\nu}, \nu)$-self-concordant.*

Lemma 10 (Proposition 4 from Sun and Tran-Dinh [2019])**.** *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is (γ, ν)-self-concordant with $\nu \geq 2$, and ∇f is Lipschitz continuous with Lipschitz constant $L \geq 0$ in ℓ_2-norm. Then f is $(L^{\nu/2-1}\gamma, 2)$-self-concordant.*

Lemma 11 (Proposition 8 from Sun and Tran-Dinh [2019]). *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is $(\gamma, 2)$-self-concordant. For any $x, y \in \mathbb{R}^p$, we have*

$$\exp\left(-\gamma\|x-y\|_2\right)\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \exp\left(\gamma\|x-y\|_2\right)\nabla^2 f(x).$$

Lemma 12 (Lemma 2 of Sun and Tran-Dinh [2019]). *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is $(\gamma, 2)$-self-concordant. For $x, y \in \mathbb{R}^p$, the matrix $H(x, y)$ defined by*

$$H(x,y) := \nabla^2 f(x)^{-1/2}\left(\int_0^1 \left(\nabla^2 f(x + t(y-x)) - \nabla^2 f(x)\right) dt\right)\nabla^2 f(x)^{-1/2}$$

satisfies

$$\|H(x,y)\|_2 \le \left(\frac{3}{2} + \frac{\gamma\|x-y\|_2}{3}\right)\gamma\|x-y\|_2 \exp\left(\gamma\|x-y\|_2\right).$$

Lemma 13 (Proposition 9 of Sun and Tran-Dinh [2019]). *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is $(\gamma, 2)$-self-concordant. For any $x, y \in \mathbb{R}^p$, we have*

$$\frac{1-\exp(-\gamma\|y-x\|_2)}{\gamma\|y-x\|_2^2}\|y-x\|^2_{\nabla^2 f(y)} \le \langle \nabla f(y) - \nabla f(x), y - x\rangle.$$

Furthermore, by the Cauchy-Schwarz inequality, the right-hand expression is upper-bounded by $\|\nabla f(y) - \nabla f(x)\|_{\nabla^2 f(y)^{-1}}\|y-x\|_{\nabla^2 f(y)}$, so

$$\frac{1-\exp(-\gamma\|y-x\|_2)}{\gamma\|y-x\|_2^2}\|y-x\|_{\nabla^2 f(y)} \le \|\nabla f(y) - \nabla f(x)\|_{\nabla^2 f(y)^{-1}}.$$

Lemma 14 (Proposition 10 of Sun and Tran-Dinh [2019]). *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is $(\gamma, 2)$-self-concordant. For any $x, y \in \mathbb{R}^p$, we have*

$$\omega(-\gamma\|x-y\|_2)\cdot\|x-y\|^2_{\nabla^2 f(x)} \le f(y) - f(x) - \langle\nabla f(x), y-x\rangle \le \omega(\gamma\|x-y\|_2)\cdot\|x-y\|^2_{\nabla^2 f(x)},$$

where $\omega(t) := \frac{\exp(t)-t-1}{t^2}$.

C.2 Stability of $(\gamma, 2)$-self-concordant functions

The following Hessian stability condition was introduced in Karimireddy et al. [2018], where it was used to prove global convergence of the iterates in (non-noisy) Newton's method for generalized self-concordant functions.

Condition 4 (Hessian stability). *Suppose $\Theta_0 \subseteq \mathbb{R}^p$. For any $\theta_1, \theta_2 \in \Theta_0$ and $\theta_1 \neq \theta_2$, assume $\|\theta_1 - \theta_2\|_{\nabla^2\mathcal{L}_n(\theta_1)} > 0$ and there exists a constant $\tau_0 \ge 1$ such that*

$$\tau_0 := \sup_{\theta_1,\theta_2\in\Theta_0} \frac{\|\theta_1-\theta_2\|^2_{\nabla^2\mathcal{L}_n(\theta_2)}}{\|\theta_1-\theta_2\|^2_{\nabla^2\mathcal{L}_n(\theta_1)}} < \infty.$$

In practice, we will take $\Theta_0 = \mathcal{B}_R(\hat{\theta})$, where R is a suitably chosen radius such that $\{\theta^{(k)}\} \subseteq \Theta_0$. Condition 4 is a stability assumption that allows us to show global convergence results without assuming local strong convexity. An important consequence is that it implies upper and lower bounds that are similar to local strong convexity and smoothness, as we can see in the following lemma:

Lemma 15. *Given Condition 4, for any $\theta_1, \theta_2 \in \Theta_0$, we have the following upper and lower bounds:*

$$\mathcal{L}_n(\theta_1) \le \mathcal{L}_n(\theta_2) + \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \frac{\tau_0}{2} \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)}, \tag{15}$$

$$\mathcal{L}_n(\theta_1) \ge \mathcal{L}_n(\theta_2) + \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \frac{1}{2\tau_0} \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)}. \tag{16}$$

Proof. We note that this corresponds to Lemma 2 in Karimireddy et al. [2018], but their arguments have a mistake which we have corrected here. A second-order Taylor expansion shows that

$$\mathcal{L}_n(\theta_1) = \mathcal{L}_n(\theta_1) + \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \frac{1}{2} \|\theta_1 - \theta_2\|^2_{\int_0^1 \nabla^2 \mathcal{L}_n(t\theta_1 + (1-t)\theta_2) \mathrm{d}t}. \tag{17}$$

Let $\bar{\theta}_t = t\theta_1 + (1-t)\theta_2$, and note that by convexity of $\mathcal{L}_n$, we have $\bar{\theta}_t \in \Theta_0$, for all $t \in [0,1]$. Therefore, Condition 4 ensures that

$$t^2 \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\bar{\theta}_t)} = \|\bar{\theta}_t - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\bar{\theta}_t)} \le \tau_0 \|\bar{\theta}_t - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)} = \tau_0 t^2 \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)},$$

so

$$\|\theta_1 - \theta_2\|^2_{\int_0^1 \nabla^2 \mathcal{L}_n(t\theta_1 + (1-t)\theta_2) \mathrm{d}t} \le \sup_{t \in [0,1]} \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\bar{\theta}_t)} \le \tau_0 \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)}. \tag{18}$$

Combining inequalities (17) and (18) yield the desired upper bound (15). An analogous argument establishes the lower bound. Indeed, Condition 4 also ensures that

$$\tau_0 t^2 \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\bar{\theta}_t)} = \tau_0 \|\bar{\theta}_t - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\bar{\theta}_t)} \ge \|\bar{\theta}_t - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)} = t^2 \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)},$$

so

$$\|\theta_1 - \theta_2\|^2_{\int_0^1 \nabla^2 \mathcal{L}_n(t\theta_1 + (1-t)\theta_2) \mathrm{d}t} \ge \inf_{t \in [0,1]} \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\bar{\theta}_t)} \ge \frac{1}{\tau_0} \|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)}. \tag{19}$$

Combining inequalities (17) and (19) proves the lower bound (16). □

We now show that Condition 4 is implied by self-concordance:

Lemma 16. *Suppose $\mathcal{L}_n : \mathbb{R}^p \to \mathbb{R}$ is $(\gamma, 2)$-self-concordant.*

(i) For any $r > 0$, we have

$$\tau_{1,r} := \inf_{\theta \in \mathcal{B}_r(\theta_0)} \lambda_{\min}(\nabla^2 \mathcal{L}_n(\theta)) \ge \exp(-\gamma r) \lambda_{\min}(\nabla^2 \mathcal{L}_n(\theta_0)).$$

(ii) Suppose $\nabla^2 \mathcal{L}_n(\theta_0) \succ 0$ for some $\theta_0 \in \Theta_0$, and in addition, $\mathrm{diam}(\Theta_0) \le D$. Then $\mathcal{L}_n$ satisfies Condition 4 with $\tau_0 = \exp(\gamma D)$.

Proof. Suppose $\theta \in \mathcal{B}_r(\theta_0)$. By Lemma 11, we have

$$\nabla^2 \mathcal{L}_n(\theta) \succeq \exp(-\gamma r) \nabla^2 \mathcal{L}_n(\theta_0),$$

which immediately implies (i).

For (ii), note that Lemma 11 also implies that for $\theta_1, \theta_2 \in \Theta_0$, we have

$$\nabla^2 \mathcal{L}_n(\theta_2) \preceq \exp(\gamma D) \nabla^2 \mathcal{L}_n(\theta_1). \tag{20}$$

Hence, left- and right-multiplying by $(\theta_1 - \theta_2)$ and rearranging gives the upper bound

$$\frac{\|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_2)}}{\|\theta_1 - \theta_2\|^2_{\nabla^2 \mathcal{L}_n(\theta_1)}} \le \exp(\gamma D).$$

Further note that for $\theta_1 \neq \theta_0$, we have $\|\theta_1 - \theta_2\|_{\nabla^2 \mathcal{L}_n(\theta_1)} > 0$, since

$$\nabla^2 \mathcal{L}_n(\theta_1) \succeq \exp(-\gamma D) \nabla^2 \mathcal{L}_n(\theta_0) \succ 0,$$

by inequality (20). □

D Proofs for noisy gradient descent

In this appendix, we provide the proofs of the two main results on the convergence of noisy gradient descent, Theorem 2 and Proposition 1.

D.1 Proof of Theorem 2

We first present the main argument, followed by statements and proofs of supporting lemmas in succeeding subsections.

D.1.1 Main argument

It is easy to see that the Gaussian mechanism and post-processing guarantee that every iteration of the algorithm is $\frac{\mu}{\sqrt{K}}$-GDP: Fixing $\theta^{(k)}$, the global sensitivity for the gradient iterates $\theta^{(k+1)}$ in equation (4) is clearly bounded by $\frac{2\eta B}{n}$, and then we simply apply the Gaussian mechanism (Theorem 1) to obtain the noisy gradient descent iterates (5). It follows from Corollary 2 in Dong et al. [2021] that the entire algorithm is μ-GDP after K iterations. This proves (i).

Let us now turn to the proof of (ii). We denote $N_k = \frac{B\sqrt{K}}{\mu n} Z_k$, so the noisy gradient updates can be rewritten as $\theta^{(k+1)} = \theta^{(k)} - \eta \nabla \mathcal{L}_n(\theta^{(k)}) + \eta N_k$. From Lemma 4 and a union bound, we have that with probability at least $1 - \xi$,

$$\|N_k\|_2 \leq \frac{\{4\sqrt{p} + 2\sqrt{2\log(K/\xi)}\} B\sqrt{K}}{\mu n} := r_{priv}, \tag{21}$$

for all $k < K$. For the remainder of the argument, we will assume the bound (21) holds, and prove that the desired conditions hold deterministically.

From Lemmas 18 and 19 and the local strong convexity assumption, we have

$$\begin{aligned} \|\theta^{(k)} - \hat{\theta}\|_2^2 &\leq \frac{1}{\tau_1} (2\eta)^k \Delta_0 + \frac{3 r r_{priv}}{2(1-\kappa)\tau_1} \\ &\leq \frac{3 r r_{priv}}{(1-\kappa)\tau_1}, \end{aligned} \tag{22}$$

where $r_{priv} = \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\} B\sqrt{K}}{\mu n}$ and the last inequality holds as long as $k \geq k_0 = \lceil k_0' \rceil$, where

$$k_0' = \frac{\log(1/\Delta_0) + \log\left(\frac{3r\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\} B\sqrt{K}}{2(1-\kappa)\mu n}\right)}{\log(2\eta)}.$$

We note that the bound (22) is looser than the desired result. The rest of our argument will improve this estimate for later iterates $k > K$ (with $K > k_0$) by refining the argument of Lemma 19 leveraging (22).

As derived in inequality (40) in the proof of Lemma 19, we can show that

$$\Delta_{k+1} \leq (1 - 2\gamma)\Delta_k + \|N_k\|_2 \left(\|\hat{\theta} - \theta^{(k)}\|_2 + \|\theta^{(k+1)} - \theta^{(k)}\|_2 \right), \tag{23}$$

where $\Delta_k = \mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})$ and $\gamma \in (0, \frac{1}{2})$ is chosen to satisfy $\gamma \leq 2\eta\tau_1$. Furthermore, noting that smoothness, (22) and $k \geq k_0$, imply that

$$\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 = \|\nabla \mathcal{L}_n(\theta^{(k)}) - \nabla \mathcal{L}_n(\hat{\theta})\|_2 \leq 2\tau_2 \|\theta^{(k)} - \hat{\theta}\|_2 \leq 2\tau_2 \sqrt{\frac{3 r r_{priv}}{(1-\kappa)\tau_1}},$$

we see that the triangle inequality and inequality (21) give

$$\|\theta^{(k+1)} - \theta^{(k)}\|_2 = \eta\|\nabla\mathcal{L}_n(\theta^{(k)}) + N_k\|_2 \leq \eta r_{priv} + 2\tau_2\eta\sqrt{\frac{3r}{(1-\kappa)\tau_1}}r_{priv}^{1/2}. \tag{24}$$

Consequently, from inequalities (22), (23), (24) and $k \geq k_0$ we obtain

$$\Delta_{k+1} \leq (1-2\gamma)\Delta_k + r_{priv}\left(\eta r_{priv} + (2\tau_2\eta+1)\sqrt{\frac{3r}{(1-\kappa)\tau_1}}r_{priv}^{1/2}\right).$$

Taking $2\gamma = 1-\kappa$, we have that for $j = k_0+1, \ldots, K$,

$$\begin{aligned}\Delta_{j+k_0} &\leq \kappa\Delta_{j-1+k_0} + (2\tau_2\eta+1)\sqrt{\frac{3r}{(1-\kappa)\tau_1}}r_{priv}^{3/2} + \eta r_{priv}^2 \\ &\leq \kappa^j\Delta_{k_0} + \left(2\tau_2\eta + \frac{3}{2}\right)\sqrt{\frac{3r}{(1-\kappa)\tau_1}}r_{priv}^{3/2},\end{aligned} \tag{25}$$

where the last inequality used the fact that $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu\cdot\frac{1}{4\eta^2}\frac{3r}{(1-\kappa)\tau_1}} \iff \eta r_{priv}^{1/2} \leq \frac{1}{2}\sqrt{\frac{3r}{(1-\kappa)\tau_1}}$.
Taking $k = j + k_0$, local strong convexity and inequality (25) show that

$$\begin{aligned}\|\theta^{(k)} - \hat{\theta}\|_2^2 &\leq \frac{1}{\tau_1}\kappa^j\Delta_{k_0} + \left(2\tau_2\eta + \frac{3}{2}\right)\sqrt{\frac{3r}{(1-\kappa)\tau_1^3}}r_{priv}^{3/2} \\ &\leq (2\tau_2\eta+2)\sqrt{\frac{3r}{(1-\kappa)\tau_1^3}}r_{priv}^{3/2},\end{aligned} \tag{26}$$

where the last inequality holds as long as $k \geq k_1 + k_0$ with $k_1 = \lceil k_1' \rceil$ and

$$k_1' = \frac{\log(1/\Delta_{k_0}) + \log\left(\frac{1}{2}\sqrt{\frac{3r}{(1-\kappa)\tau_1}}r_{priv}^{3/2}\right)}{\log(\kappa)}.$$

Thus, we see $k_1 + k_0$ iterations of the algorithm improve the convergence rate of the iterates from $O\left(r_{priv}^{1/2}\right)$, obtained in (22) after k_0 iterations, to $O\left(r_{priv}^{3/4}\right)$. Repeating the same argument m times, we will show that successive iterates of the algorithm in fact improve the estimation error bound successively as $O\left(r_{priv}^{1/2}\right), O\left(r_{priv}^{3/4}\right), O\left(r_{priv}^{7/8}\right), \ldots, O\left(r_{priv}^{1-1/2^m}\right)$. This will be enough to prove the desired result, since we can take $m = \log_2 n$, which gives the rate $O\left(r_{priv}^{1-1/n}\right)$.

Letting $C_0 = \sqrt{\frac{3r}{(1-\kappa)\tau_1}}$, we see that inequality (22) says that for $k > k_0$, we have $\|\theta^{(k)} - \hat{\theta}\|_2 \leq C_0 r_{priv}^{1/2}$. We then saw that for $k > k_1 + k_0$ this bound can be further refined to

$$\|\theta^{(k)} - \hat{\theta}\|_2 \leq \sqrt{\frac{2\tau_2\eta+2}{\tau_1}C_0}r_{priv}^{3/4} := C_1 r_{priv}^{3/4}.$$

The same arguments used to obtain inequality (25) show that for $j = k_0 + k_1 + 1, \ldots, K$ we have

$$\begin{aligned}\Delta_{j+k_0+k_1} &\leq \kappa\Delta_{j-1+k_0} + (2\tau_2\eta+1)C_1 r_{priv}^{1+3/4} + \eta r_{priv}^2 \\ &\leq \kappa^{j+k_1}\Delta_{k_0} + \left(2\tau_2\eta + \frac{3}{2}\right)C_1 r_{priv}^{7/4},\end{aligned} \tag{27}$$

where the last inequality uses $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu\frac{1}{2^4}C_1^4}\eta^4$, which is implied by the minimum sample size assumption $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu\cdot\frac{1}{2^2}C_0^2}\eta^2$ used to show inequality (25), since $\frac{1}{2^2\eta^2}C_0^2 < \frac{1}{2^4\eta^4}C_1^4$. Similar to inequality (26), taking $k = j + k_0 + k_1$, local strong convexity and inequality (27) show that

$$\begin{aligned}\|\theta^{(k)} - \hat{\theta}\|_2^2 &\leq \frac{1}{\tau_1}\kappa^j\Delta_{k_0+k_1} + \frac{2\tau_2\eta + 3/2}{\tau_1}C_1 r_{priv}^{7/4} \\ &\leq \frac{2\tau_2\eta + 2}{\tau_1}C_1 r_{priv}^{7/4}, \end{aligned} \tag{28}$$

where the last inequality holds as long as $k \geq k_2 + k_1 + k_0$ with $k_2 = \lceil k_2' \rceil$ and

$$k_2' = \frac{\log(1/\Delta_{k_0+k_1}) + \log\left((2\tau_2\eta + 2)C_1 r_{priv}^{7/4}\right)}{\log(\kappa)}.$$

We conclude from inequality (28) that for $k > k_0 + k_1 + k_2$ we have

$$\|\theta^{(k)} - \hat{\theta}\|_2 \leq \sqrt{\frac{2\tau_2\eta + 2}{\tau_1}C_1 r_{priv}^{7/8}} = C_2 r_{priv}^{1-1/2^3}.$$

Iterating this argument, a tedious but straightforward calculation shows that we can obtain the following recurrence: Let $C_0 = \sqrt{\frac{3r}{(1-\kappa)\tau_1}}$, and define $C_m := \sqrt{\frac{2\tau_2\eta+2}{\tau_1}C_{m-1}}$, $k_m := \lceil k_m' \rceil$, and

$$k_m' = \frac{\log(1/\Delta_{\sum_{j=0}^{m-1}k_j}) + \log\left((2\tau_2\eta + 2)C_{m-1} r_{priv}^{2-1/2^m}\right)}{\log(\kappa)}.$$

Note that we also have

$$C_m = \left(\frac{2\tau_2\eta + 2}{\tau_1}\right)^{\sum_{j=1}^m 1/2^j} C_0^{1/2^m} = \left(\frac{2\tau_2\eta + 2}{\tau_1}\right)^{1-1/2^m} C_0^{1/2^m}. \tag{29}$$

Then, taking $k > \sum_{j=0}^m k_j$ and $m = \log_2 n - 1$, we have

$$\|\theta^{(k)} - \hat{\theta}\|_2 \leq C_m r_{priv}^{1-1/2^{m+1}} = \left(\frac{2\tau_2\eta + 2}{\tau_1}\right)^{1+1/2^m} C_0^{1/(\log_2 n - 1)} r_{priv}^{1-1/n} \leq C r_{priv}, \tag{30}$$

where C is some positive constant. We note that (30) implicitly used the fact that the minimum sample size requirement $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu\cdot\frac{1}{(2\eta)^2}C_0^2}$ also implies that $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu\cdot\frac{1}{(2\eta)^{2^{m+1}}}C_{m-1}^{2^m}}$ when $\eta \leq \frac{1}{2}$. To see this, it suffices to check that $C_0^2 \leq \frac{1}{(2\eta)^{2^m}}C_{m-1}^{2^m}$. The latter holds true since inequality (29) and $\tau_2 \geq \tau_1$ show that $C_{m-1} \geq (2\eta)^{1-1/2^{m-1}} C_0^{1/2^{m-1}}$. This last inequality implies the desired inequality since clearly

$$C_0^2 \leq \frac{1}{(2\eta)^{2^m}}\left((2\eta)^{1-1/2^{m-1}} C_0^{1/2^{m-1}}\right)^{2^m} = \frac{1}{(2\eta)^2}C_0^2.$$

This completes the proof.

D.1.2 Supporting lemmas

Lemma 17. *Suppose $\mathcal{L}_n$ is convex in $\Theta \subseteq \mathbb{R}^p$ and locally τ_1-strongly convex in $\mathcal{B}_r(\theta_0)$. If $\hat{\theta} \in \mathcal{B}_{r/2}(\theta_0)$ and $\mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta}) \leq \frac{r^2}{4}\tau_1$, then $\|\theta - \hat{\theta}\|_2 \leq \frac{r}{2}$ and $\theta \in \mathcal{B}_r(\theta_0)$.*

Proof. Define

$$\Delta := \inf_{\theta \notin \mathcal{B}_{r/2}(\hat{\theta})} \mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta}).$$

By convexity of $\mathcal{L}_n$, the infimum must be achieved at some point θ_r^* on the boundary of the ball $\mathcal{B}_{r/2}(\hat{\theta})$. Therefore, for any parameter θ such that $\mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta}) < \Delta$, we must have $\|\theta - \hat{\theta}\|_2 \leq \frac{r}{2}$, and hence also $\theta \in \mathcal{B}_r(\theta_0)$.

We now claim that $\Delta \geq \frac{r^2}{4}\tau_1$, from which the desired result follows. By the triangle inequality, we have $\theta_r^* \in \mathcal{B}_r(\theta_0)$. Thus, by strong convexity,

$$\Delta = \mathcal{L}_n(\theta_r^*) - \mathcal{L}_n(\hat{\theta}) \geq \tau_1 \|\theta_r^* - \hat{\theta}\|_2^2 = \frac{r^2}{4}\tau_1,$$

as claimed. □

Lemma 18. *Suppose $\mathcal{L}_n(\theta)$ is twice-differentiable almost everywhere in $\mathcal{B}_r(\theta_0)$ and satisfies LSC and strong smoothness with parameters τ_1 and $\tau_2 \leq \frac{1}{2\eta}$. Also suppose the sample size satisfies $n = \Omega\left(\frac{\sqrt{K\log(K/\xi)}}{\mu}\right)$ and inequality (21) holds, and suppose $\hat{\theta} \in \mathcal{B}_{r/2}(\theta_0)$ and $\mathcal{L}_n(\theta^{(0)}) \leq \mathcal{L}_n(\hat{\theta}) + \tau_1\frac{r^2}{4}$. Then*

$$\mathcal{L}_n(\theta^{(k)}) \leq \mathcal{L}_n(\hat{\theta}) + \tau_1\frac{r^2}{4} \quad \text{and} \quad \|\theta^{(k)} - \hat{\theta}\|_2 \leq \frac{r}{2}, \quad \forall k \leq K.$$

Proof. We will induct on k. Note that by Lemma 17, we have $\|\theta^{(0)} - \hat{\theta}\|_2 \leq \frac{r}{2}$.

For the inductive step, suppose $\mathcal{L}_n(\theta^{(k)}) \leq \mathcal{L}_n(\hat{\theta}) + \tau_1\frac{r^2}{4}$ and $\|\theta^{(k)} - \hat{\theta}\|_2 \leq \frac{r}{2}$ for some $0 \leq k < n^\alpha$. Since $\mathcal{L}_n$ is almost everywhere twice-differentiable, we have

$$\mathcal{L}_n(\theta - \eta\Delta) = \mathcal{L}_n(\theta) - \eta\langle\nabla\mathcal{L}_n(\theta), \Delta\rangle + \frac{\eta^2}{2}\langle\textstyle\int_0^1 \nabla^2\mathcal{L}_n(\theta - t\eta\Delta)\mathrm{d}t\Delta, \Delta\rangle, \tag{31}$$

for any $\Delta \in \mathbb{R}^p$. Recalling that $N_k = \frac{B\sqrt{K}}{\mu n}Z_k$ and $\theta^{(k+1)} = \theta^{(k)} - \eta\nabla\mathcal{L}_n(\theta^{(k)}) + \eta N_k$, applying equation (31) with $\theta = \theta^{(k)}$ and $\Delta = \nabla\mathcal{L}_n(\theta^{(k)}) - N_k$ leads to

$$\begin{aligned}
\mathcal{L}_n(\theta^{(k+1)}) &= \mathcal{L}_n(\theta^{(k)}) - \eta\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + \eta\langle\nabla\mathcal{L}_n(\theta^{(k)}), N_k\rangle \\
&\quad + \frac{\eta^2}{2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_{H_k}^2 - \eta^2\langle N_k, H_k\nabla\mathcal{L}_n(\theta^{(k)})\rangle + \frac{\eta^2}{2}\|N_k\|_{H_k}^2 \\
&\leq \mathcal{L}_n(\theta^{(k)}) - \eta\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + \eta\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2\|N_k\|_2 \\
&\quad + \frac{\eta^2}{2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_{H_k}^2 + \eta^2\|N_k\|_2\|H_k\|_2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2 + \frac{\eta^2}{2}\|N_k\|_{H_k}^2,
\end{aligned} \tag{32}$$

where $H_k = \int_0^1 \nabla^2\mathcal{L}_n\left(\theta^{(k)} - t\eta\nabla\mathcal{L}_n(\theta^{(k)}) + t\eta N_k\right)\mathrm{d}t$. By inequality (32) and the Cauchy-Schwarz inequality, we can guarantee that $\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\theta^{(k)})$ holds if the following inequality is satisfied:

$$\|\nabla\mathcal{L}_n(\theta^{(k)})\|_{I-\frac{\eta}{2}H_k}^2 \geq \|\nabla\mathcal{L}_n(\theta^{(k)})\|_2\|N_k\|_2 + \eta\|N_k\|_2\|H_k\|_2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2 + \frac{\eta}{2}\|N_k\|_2^2\|H_k\|_2. \tag{33}$$

The descent condition $\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\theta^{(k)})$, together with the inductive hypothesis, would then imply that

$$\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) \leq \mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \leq \tau_1 \frac{r^2}{4},$$

so from Lemma 17, we could conclude that $\theta^{(k+1)} \in \mathcal{B}_{r/2}(\hat{\theta})$, as wanted.

Returning to inequality (33), note that $\|\nabla \mathcal{L}_n(\theta^{(k)})\|^2_{I-\frac{\eta}{2}H_k} \geq \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2(1-\eta\tau_2)$, since by τ_2-smoothness, we have $\lambda_{\max}(H_k) \leq 2\tau_2$. Using this bound on the left-hand side of inequality (33) and the upper bounds $\sup_\theta \|\nabla \mathcal{L}_n(\theta)\|_2 \leq B$ and $\lambda_{\max}(H_k) \leq 2\tau_2$ on the right-hand side of inequality (33), we have the stronger sufficient condition

$$\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2(1-\eta\tau_2) \geq \|N_k\|_2 B\,(1+2\eta\tau_2) + \eta\tau_2\|N_k\|_2^2,$$

which is guaranteed to be satisfied when

$$\|N_k\| \leq \frac{-B(1+2\eta\tau_2) + \sqrt{B^2(1+2\eta\tau_2)^2 + 4\eta\tau_2(1-\eta\tau_2)\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2}}{2\eta\tau_2},$$

or equivalently,

$$\begin{aligned}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 &\geq \sqrt{\frac{\left(2\eta\tau_2\|N_k\|_2 + B(1+2\eta\tau_2)\right)^2 - B^2(1+2\eta\tau_2)^2}{4\eta\tau_2(1-\eta\tau_2)}} \\ &= \sqrt{\frac{\eta\tau_2\|N_k\|_2^2 + B(1+2\eta\tau_2)\|N_k\|_2}{1-\eta\tau_2}}. \end{aligned} \tag{34}$$

Combining inequality (34) with the upper bound (21) on $\|N_k\|_2$, we see that inequality (33) holds with high probability, provided

$$\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \geq \sqrt{\frac{\eta\tau_2 r_{priv}^2 + B(1+2\eta\tau_2) r_{priv}}{1-\eta\tau_2}} = \bar{r}_{priv}. \tag{35}$$

We now argue that if $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 < \bar{r}_{priv}$, we still have $\|\theta^{(k+1)} - \hat{\theta}\|_2 \leq \frac{r}{2}$ and $\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\hat{\theta}) + \tau_1 \frac{r^2}{4}$. Indeed, since $\theta^{(k)} \in \mathcal{B}_r(\theta_0)$ by assumption, local strong convexity implies that

$$\tau_1\|\theta^{(k)} - \hat{\theta}\|_2^2 \leq \mathcal{L}_n(\hat{\theta}) - \mathcal{L}_n(\theta^{(k)}) - \langle \nabla \mathcal{L}_n(\theta^{(k)}), \hat{\theta} - \theta^{(k)} \rangle \leq \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \|\theta^{(k)} - \hat{\theta}\|_2,$$

so

$$\bar{r}_{priv} > \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \geq \tau_1 \|\theta^{(k)} - \hat{\theta}\|_2.$$

Hence, by the triangle inequality and inequality (21), we have

$$\begin{aligned}\|\theta^{(k+1)} - \hat{\theta}\|_2 &\leq \|\theta^{(k)} - \hat{\theta}\|_2 + \|\theta^{(k+1)} - \theta^{(k)}\|_2 \\ &\leq \frac{\bar{r}_{priv}}{\tau_1} + \eta\|\nabla \mathcal{L}_n(\theta^{(k)}) + N_k\|_2 \\ &\leq \left(\frac{1}{\tau_1} + \eta\right)\bar{r}_{priv} + \eta r_{priv} \\ &\leq \left(\frac{\tau_1}{\tau_2}\right)^{1/2} \frac{r}{2} \leq \frac{r}{2} \end{aligned} \tag{36}$$

with high probability, when n is sufficiently large. Consequently, $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 < \bar{r}_{priv}$ also implies that $\theta^{(k+1)} \in \mathcal{B}_{r/2}(\hat{\theta})$; furthermore, by smoothness, we also have

$$\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) \le \tau_2 \|\theta^{(k+1)} - \hat{\theta}\|_2^2 \le \tau_2 \cdot \frac{\tau_1}{\tau_2} \frac{r^2}{4} \le \tau_1 \frac{r^2}{4}.$$

Note that the penultimate inequality in the chain (36) holds under the sample size condition $n = \Omega\left(\frac{\sqrt{K \log(K/\xi)}}{\mu}\right)$. Indeed, $\eta r_{priv} \le \frac{r}{4}\sqrt{\frac{\tau_1}{\tau_2}}$ if $n \ge \eta \frac{4\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{r\mu\sqrt{\tau_1/\tau_2}}$ and $\left(\frac{1}{\tau_1}+\eta\right)\bar{r}_{priv} \le \frac{r}{4}\sqrt{\frac{\tau_1}{\tau_2}}$ if

$$r_{priv} \le \frac{\frac{-B(1+2\eta\tau_2)}{1-\eta\tau_2} + \sqrt{\frac{B^2(1+2\eta\tau_2)^2}{(1-\eta\tau_2)^2} + \frac{r^2}{4}\frac{\eta\tau_1}{(1-\eta\tau_2)(\frac{1}{\tau_1}+\eta)^2}}}{\frac{2\eta\tau_2}{1-\eta\tau_2}}.$$

□

In fact, as the following result shows, the assumptions in Lemma 18 also imply that the optimality gap of successive iterates decreases at a geometric rate (up to an error term).

Lemma 19. *Let $\Delta_k = \mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})$, and suppose the conditions of Lemma 18 hold. Then*

$$\Delta_k = \kappa^k \Delta_0 + \frac{3r\{4\sqrt{p} + 2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{2(1-\kappa)\mu n}, \quad \forall k \le K,$$

where κ is a constant depending on τ_1 and η.

Proof. Note that $\theta^{(k+1)}$ can be viewed as the minimizer of

$$Q_k(\theta) := \mathcal{L}_n(\theta^{(k)}) + \langle \nabla \mathcal{L}_n(\theta^{(k)}) - N_k, \theta - \theta^{(k)} \rangle + \frac{1}{2\eta}\|\theta - \theta^{(k)}\|_2^2.$$

Denote $\theta_\gamma = \gamma\hat{\theta} + (1-\gamma)\theta^{(k)}$, for a parameter $\gamma \in (0,1)$ to be chosen later. By convexity of Q_k and local strong convexity, we have

$$\begin{aligned}
Q_k(\theta^{(k+1)}) &\le Q_k(\theta_\gamma) \le (1-\gamma)Q_k(\theta^{(k)}) + \gamma Q_k(\hat{\theta}) \\
&= (1-\gamma)\mathcal{L}_n(\theta^{(k)}) + \gamma \mathcal{L}_n(\hat{\theta}) + \gamma\langle \nabla \mathcal{L}_n(\theta^{(k)}) - N_k, \hat{\theta} - \theta^{(k)} \rangle + \frac{\gamma}{2\eta}\|\theta^{(k)} - \hat{\theta}\|_2^2 \\
&\le \mathcal{L}_n(\theta^{(k)}) - \gamma\left(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})\right) + \frac{\gamma}{2\eta}\|\theta^{(k)} - \hat{\theta}\|_2^2 \\
&\quad + \gamma\left(\mathcal{L}_n(\hat{\theta}) - \mathcal{L}_n(\theta^{(k)})\right) - \gamma\tau_1\|\theta^{(k)} - \hat{\theta}\|_2^2 - \gamma\langle N_k, \hat{\theta} - \theta^{(k)} \rangle \\
&\le \mathcal{L}_n(\theta^{(k)}) - 2\gamma\Delta_k + \frac{\gamma}{2\eta}\|\theta^{(k)} - \hat{\theta}\|_2^2 - \gamma\langle N_k, \hat{\theta} - \theta^{(k)} \rangle. \qquad (37)
\end{aligned}$$

Furthermore, by local smoothness and the fact that $\tau_1 \le \tau_2 \le \frac{1}{2\eta}$, we have

$$\begin{aligned}
\mathcal{L}_n(\theta^{(k+1)}) &\le \mathcal{L}_n(\theta^{(k)}) + \langle \nabla \mathcal{L}_n(\theta^{(k)}), \theta^{(k+1)} - \theta^{(k)} \rangle + \tau_2\|\theta^{(k+1)} - \theta^{(k)}\|_2^2 \\
&\le Q_k(\theta^{(k+1)}) + \langle N_k, \theta^{(k+1)} - \theta^{(k)} \rangle. \qquad (38)
\end{aligned}$$

Combining inequalities (37) and (38), we obtain

$$\mathcal{L}_n(\theta^{(k+1)}) \le \mathcal{L}_n(\theta^{(k)}) - 2\gamma\Delta_k + \frac{\gamma}{2\eta}\|\theta^{(k)} - \hat{\theta}\|_2^2 - \gamma\langle N_k, \hat{\theta} - \theta^{(k)} \rangle + \langle N_k, \theta^{(k+1)} - \theta^{(k)} \rangle. \qquad (39)$$

Furthermore, subtracting $\mathcal{L}_n(\hat{\theta})$ from both sides of inequality (39), we see that local strong convexity, taking $\gamma \le 2\eta\tau_1$, and using the bound (21) implies that

$$\begin{aligned}
\Delta_{k+1} &\le \Delta_k(1-2\gamma) + \frac{\gamma}{2\eta}\|\theta^{(k)} - \hat{\theta}\|_2^2 - \gamma\langle N_k, \hat{\theta} - \theta^{(k)}\rangle + \langle N_k, \theta^{(k+1)} - \theta^{(k)}\rangle \\
&\le (1-2\gamma)\Delta_k - \gamma\langle N_k, \hat{\theta} - \theta^{(k)}\rangle + \langle N_k, \theta^{(k+1)} - \theta^{(k)}\rangle \\
&\le (1-2\gamma)\Delta_k + \|N_k\|_2\left(\|\hat{\theta} - \theta^{(k)}\|_2 + \|\theta^{(k+1)} - \theta^{(k)}\|_2\right) \\
&\le (1-2\gamma)\Delta_k + r_{priv}\frac{3r}{2}, \qquad (40)
\end{aligned}$$

where we have used the conclusion of Lemma 18 in the last inequality. Iterating the bound (40) and writing $\kappa = 1 - 2\gamma$, we see that

$$\begin{aligned}
\Delta_k &\le \kappa^k\Delta_0 + \frac{3r_{priv}r}{2}(1 + \kappa + \kappa^2 + \cdots + \kappa^{k-1}) \\
&\le \kappa^k\Delta_0 + \frac{3r_{priv}r}{2(1-\kappa)}.
\end{aligned}$$

This shows the desired result. □

D.2 Proof of Proposition 1

We again begin by presenting the main argument, followed by statements and proofs of supporting lemmas.

D.2.1 Main argument

Let $N_k = \frac{B\sqrt{K}}{\mu n} Z_k$. By Lemma 4 and a union bound, we have

$$\max_{k<K_0} \|N_k\|_2 \le \frac{\{4\sqrt{p} + 2\sqrt{2\log(K_0/\xi_0)}\}B\sqrt{K_0}}{\mu n} := r_{priv}, \qquad (41)$$

with probability at least $1 - \xi_0$. We will assume this inequality holds for the rest of the proof, and argue that the desired conditions hold deterministically.

Recall the form of the updates (5). By strong smoothness and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathcal{L}_n(\theta^{(k+1)}) &\le \mathcal{L}_n(\theta^{(k)}) - \eta\langle\nabla\mathcal{L}_n(\theta^{(k)}), \nabla\mathcal{L}_n(\theta^{(k)}) - N_k\rangle + \tau_2\eta^2\|\nabla\mathcal{L}_n(\theta^{(k)}) - N_k\|_2^2 \\
&\le \mathcal{L}_n(\theta^{(k)}) - \eta\left(1 - \tau_2\eta\right)\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + \eta\left(1 + 2\tau_2\eta\right)\|N_2\|_2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2 + \tau_2\eta^2\|N_k\|_2^2.
\end{aligned}$$

Using the bound (41) and the fact that $\eta \le \frac{1}{2\tau_2}$, we then have

$$\mathcal{L}_n(\theta^{(k+1)}) \le \mathcal{L}_n(\theta^{(k)}) - \frac{\eta}{2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + Cr_{priv}, \qquad (42)$$

for all $k \le K_0$, where $C \ge 2\eta B + \frac{\eta}{2}r_{priv}$. Furthermore, convexity implies that

$$\mathcal{L}_n(\theta^{(k)}) \le \mathcal{L}_n(\hat{\theta}) + \langle\nabla\mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta}\rangle. \qquad (43)$$

Combining inequalities (42) and (43), we obtain

$$\mathcal{L}_n(\theta^{(k+1)}) \le \mathcal{L}_n(\hat{\theta}) + \langle\nabla\mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta}\rangle - \frac{\eta}{2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + Cr_{priv}.$$

This in turn shows that

$$\begin{aligned}
\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) &\leq \frac{1}{2\eta}\left(2\eta\langle\nabla\mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta}\rangle - \eta^2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2\right) + Cr_{priv} \\
&= \frac{1}{2\eta}\left(\|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(k)} - \eta\nabla\mathcal{L}_n(\theta^{(k)}) - \hat{\theta}\|_2^2\right) + Cr_{priv} \\
&= \frac{1}{2\eta}\left(\|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(k+1)} - \hat{\theta} - \eta N_k\|_2^2\right) + Cr_{priv} \\
&= \frac{1}{2\eta}\left(\|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(k+1)} - \hat{\theta}\|_2^2 - \eta^2\|N_k\|_2^2 + 2\eta\langle N_k, \theta^{(k+1)} - \hat{\theta}\rangle\right) + Cr_{priv} \\
&\leq \frac{1}{2\eta}\left(\|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(k+1)} - \hat{\theta}\|_2^2\right) + \|N_k\|_2\|\theta^{(k+1)} - \hat{\theta}\|_2 + Cr_{priv}.
\end{aligned}$$

Summing over k, we obtain

$$\begin{aligned}
\sum_{k=1}^{K_0}\left(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})\right) &\leq \frac{1}{2\eta}\left(\|\theta^{(0)} - \hat{\theta}\|_2^2 - \|\theta^{(K_0)} - \hat{\theta}\|_2^2\right) + K_0 \max_{k<K_0}\|N_k\|_2\|\theta^{(k+1)} - \hat{\theta}\|_2 + K_0 Cr_{priv} \\
&\leq \frac{1}{2\eta}\|\theta^{(0)} - \hat{\theta}\|_2^2 + K_0 r_{priv}(\|\theta^{(0)} - \hat{\theta}\|_2 + 1) + K_0 Cr_{priv}, \qquad (44)
\end{aligned}$$

where the last inequality is a consequence of Lemma 20.

By inequality (42), we have

$$\mathcal{L}_n(\theta^{(K_0)}) \leq \mathcal{L}_n(\theta^{(k)}) + CK_0 r_{priv}$$

for all $k < K_0$, so

$$K_0\left(\mathcal{L}_n(\theta^{(K_0)}) - \mathcal{L}_n(\hat{\theta})\right) \leq \sum_{k=1}^{K_0}\left(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})\right) + CK_0^2 r_{priv}.$$

Combined with inequality (44), this gives

$$\begin{aligned}
\mathcal{L}_n(\theta^{(K_0)}) - \mathcal{L}_n(\hat{\theta}) &\leq \frac{1}{K_0}\sum_{k=1}^{K_0}\left(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})\right) + CK_0 r_{priv} \\
&\leq \frac{1}{K_0}\frac{1}{2\eta}\|\theta^{(0)} - \hat{\theta}\|_2^2 + r_{priv}(\|\theta^{(0)} - \hat{\theta}\|_2 + 1) + K_0 Cr_{priv} \\
&\leq \frac{1}{K_0}\frac{1}{2\eta}R^2 + (R + CK_0 + 1)r_{priv} \\
&\leq \Delta,
\end{aligned}$$

where we have taken $K_0 = \frac{R^2}{\eta\Delta}$ and

$$n \geq \frac{2(R + CK_0 + 1)\{4\sqrt{p} + 2\sqrt{2\log(K_0/\xi_0)}\}B\sqrt{K_0}}{\Delta\mu}.$$

This last inequality establishes the desired result.

D.2.2 Supporting lemmas

Lemma 20. *Suppose $\mathcal{L}_n$ is convex and τ_2-smooth and let $\theta^{(k)}$ denote the updates (5), with $\eta \le \frac{1}{2\tau_2}$. Let $\|\theta^{(0)} - \hat{\theta}\|_2 = R$. Also suppose inequality (41) holds and $(K_0 C' + 2(K_0 - 1)\eta)\, r_{priv} \le 1$, where $C' = 2\eta R + 2\eta^2(B+1)$. Then*

$$\|\theta^{(k+1)} - \hat{\theta}\|_2^2 \le R^2 + \big((k+1)C' r_{priv} + 2k\eta\big)\, r_{priv} \le R^2 + 1,$$

for all $k < K_0$.

Proof. We first note that since $\mathcal{L}_n$ is convex and τ_2-smooth, we have

$$\langle \nabla \mathcal{L}_n(\theta_1) - \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle \ge \frac{1}{2\tau_2} \|\nabla \mathcal{L}_n(\theta_1) - \nabla \mathcal{L}_n(\theta_2)\|_2^2, \quad \forall \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^p. \tag{45}$$

Applying inequality (45) to the pair $(\theta^{(k)}, \hat{\theta})$, we then have

$$\frac{1}{2\tau_2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 \le \langle \nabla \mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta} \rangle.$$

Then using the Cauchy-Schwarz inequality and the fact that $\eta \le \frac{1}{2\tau_2}$, we obtain

$$\begin{aligned}
\|\theta^{(k+1)} - \hat{\theta}\|_2^2 &= \|\theta^{(k)} - \eta \nabla \mathcal{L}_n(\theta^{(k)}) + \eta N_k - \hat{\theta}\|_2^2 \\
&= \|\theta^{(k)} - \hat{\theta}\|_2^2 - 2\eta \langle \theta^{(k)} - \hat{\theta}, \nabla \mathcal{L}_n(\theta^{(k)}) - N_k \rangle + \eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)}) - N_k\|_2^2 \\
&\le \|\theta^{(k)} - \hat{\theta}\|_2^2 - \frac{\eta}{\tau_2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + 2\eta \|\theta^{(k)} - \hat{\theta}\| \|N_k\|_2 + \eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)}) - N_k\|_2^2 \\
&\le \|\theta^{(k)} - \hat{\theta}\|_2^2 - \frac{\eta}{\tau_2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + 2\eta \|\theta^{(k)} - \hat{\theta}\| \|N_k\|_2 \\
&\qquad + \eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + 2\eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \|N_k\|_2 + \eta^2 \|N_k\|_2^2 \\
&\le \|\theta^{(k)} - \hat{\theta}\|_2^2 + 2\eta \|\theta^{(k)} - \hat{\theta}\|_2 \|N_k\|_2 - \frac{\eta}{2\tau_2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + 2\eta^2 B \|N_k\|_2 + \eta^2 \|N_k\|_2^2.
\end{aligned} \tag{46}$$

It is easy to see from the condition (41) and inequality (46) that

$$\|\theta^{(1)} - \hat{\theta}\|_2^2 \le R^2 + 2\eta R r_{priv} + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \le R^2 + C' r_{priv}, \tag{47}$$

where $C' = 2\eta R + 2\eta^2(B+1) \ge 2\eta R + 2\eta^2 B + \eta^2 r_{priv}$. Furthermore, inequalities (46) and (47) also imply that

$$\begin{aligned}
\|\theta^{(2)} - \hat{\theta}\|_2^2 &\le \|\theta^{(1)} - \hat{\theta}\|_2^2 + 2\eta \|\theta^{(1)} - \hat{\theta}\|_2 r_{priv} + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \\
&\le (R^2 + C' r_{priv}) + 2\eta r_{priv} (R + \sqrt{C' r_{priv}}) + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \\
&\le R^2 + 2C' r_{priv} + 2\eta r_{priv} \sqrt{C' r_{priv}} \\
&\le R^2 + (2C' + 2\eta) r_{priv},
\end{aligned}$$

where the last inequality uses the fact that $C' r_{priv} \le 1$. A similar argument shows that

$$\begin{aligned}
\|\theta^{(3)} - \hat{\theta}\|_2^2 &\le \|\theta^{(2)} - \hat{\theta}\|_2^2 + 2\eta \|\theta^{(2)} - \hat{\theta}\|_2 r_{priv} + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \\
&\le (R^2 + 2C' r_{priv}) + 2\eta r_{priv} + 2\eta r_{priv} (R + \sqrt{(2C' + 2\eta) r_{priv}}) + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \\
&\le R^2 + 3C' r_{priv} + 2\eta r_{priv} (1 + \sqrt{(2C' + 2\eta) r_{priv}}) \\
&\le R^2 + 3C' r_{priv} + 4\eta r_{priv},
\end{aligned}$$

where the last inequality uses the assumption $(2C'+2\eta)r_{priv} \le 1$. More generally, one can recursively verify that as long as $\left((k+1)C' + 2k\eta\right) r_{priv} \le 1$, we also have

$$\|\theta^{(k+1)} - \hat{\theta}\|_2^2 \le R^2 + \left((k+1)C' r_{priv} + 2k\eta\right) r_{priv}.$$

□

E Proofs for noisy Newton's method

In this appendix, we provide proofs of Theorems 3 and 4 and Proposition 2.

E.1 Proof of Theorem 3

We begin by presenting the main argument, followed by statements and proofs of supporting lemmas.

E.1.1 Main argument

The arguments for part (i) are very similar to those presented for Theorem 2. In particular, every iteration of the algorithm is $\frac{\mu}{\sqrt{K}}$-GDP by the Gaussian mechanism, post-processing, and the composition theorem of [Dong et al., 2021, Corollary 1]. The same composition result shows that the whole algorithm is μ-GDP after K iterations. This proves (i).

We now turn to the proof of (ii). Let $\tilde{W}_k = \frac{2\bar{B}\sqrt{2K}}{\mu n} W_k$ and $\tilde{Z}_k = \frac{2B\sqrt{2K}}{\mu n} Z_k$, and note that the Neumann series formula leads to the identity

$$\left\{\nabla^2 \mathcal{L}_n(\theta^{(k)}) + \tilde{W}_k\right\}^{-1} = \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \sum_{j=0}^{\infty} \left[-\tilde{W}_k \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1}\right]^j.$$

Hence,

$$\begin{aligned}
\theta^{(k+1)} &= \theta^{(k)} - \eta \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \sum_{j=0}^{\infty} \left[-\tilde{W}_k \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1}\right]^j \left\{\nabla \mathcal{L}_n(\theta^{(k)}) + \tilde{Z}_k\right\} \\
&= \theta^{(k)} - \eta \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)}) + \eta \tilde{N}_k,
\end{aligned}$$

where

$$\tilde{N}_k = \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \left(\tilde{Z}_k + \sum_{j=1}^{\infty} \left[-\tilde{W}_k \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1}\right]^j \left\{\nabla \mathcal{L}_n(\theta^{(k)}) + \tilde{Z}_k\right\} \right). \tag{48}$$

Note that Lemmas 4 and 5 and a union bound imply that with probability at least $1-\xi$, we have $\max_{k<K} \|Z_k\|_2 \le 4\sqrt{p} + 2\sqrt{2\log(2K/\xi)}$ and $\max_{k<K} \|W_k\|_2 \le \sqrt{2p\log(4Kp/\xi)}$. We will assume that these inequalities hold and argue deterministically for the rest of the proof.

Lemma 21 is our main workhorse. The bound (50), together with the starting value condition $\|\nabla \mathcal{L}_n(\theta^{(0)})\|_2 \le \frac{\tau_1^2}{L}$, implies that

$$\frac{L}{2\tau_1^2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 < \left(\frac{L}{2\tau_1^2} \|\nabla \mathcal{L}_n(\theta^{(0)})\|_2\right)^{2^k} + 3C\tilde{r}_{priv}, \tag{49}$$

for all $k \le K$. Indeed, we can prove inequality (49) by induction: The base case $k = 1$ holds by inequality (50), and if we assume the bound holds for some $k \ge 1$, then

$$\begin{aligned}
\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k+1)})\|_2 &\le \left(\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2\right)^2 + C\tilde{r}_{priv} \\
&\le \left\{\left(\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(0)})\|_2\right)^{2^k} + 3C\tilde{r}_{priv}\right\}^2 + C\tilde{r}_{priv} \\
&\le \left(\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(0)})\|_2\right)^{2^{k+1}} + C\tilde{r}_{priv}\left(\frac{3}{2} + 9C\tilde{r}_{priv}\right) + C\tilde{r}_{priv} \\
&< \left(\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(0)})\|_2\right)^{2^{k+1}} + 3C\tilde{r}_{priv},
\end{aligned}$$

where the first inequality comes from inequality (50); the second inequality follows by the induction hypothesis; and the last inequality uses the fact that the sample size condition $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$ guarantees that $C\tilde{r}_{priv} < \frac{1}{18}$. This completes the inductive step.

Note that inequality (49) implies that

$$\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(K)})\|_2 < \left(\frac{1}{2}\right)^{2^K} + 3C\tilde{r}_{priv}.$$

Hence, for $K \ge \frac{1}{\log(2)}\log\left(\frac{\log(C\tilde{r}_{priv})}{\log(1/2)}\right)$, we have

$$4C\tilde{r}_{priv} \ge \frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(K)})\|_2 \ge \frac{L}{2\tau_1^2}\cdot 2\tau_1\|\theta^{(K)} - \hat{\theta}\|_2,$$

where the last inequality holds by LSC. Rearranging yields the desired result.

E.1.2 Supporting lemmas

In the statements and proofs of these lemmas, we work under the assumption that $\max_{k<K}\|Z_k\|_2 \le 4\sqrt{p} + 2\sqrt{2\log(2K/\xi)}$ and $\max_{k<K}\|W_k\|_2 \le \sqrt{2p\log(4Kp/\xi)}$, in addition to the assumptions stated in Theorem 3.

Lemma 21. *For all $0 \le k < K$, we have*

(a) $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \le \min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\}$,

(b) $\|\theta^{(k)} - \hat{\theta}\|_2 \le r$, *and*

(c)

$$\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k+1)}\|_2 \le \left(\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2\right)^2 + C\tilde{r}_{priv}, \tag{50}$$

where $\tilde{r}_{priv} := \frac{\bar{B}B\sqrt{Kp\log(Kp/\xi)}}{\mu n\tau_1^2}$.

In addition, $\|\theta^{(K)} - \hat{\theta}\|_2 \le r$.

Proof. We induct on k. For the inductive hypothesis, consider the case $k = 0$. Note that (a) holds by assumption. Lemma 22 then implies that (b) holds, and then Lemma 23 implies that (c) holds, as well.

Turning to the inductive step, assume that (a), (b), and (c) all hold for some $k \geq 0$. We will show that $\|\nabla \mathcal{L}_n(\theta^{(k+1)})\|_2 \leq \min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\}$; then Lemmas 22 and 23 imply that statements (b) and (c) hold, as well. Note that by the inductive hypothesis, we have

$$\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \leq \frac{1}{2}.$$

Then by (a) and (c), we have

$$\begin{aligned}
\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k+1)}\|_2 &\leq \left(\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2\right)^2 + C\tilde{r}_{priv} \\
&\leq \frac{1}{2}\left(\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2\right) + C\tilde{r}_{priv} \\
&\leq \frac{1}{2}\cdot\frac{L}{2\tau_1^2}\min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\} + C\tilde{r}_{priv} \\
&\leq \frac{L}{2\tau_1^2}\min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\}, \qquad (51)
\end{aligned}$$

using the sample size assumption $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$. This completes the inductive step.

Finally, note that the argument used to establish inequality (51) above shows that we can also deduce statements (a) and (b) for $k = K$. □

Lemma 22. *Suppose* $\|\nabla \mathcal{L}_n(\theta)\|_2 < 2\tau_1 r$. *Then* $\|\theta - \hat{\theta}\|_2 \leq r$.

Proof. We prove the contrapositive. Suppose $\|\theta - \hat{\theta}\|_2 > r$, and let $\widetilde{\theta}$ denote the point on the boundary of $\mathcal{B}_r(\hat{\theta})$ lying on the segment between θ and $\hat{\theta}$. By the LSC condition, we have

$$\langle \nabla \mathcal{L}_n(\widetilde{\theta}), \widetilde{\theta} - \hat{\theta}\rangle \geq 2\tau_1\|\widetilde{\theta} - \hat{\theta}\|_2^2.$$

Rearranging and defining $v = \frac{\widetilde{\theta}-\hat{\theta}}{\|\widetilde{\theta}-\hat{\theta}\|_2}$, we then have

$$\langle \nabla \mathcal{L}_n(\widetilde{\theta}), v\rangle \geq 2\tau_1 r.$$

Further note that if we define $f(t) = \langle \nabla \mathcal{L}_n(\hat{\theta} + tv), v\rangle$ for $t \geq 0$, then $f'(t) = v^T \nabla^2 \mathcal{L}_n(\hat{\theta} + tv)v \geq 0$, so $f(t)$ is increasing. This implies that

$$\|\nabla \mathcal{L}_n(\theta)\|_2 \geq \langle \nabla \mathcal{L}_n(\theta), v\rangle \geq \langle \nabla \mathcal{L}_n(\widetilde{\theta}), v\rangle \geq 2\tau_1 r,$$

giving the desired result. □

Lemma 23. *Suppose* $\theta^{(k)} \in \mathcal{B}_r(\hat{\theta})$. *Then inequality* (50) *holds.*

Proof. We begin by establishing a bound on $\|\tilde{N}_k\|_2$. By LSC, we have $\|\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})\right\}^{-1}\|_2 \le \frac{1}{2\tau_1}$. Furthermore, by our sample size assumption, we have $\max_{k<K}\|\tilde{W}_k\|_2 \le \tau_1$. Then the series expansion (48) and the bounds on $\|Z_k\|_2$ and $\|W_k\|_2$ imply that for all $k < K$, we have

$$\begin{aligned}
\|\tilde{N}_k\|_2 &\le \frac{1}{2\tau_1}\|\tilde{Z}_k\|_2 + \sum_{j=1}^{\infty}\frac{\|\tilde{W}_k\|_2^j}{(2\tau_1)^{(j+1)}}\left\{\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2 + \|\tilde{Z}_k\|_2\right\} \\
&\le \frac{1}{2\tau_1}\|\tilde{Z}_k\|_2 + \frac{1}{2\tau_1}\frac{\|\tilde{W}_k\|_2}{2\tau_1 - \|\tilde{W}_k\|_2}\left\{\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2 + \|\tilde{Z}_k\|_2\right\} \\
&\le \frac{2B\sqrt{2K}}{2\mu n\tau_1}\|Z_k\|_2 + \frac{1}{2\tau_1}\frac{\|\tilde{W}_k\|_2}{2\tau_1 - \|\tilde{W}_k\|_2}\left\{B + \frac{2B\sqrt{2K}}{\mu n}\|Z_k\|_2\right\} \\
&\le \frac{B\sqrt{2K}(4\sqrt{p} + 2\sqrt{2\log(2K/\xi)})}{\mu n\tau_1} \\
&\quad + \frac{2\bar{B}\sqrt{2K}}{2\tau_1\mu n}\frac{\|W_k\|_2}{\tau_1}\left(B + \frac{2B\sqrt{2K}(4\sqrt{p} + 2\sqrt{2\log(2K/\xi)})}{\mu n}\right) \\
&\le C_0\frac{B\sqrt{K}\{\sqrt{p} + \sqrt{\log(2K/\xi)}\} + \bar{B}B\tau_1^{-1}\sqrt{Kp\log(Kp/\xi)}}{\mu n\tau_1} = C'\tilde{r}_{priv},
\end{aligned} \tag{52}$$

where $C_0 > 0$ is a constant.

To establish inequality (50), we write $\theta^{k+1} = \theta^{(k)} + \Delta\theta^{(k)} + \tilde{N}_k$, where

$$\Delta\theta^{(k)} = -\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})\right\}^{-1}\nabla\mathcal{L}_n(\theta^{(k)}).$$

Using Condition 3, the LSC condition, and the triangle inequality, we then obtain

$$\begin{aligned}
&\|\nabla\mathcal{L}_n(\theta^{(k+1)})\|_2 \\
&= \left\|\nabla\mathcal{L}_n\left(\theta^{(k)} + \Delta\theta^{(k)} + \tilde{N}_k\right)\right\|_2 \\
&= \left\|\nabla\mathcal{L}_n\left(\theta^{(k)} + \Delta\theta^{(k)} + \tilde{N}_k\right) - \nabla\mathcal{L}_n(\theta^{(k)}) - \nabla^2\mathcal{L}_n(\theta^{(k)})\left(\Delta\theta^{(k)} - \tilde{N}_k + \tilde{N}_k\right)\right\|_2 \\
&= \left\|\int_0^1\left\{\nabla^2\mathcal{L}_n\left(\theta^{(k)} + t(\Delta\theta^{(k)} + \tilde{N}_k)\right) - \nabla^2\mathcal{L}_n(\theta^{(k)})\right\}(\Delta\theta^{(k)} + \tilde{N}_k)\mathrm{d}t + \nabla^2\mathcal{L}_n(\theta^{(k)})\tilde{N}_k\right\|_2 \\
&\le \left\|\int_0^1 Lt\|\Delta\theta^{(k)} - \tilde{N}_k\|_2^2\mathrm{d}t + \nabla^2\mathcal{L}_n(\theta^{(k)})\tilde{N}_k\right\|_2 \\
&\le \frac{L}{2}\|\Delta\theta^{(k)} + \tilde{N}_k\|_2^2 + \|\nabla^2\mathcal{L}_n(\theta^{(k)})\tilde{N}_k\|_2 \\
&\le \frac{L}{2}\|\Delta\theta^{(k)}\|_2^2 + \left(L\|\Delta\theta^{(k)}\|_2 + \bar{B}\right)\|\tilde{N}_k\|_2 + \frac{L}{2}\|\tilde{N}_k\|_2^2 \\
&\le \frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + \left(L\|\Delta\theta^{(k)}\|_2 + \bar{B}\right)\|\tilde{N}_k\|_2 + \frac{L}{2}\|\tilde{N}_k\|_2^2.
\end{aligned}$$

Since LSC guarantees that $\|\Delta\theta^{(k)}\|_2 \le \frac{B}{\tau_1}$, the desired result follows from the bound (52) on $\|\tilde{N}_k\|_2$ and the sample size condition. □

E.2 Proof of Theorem 4

We begin by presenting the main argument, followed by the proofs of the supporting lemmas.

E.2.1 Main argument

It is easy to see that (i) follows by the same arguments as in Theorem 3. For (ii), we borrow arguments from the proof of Theorem 3 of Sun and Tran-Dinh [2019] and follow a similar approach as in our analysis of noisy Newton's method under LSC in Theorem 3. We relegate the statements and proofs of the key auxiliary lemmas to Appendix E.2.2.

As in the proof of Theorem 3, we first control the additive noise introduced in the first K steps via a union bound, and then argue deterministically for the rest of the proof. Let $\tilde{W}_k = \frac{2\bar{B}\sqrt{2K}}{\mu n} W_k$ and $\tilde{Z}_k = \frac{2B\sqrt{2K}}{\mu n} Z_k$. Then Lemmas 4 and 5 and a union bound imply that with probability at least $1 - \xi$, we have $\max_{k<K} \|Z_k\|_2 \le 4\sqrt{p} + 2\sqrt{2\log(2K/\xi)}$ and $\max_{k<K} \|W_k\|_2 \le \sqrt{2p\log(4Kp/\xi)}$. We will assume that these inequalities hold for the rest of the proof.

Define

$$\lambda_k := \lambda_{\min}^{-1/2}\left(\nabla^2 \mathcal{L}_n(\theta^{(k)})\right)\lambda(\theta^{(k)}) = \lambda_{\min}^{-1/2}\left(\nabla^2 \mathcal{L}_n(\theta^{(k)})\right)\|\nabla \mathcal{L}_n(\theta^{(k)})\|_{\nabla^2 \mathcal{L}_n(\theta^{(k)})^{-1}}.$$

Lemma 26 shows that the sequence $\{\lambda_k\}$ (approximately) converges at a quadratic rate. We now show that this implies a similar convergence statement about the iterates $\|\theta^{(k)} - \hat{\theta}\|_2$. Applying Lemma 13 with $x = \hat{\theta}$ and $y = \theta^{(k)}$, we have

$$\frac{1 - \exp(-\gamma\|\theta^{(k)} - \hat{\theta}\|_2)}{\gamma\|\theta^{(k)} - \hat{\theta}\|_2^2}\|\hat{\theta}^{(k)} - \hat{\theta}\|_{\nabla^2 \mathcal{L}_n(\theta^{(k)})} \le \|\nabla \mathcal{L}_n(\theta^{(k)})\|_{\nabla^2 \mathcal{L}_n(\theta^{(k)})^{-1}} = \lambda(\theta^{(k)}).$$

By Lemma 24, we have $\|\theta^{(k)} - \hat{\theta}\|_2 \le \frac{1}{\gamma}$. Thus, the left-hand expression is lower-bounded by $\frac{1}{2}\|\hat{\theta}^{(k)} - \hat{\theta}\|_{\nabla^2 \mathcal{L}_n(\theta^{(k)})}$, implying that

$$\lambda_{\min}\left(\nabla^2 \mathcal{L}_n(\theta^{(k)})\right)^{1/2}\frac{\|\hat{\theta}^{(k)} - \hat{\theta}\|_2}{2} \le \frac{1}{2}\|\hat{\theta}^{(k)} - \hat{\theta}\|_{\nabla^2 \mathcal{L}_n(\theta^{(k)})} \le \lambda(\theta^{(k)}).$$

Hence, by Lemma 26, we have

$$\|\hat{\theta}^{(k)} - \hat{\theta}\|_2 \le 2\lambda_k \le \frac{1}{\gamma}\left((2\gamma\lambda_0)^{2^k} + 8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}\right) \le \frac{1}{\gamma}\left(\frac{1}{8}\right)^{2^k} + 8\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}},$$

for all $k \le K$, where $\zeta = C'\frac{\bar{B}B\sqrt{Kp\log(Kp/\xi)}}{\mu n \tau_{1,2/\gamma}^2}$ is as defined in Lemma 24. (Note that the condition $8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \le \frac{1}{4}$ holds by our sample size assumption.) Finally, taking $K = \Omega(\log\log n)$ yields the desired error bound.

E.2.2 Supporting lemmas

In the statements and proofs of the following lemmas, we are working under the assumption that $\max_{k<K} \|Z_k\|_2 \le 4\sqrt{p} + 2\sqrt{2\log(2K/\xi)}$ and $\max_{k<K} \|W_k\|_2 \le \sqrt{2p\log(4Kp/\xi)}$, in addition to the assumptions stated in Theorem 4.

Lemma 24. *For all $0 \le k < K$, we have*

(a) $\lambda_k \le \frac{1}{16\gamma}$,

(b) $\|\theta^{(k)} - \hat{\theta}\|_2 \le \frac{1}{\gamma}$, *and*

(c)

$$\lambda_{k+1} \le 2\gamma\left(\lambda_k + \zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}\right)^2 + \exp\left(\frac{1}{8}\right)\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}, \tag{53}$$

where $\zeta := C'\frac{\bar{B}B\sqrt{Kp\log(Kp/\xi)}}{\mu n\tau_{1,2/\gamma}^2}$.

In addition, $\|\theta^{(K)} - \hat{\theta}\|_2 \le \frac{1}{\gamma}$.

Proof. We induct on k. For the inductive hypothesis, consider the case $k = 0$. Note that (a) holds by assumption. Lemma 25 then implies that (b) and (c) hold, as well.

Turning to the inductive step, assume that (a), (b), and (c) all hold for some $k \ge 0$. It clearly suffices to show that $\lambda_{k+1} \le \frac{1}{16\gamma}$; then Lemma 25 implies that statements (b) and (c) also hold. Note that by our sample size assumption $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$, we can make $\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \le \frac{1}{32\gamma}$, so statement (c) of the inductive hypothesis implies that

$$\lambda_{k+1} \le 2\gamma\left(\frac{1}{8\gamma}\right)^2 + \frac{1}{32\gamma}\exp\left(\frac{1}{8}\right) \le \frac{1}{16\gamma},$$

as wanted.

Finally, note that the argument above shows that we can also deduce statement (a) for $k = K$, and statement (b) then follows by Lemma 25. □

Lemma 25. *Suppose* $\lambda_k \le \frac{1}{16\gamma}$. *Then* $\|\theta^{(k)} - \hat{\theta}\|_2 \le \frac{1}{\gamma}$ *and inequality* (53) *holds.*

Proof. We first show that $\theta^{(k)} \in \mathcal{B}_{1/\gamma}(\hat{\theta})$. By Lemma 13, we have

$$\begin{aligned}\frac{1-\exp(-\gamma\|\theta^{(k)} - \hat{\theta}\|_2)}{\gamma\|\theta^{(k)} - \hat{\theta}\|_2} \cdot \lambda_{\min}(\theta^{(k)})^{1/2}\|\theta^{(k)} - \hat{\theta}\|_2 &\le \frac{1-\exp(-\gamma\|\theta^{(k)} - \hat{\theta}\|_2)}{\gamma\|\theta^{(k)} - \hat{\theta}\|_2} \cdot \|\theta^{(k)} - \hat{\theta}\|_{\nabla^2\mathcal{L}_n(\theta^{(k)})} \\ &\le \|\nabla\mathcal{L}_n(\theta^{(k)})\|_{\nabla^2\mathcal{L}_n(\theta^{(k)})^{-1}},\end{aligned}$$

implying that

$$\frac{1-\exp(-\gamma\|\theta^{(k)} - \hat{\theta}\|_2)}{\gamma} \le \lambda_k \le \frac{1}{16\gamma}.$$

Rearranging, we see that

$$\gamma\|\theta^{(k)} - \hat{\theta}\|_2 \le \log\left(\frac{16}{15}\right) \le 1,$$

implying the desired result.

Turning to inequality (53), let $v = \theta^{(k+1)} - \theta^{(k)}$. We claim that under the assumptions, we have

$$\|v\|_{\nabla^2\mathcal{L}_n(\theta^{(k)})} \le \lambda(\theta^{(k)}) + \sqrt{\bar{B}}\zeta. \tag{54}$$

Indeed, the same logic applied in the proof of Lemma 23, together with the fact that $\theta^{(k)} \in \mathcal{B}_{1/\gamma}(\hat{\theta})$ and the lower bound on the Hessian guaranteed by Lemma 16, implies that $\|\tilde{N}_k\|_2 \le \zeta$, where

$$\tilde{N}_k = v + \left\{\nabla^2\mathcal{L}_n(\theta^{(k)})\right\}^{-1}\nabla\mathcal{L}_n(\theta^{(k)}).$$

Hence, by the triangle inequality,

$$\|v\|_{\nabla^2\mathcal{L}_n(\theta^{(k)})} \le \lambda(\theta^{(k)}) + \sqrt{\bar{B}}\|\tilde{N}_k\|_2 \le \lambda(\theta^{(k)}) + \sqrt{\bar{B}}\zeta.$$

Now let

$$G_k := \int_0^1 \left(\nabla^2\mathcal{L}_n(\theta^{(k)} + tv) - \nabla^2\mathcal{L}_n(\theta^{(k)})\right) dt,$$
$$H_k := \nabla^2\mathcal{L}_n(\theta^{(k)})^{-1/2} G_k \nabla^2\mathcal{L}_n(\theta^{(k)})^{-1/2}.$$

By Lemma 12, we have

$$\|H_k\|_2 \le \left(\frac{3}{2} + \frac{\gamma\|v\|_2}{3}\right)\gamma\|v\|_2 \exp\left(\gamma\|v\|_2\right). \tag{55}$$

Furthermore,

$$\begin{aligned}\nabla\mathcal{L}_n(\theta^{(k+1)}) &= \nabla\mathcal{L}_n(\theta^{(k+1)}) - \nabla\mathcal{L}_n(\theta^{(k)}) - \nabla^2\mathcal{L}_n(\theta^{(k)})(v - \tilde{N}_k)\\ &= G_k v + \nabla^2\mathcal{L}_n(\theta^{(k)})\tilde{N}_k.\end{aligned}$$

Hence,

$$\begin{aligned}\|\nabla\mathcal{L}_n(\theta^{(k+1)})\|_{\nabla^2\mathcal{L}_n(\theta^{(k)})^{-1}} &\le \left(v^T G_k \nabla^2\mathcal{L}_n(\theta^{(k)})^{-1} G_k v\right)^{1/2} + \sqrt{\bar{B}}\zeta\\ &= \left(v^T\nabla^2\mathcal{L}_n(\theta^{(k)})^{1/2} H_k^2 \nabla^2\mathcal{L}_n(\theta^{(k)})^{1/2} v\right)^{1/2} + \sqrt{\bar{B}}\zeta\\ &\le \|H_k\|_2\|v\|_{\nabla^2\mathcal{L}_n(\theta^{(k)})} + \sqrt{\bar{B}}\zeta.\end{aligned} \tag{56}$$

Now note that by Lemma 11, with $x = \theta^{(k)}$ and $y = \theta^{(k+1)}$, we have

$$\lambda(\theta^{(k+1)})^2 \le \exp(\gamma\|v\|_2)\|\nabla\mathcal{L}_n(\theta^{(k+1)})\|^2_{\nabla^2\mathcal{L}_n(\theta^{(k)})^{-1}}.$$

Combined with inequalities (54), (55), and (56), this implies that

$$\begin{aligned}\lambda(\theta^{(k+1)}) &\le \exp\left(\frac{\gamma}{2}\|v\|_2\right)\left(\|H_k\|_2\|v\|_{\nabla^2\mathcal{L}_n(\theta^{(k)})} + \sqrt{\bar{B}}\zeta\right)\\ &\le \exp\left(\frac{\gamma}{2}\|v\|_2\right)\left(\frac{3}{2} + \frac{\gamma\|v\|_2}{3}\right)\gamma\|v\|_2\exp\left(\gamma\|v\|_2\right)\left(\lambda(\theta^{(k)}) + \sqrt{\bar{B}}\zeta\right) + \exp\left(\frac{\gamma}{2}\|v\|_2\right)\sqrt{\bar{B}}\zeta.\end{aligned} \tag{57}$$

By inequality (54), we also have

$$\lambda_{\min}^{1/2}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)\|v\|_2 \le \lambda(\theta^{(k)}) + \sqrt{\bar{B}}\zeta,$$

so rearranging,

$$\|v\|_2 \le \lambda_k + \lambda_{\min}^{-1/2}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)\sqrt{\bar{B}}\zeta. \tag{58}$$

Furthermore, Lemma 11 also implies that

$$\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right) \le \lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k+1)})\right)\exp(\gamma\|v\|_2),$$

so

$$\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k+1)})\right)^{-1/2} \le \lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)^{-1/2}\exp\left(\frac{\gamma}{2}\|v\|_2\right).$$

Combined with inequalities (57) and (58), it follows that

$$\begin{aligned}
\lambda_{k+1} \le \exp\left(\frac{\gamma}{2}\|v\|_2\right)\left(\frac{3}{2}+\frac{\gamma\|v\|_2}{3}\right)\gamma\exp\left(\gamma\|v\|_2\right)\left(\lambda(\theta^{(k)})+\sqrt{\bar{B}}\zeta\right) \\
\cdot\left(\lambda_k+\lambda_{\min}^{-1/2}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)\sqrt{\bar{B}}\zeta\right) \\
\cdot\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)^{-1/2}\exp\left(\frac{\gamma}{2}\|v\|_2\right) \\
+\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)^{-1/2}\exp\left(\gamma\|v\|_2\right)\sqrt{\bar{B}}\zeta \\
=\gamma\exp\left(2\gamma\|v\|_2\right)\left(\frac{3}{2}+\frac{\gamma\|v\|_2}{3}\right)\left(\lambda_k+\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)^{-1/2}\sqrt{\bar{B}}\zeta\right)^2 \\
+\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)^{-1/2}\exp\left(\gamma\|v\|_2\right)\sqrt{\bar{B}}\zeta.
\end{aligned}$$

We can check that if $\gamma\|v\|_2 \le \frac{1}{8}$, then

$$\exp(2\gamma\|v\|_2)\left(\frac{3}{2}+\frac{\gamma\|v\|_2}{3}\right) \le 2,$$

implying that

$$\lambda_{k+1} \le 2\gamma\left(\lambda_k+\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)^{-1/2}\sqrt{\bar{B}}\zeta\right)^2+\exp\left(\frac{1}{8}\right)\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right)^{-1/2}\sqrt{\bar{B}}\zeta. \tag{59}$$

Note that since $\|\theta^{(k)}-\hat{\theta}\|_2 \le \frac{1}{\gamma}$ and $\hat{\theta} \in \mathcal{B}_{1/\gamma}(\theta_0)$ by assumption, we have $\lambda_{\min}\left(\nabla^2\mathcal{L}_n(\theta^{(k)})\right) \ge \tau_{1,2/\gamma}$. Combining this with inequality (59) yields inequality (53).

Finally, note that if $\lambda_k \le \frac{1}{16\gamma}$, then by inequality (58) and the sample size condition, we indeed have $\|v\|_2 \le \frac{1}{8\gamma}$. □

Lemma 26. *Suppose* $8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \le \frac{1}{4}$ *and* $\lambda_0 \le \frac{1}{16\gamma}$, *with* ζ *as defined in Lemma 24. Then for* $k \le K$, *we have*

$$2\gamma\lambda_k \le (2\gamma\lambda_0)^{2^k}+8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}. \tag{60}$$

Proof. We will use induction to show that inequality (60) holds for all $k \le K$. The base case, $k = 0$, is obvious. For the inductive step, assume the inequality holds for some $k \ge 0$.

By Lemma 24, we have

$$\lambda_{k+1} \le 2\gamma\left(\lambda_k+\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}\right)^2+\exp\left(\frac{1}{8}\right)\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \tag{61}$$

$$\begin{aligned}
&=2\gamma\lambda_k^2+4\gamma\lambda_k\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}+2\gamma\zeta^2\frac{\bar{B}}{\tau_{1,2/\gamma}}+\exp\left(\frac{1}{8}\right)\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \\
&\le 2\gamma\lambda_k^2+\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}\left(\frac{1}{4}+\frac{1}{4}+\exp\left(\frac{1}{8}\right)\right) \\
&=2\gamma\lambda_k^2+2\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}},
\end{aligned} \tag{62}$$

using the fact that $\lambda_k \le \frac{1}{16\gamma}$ from Lemma 24 and the assumption $2\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \le \frac{1}{4}$.

Hence,

$$\begin{aligned}
2\gamma\lambda_{k+1} &\le 2\gamma\left(2\gamma\lambda_k^2 + 2\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}\right) \\
&= (2\gamma\lambda_k)^2 + 4\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \\
&\le \left((2\gamma\lambda_0)^{2^k} + 8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}\right)^2 + 4\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \\
&= (2\gamma\lambda_0)^{2^{k+1}} + 8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}}\left(8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} + 2\,(2\gamma\lambda_0)^{2^k}\right) + 4\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \\
&\le (2\gamma\lambda_0)^{2^{k+1}} + 8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}},
\end{aligned}$$

where the second inequality comes from the induction hypothesis and the final inequality uses the facts that $8\gamma\zeta\sqrt{\frac{\bar{B}}{\tau_{1,2/\gamma}}} \le \frac{1}{4}$ and $2\gamma\lambda_0 \le \frac{1}{8}$. □

E.3 Proof of Proposition 2

We first present the main argument, followed by statements and proofs of supporting lemmas.

E.3.1 Main argument

By Lemma 28, we know that with probability at least $1-\xi$, all iterates $\{\theta^{(k)}\}_{k=1}^K$ lie in $\Theta_0 = \mathcal{B}_{R_0}(\hat{\theta})$. Thus, we may apply Lemma 15 to the pair $(\theta^{(k)}, \theta^{(k+1)})$ to obtain

$$\mathcal{L}_n(\theta^{(k+1)}) \le \mathcal{L}_n(\theta^{(k)}) + \langle \nabla\mathcal{L}_n(\theta^{(k)}), \Delta\theta^{(k)}\rangle + \frac{\tau_0}{2}\|\Delta\theta^{(k)}\|^2_{\nabla^2\mathcal{L}_n(\theta^{(k)})},$$

where $\tau_0 = \exp(2\gamma R_0)$ (by Lemma 16), $\Delta\theta^{(k)} = \theta^{(k+1)} - \theta^{(k)} = -\eta\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})\right\}^{-1}\nabla\mathcal{L}_n(\theta^{(k)}) + \eta\tilde{N}_k$, and $\tilde{N}_k$ is as defined in equation (48). Note that

$$\begin{aligned}
\langle \nabla\mathcal{L}_n(\theta^{(k)}), \Delta\theta^{(k)}\rangle &= -\eta\langle \nabla\mathcal{L}_n(\theta^{(k)}), \{\nabla^2\mathcal{L}_n(\theta^{(k)})\}^{-1}\nabla\mathcal{L}_n(\theta^{(k)})\rangle + \eta\langle \nabla\mathcal{L}_n(\theta^{(k)}), \tilde{N}_k\rangle \\
&= -\eta\|\{\nabla^2\mathcal{L}_n(\theta^{(k)})\}^{-1}\nabla\mathcal{L}_n(\theta^{(k)})\|^2_{\nabla^2\mathcal{L}_n(\theta^{(k)})} + \eta\langle \nabla\mathcal{L}_n(\theta^{(k)}), \tilde{N}_k\rangle
\end{aligned}$$

and

$$\|\Delta\theta^{(k)}\|^2_{\nabla^2\mathcal{L}_n(\theta^{(k)})} \le 2\eta^2\left(\|\{\nabla^2\mathcal{L}_n(\theta^{(k)})\}^{-1}\nabla\mathcal{L}_n(\theta^{(k)})\|^2_{\nabla^2\mathcal{L}_n(\theta^{(k)})} + \|\tilde{N}_k\|^2_{\nabla^2\mathcal{L}_n(\theta^{(k)})}\right).$$

Thus, we obtain

$$\begin{aligned}
\mathcal{L}_n(\theta^{(k+1)}) \le \mathcal{L}_n(\theta^{(k)}) &- \left(\eta - \eta^2\tau_0\right)\|\{\nabla^2\mathcal{L}_n(\theta^{(k)})\}^{-1}\nabla\mathcal{L}_n(\theta^{(k)})\|^2_{\nabla^2\mathcal{L}_n(\theta^{(k)})} \\
&+ \eta\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2\|\tilde{N}_k\|_2 + \eta^2\tau_0\|\tilde{N}_k\|^2_{\nabla^2\mathcal{L}_n(\theta^{(k)})}. \qquad (63)
\end{aligned}$$

Furthermore, minimizing both sides of the lower bound (16) in Lemma 15 with respect to θ_1 gives

$$\mathcal{L}_n(\hat{\theta}) \ge \mathcal{L}_n(\theta_2) - \frac{\tau_0}{2}\|\{\nabla^2\mathcal{L}_n(\theta_2)\}^{-1}\nabla\mathcal{L}_n(\theta_2)\|^2_{\nabla^2\mathcal{L}_n(\theta_2)}.$$

Plugging in $\theta_2 = \theta^{(k)}$, we then obtain

$$\|\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)})\|^2_{\nabla^2 \mathcal{L}_n(\theta^{(k)})} \geq \frac{2}{\tau_0} \Big(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \Big).$$

Combining this bound with inequality (63) and using the fact that $\eta \leq \frac{1}{2\tau_0}$ then gives

$$\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\theta^{(k)}) - \frac{\eta}{\tau_0} \Big(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \Big) + \eta \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \|\tilde{N}_k\|_2 + \eta^2 \tau_0 \|\tilde{N}_k\|^2_{\nabla^2 \mathcal{L}_n(\theta^{(k)})}.$$

Hence, using the fact that $\max_{k \leq K} \|\tilde{N}_k\|_2 \leq \tilde{r}_{priv}$ from Lemma 28, we have

$$\begin{aligned} \mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) &\leq \Big(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \Big) - \frac{\eta}{\tau_0} \Big(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \Big) + \rho \\ &= \Big(1 - \frac{\eta}{\tau_0} \Big) \Big(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \Big) + \rho, \end{aligned} \tag{64}$$

where $\rho = \eta B \tilde{r}_{priv} + \eta^2 \bar{B} \tau_0 \tilde{r}^2_{priv}$. (Note that $\tau_0 \geq 1$, so $\eta \leq \frac{1}{\tau_0}$ implies that $\eta \leq \tau_0$, as well.) Iterating the bound (64), we obtain

$$\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) \leq \Big(1 - \frac{\eta}{\tau_0} \Big)^k \Big(\mathcal{L}_n(\theta^{(0)}) - \mathcal{L}_n(\hat{\theta}) \Big) + \frac{\rho \tau_0}{\eta} \Big(1 - \Big(1 - \frac{\eta}{\tau_0} \Big) \Big)^k,$$

as claimed.

E.3.2 Supporting lemmas

The following result is an analogue of Lemma 17 for self-concordant functions, showing that a parameter θ must be close to $\hat{\theta}$ if the suboptimality gap is small.

Lemma 27. *Suppose $\mathcal{L}_n$ is $(\gamma, 2)$-self-concordant. For $\theta \in \mathbb{R}$, define $\Delta = \mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta})$. Then*

$$\|\theta - \hat{\theta}\|_2 \leq g^{-1} \bigg(\frac{\gamma^2 \Delta}{\lambda_{\min}(\nabla^2 \mathcal{L}_n(\hat{\theta}))} \bigg),$$

where $g(t) := e^{-t} + t - 1$ for $t > 0$.

Proof. Let $R = \|\hat{\theta} - \theta\|_2$. Applying Lemma 14 with $x = \hat{\theta}$ and $y = \theta$, we have

$$\frac{\exp(-\gamma R) + \gamma R - 1}{(\gamma R)^2} \cdot R^2 \lambda_{\min}(\nabla^2 \mathcal{L}_n(\hat{\theta})) \leq \mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta}).$$

Rearranging then gives

$$\exp(-\gamma R) + \gamma R - 1 \leq \frac{\gamma^2 (\mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta}))}{\lambda_{\min}(\nabla^2 \mathcal{L}_n(\hat{\theta}))},$$

from which the result clearly follows by applying g^{-1}. □

Lemma 28. *Suppose $\mathcal{L}_n$ is $(\gamma, 2)$-self-concordant, and let Δ_0 and R_0 be as defined in the statement of Proposition 2. Also suppose $\eta \leq \frac{4\tau^2_{1,R_0}}{\bar{B}^2}$ and the sample size satisfies $n = \Omega\Big(\frac{\sqrt{Kp \log(Kp/\xi)}}{\mu} \Big)$. With probability at least $1 - \xi$, the noisy Newton updates (8) can be rewritten as*

$$\theta^{(k+1)} = \theta^{(k)} - \eta \Big\{ \nabla^2 \mathcal{L}_n(\theta^{(k)})) \Big\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)}) + \eta \tilde{N}_k,$$

where

$$\max_{k<K} \|\tilde{N}_k\|_2 \leq C \frac{\bar{B} B \sqrt{Kp \log(Kp/\xi)}}{\mu n \tau_{1,R_0}^2} := \tilde{r}_{priv},$$

and C is a positive constant. Furthermore, $\theta^{(k)} \in \mathcal{B}_{R_0}(\hat{\theta})$ for all $k \leq K$.

Proof. Using a similar argument as in Lemma 18, we will prove by induction that $\|\theta^{(k)} - \hat{\theta}\|_2 \leq R_0$ and $\mathcal{L}_n(\theta^{(k)}) \leq \mathcal{L}_n(\hat{\theta}) + \Delta_0$, for all $k \leq K$. First note that by Lemmas 4 and 5 and a union bound, we have $\max_{k<K} \|Z_k\|_2 \leq 4\sqrt{p} + 2\sqrt{2\log(2K/\xi)}$ and $\max_{k<K} \|W_k\|_2 \leq \sqrt{2p\log(4Kp/\xi)}$, with probability at least $1-\xi$. We will assume these bounds hold and argue deterministically for the rest of the proof.

The base case $k = 0$ holds by Lemma 27. For the inductive step, suppose the bounds hold for $\{1, \ldots, k\}$. Note that $\| \{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} \|_2 \leq \frac{1}{2\tau_{1,R_0}}$, since we have local strong convexity guaranteed by Lemma 16. As in the proof of inequality (52) in Lemma 23, the series expansion (48) and the bounds on $\|Z_k\|_2$ and $\|W_k\|_2$ imply that for all $k < K$, we have

$$\|\tilde{N}_k\|_2 \leq C_0 \frac{B\sqrt{K}\{\sqrt{p} + \sqrt{\log(2K/\xi)}\} + \bar{B} B \tau_{1,R_0}^{-1} \sqrt{Kp\log(Kp/\xi)}}{\mu n \tau_{1,R_0}} = \tilde{r}_{priv}. \tag{65}$$

Using smoothness, we now write

$$\begin{aligned}
\mathcal{L}_n(\theta^{(k+1)}) &= \mathcal{L}_n\left(\theta^{(k)} - \eta\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)}) + \eta \tilde{N}_k\right) \\
&\leq \mathcal{L}_n(\theta^{(k)}) - \eta \langle \nabla \mathcal{L}_n(\theta^{(k)}), \{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)}) - \tilde{N}_k \rangle \\
&\quad + \frac{\bar{B}\eta^2}{2} \|\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)}) - \tilde{N}_k\|_2^2 \\
&= \mathcal{L}_n(\theta^{(k)}) + \frac{\bar{B}\eta^2}{2} \|\tilde{N}_k\|_2^2 - \|\nabla \mathcal{L}_n(\theta^{(k)})\|^2_{\eta\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} - \frac{\bar{B}\eta^2}{2}\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-2}} \\
&\quad - \eta \langle \nabla \mathcal{L}_n(\theta^{(k)}) + \bar{B}\eta \{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)}), \tilde{N}_k \rangle.
\end{aligned} \tag{66}$$

Note that by the assumption $\eta \leq \frac{4\tau_{1,R_0}^2}{\bar{B}^2}$ and the facts that $\|\nabla^2 \mathcal{L}_n(\theta^{(k)})\|_2 \leq \bar{B}$ and $\|\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1}\|_2 \leq \frac{1}{2\tau_{1,R_0}}$, we have

$$\begin{aligned}
&\lambda_{\min}\left(\eta\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} - \frac{\bar{B}\eta^2}{2}\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-2}\right) \\
&\geq \eta \lambda_{\min}\left(\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1}\right) - \lambda_{\max}\left(\frac{\bar{B}\eta^2}{2}\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-2}\right) \\
&\geq \frac{\eta}{\bar{B}} - \frac{\bar{B}\eta^2}{2} \cdot \frac{1}{4\tau_{1,R_0}^2} \\
&\geq \frac{\eta}{2\bar{B}},
\end{aligned}$$

implying that

$$\|\nabla \mathcal{L}_n(\theta^{(k)})\|^2_{\eta\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1} - \frac{\bar{B}\eta^2}{2}\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-2}} \geq \frac{\eta}{2\bar{B}} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2.$$

Furthermore, we have

$$\begin{aligned}
&\left|\langle \nabla \mathcal{L}_n(\theta^{(k)}) + \bar{B}\eta\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1}\nabla \mathcal{L}_n(\theta^{(k)}), \tilde{N}_k\rangle\right| \\
&\quad \le \|\tilde{N}_k\|_2 \left\|\nabla \mathcal{L}_n(\theta^{(k)}) + \bar{B}\eta\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1}\nabla \mathcal{L}_n(\theta^{(k)})\right\|_2 \\
&\quad \le \|\tilde{N}_k\|_2 \left(B + \frac{\bar{B}\eta B}{2\tau_{1,R_0}}\right),
\end{aligned}$$

Hence, by inequality (66), the descent condition $\mathcal{L}_n(\theta^{(k+1)}) \le \mathcal{L}_n(\theta^{(k)})$ can be guaranteed as long as

$$B\eta\left(1 + \frac{\bar{B}\eta}{2\tau_{1,R_0}}\right)\|\tilde{N}_k\|_2 + \frac{\bar{B}\eta^2}{2}\|\tilde{N}_k\|_2^2 \le \frac{\eta}{2\bar{B}}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2,$$

which is in turn guaranteed if

$$\|\tilde{N}_k\| \le \frac{-B(1+\frac{\eta\bar{B}}{2\tau_{1,R_0}}) + \sqrt{B^2(1+\frac{\eta\bar{B}}{2\tau_{1,R_0}})^2 + \eta\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2}}{\eta\bar{B}},$$

or equivalently,

$$\begin{aligned}
\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 &\ge \sqrt{\frac{\left(\eta\bar{B}\|\tilde{N}_k\|_2 + B\left(1+\frac{\eta\bar{B}}{2\tau_{1,R_0}}\right)\right)^2 - B^2\left(1+\frac{\eta\bar{B}}{2\tau_{1,R_0}}\right)^2}{\eta}} \\
&= \sqrt{\eta\bar{B}^2\|\tilde{N}_k\|_2^2 + 2\bar{B}B\left(1+\frac{\eta\bar{B}}{2\tau_{1,R_0}}\right)\|\tilde{N}_k\|_2}.
\end{aligned}$$

By inequality (65), this last bound is ensured if

$$\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \ge \sqrt{\eta\bar{B}^2\tilde{r}_{priv}^2 + 2\bar{B}B\left(1+\frac{\eta\bar{B}}{2\tau_{1,R_0}}\right)\tilde{r}_{priv}} := \check{r}_{priv}. \tag{67}$$

In summary, if $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \ge \check{r}_{priv}$, then $\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) \le \mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \le \Delta_0$, so also $\theta^{(k+1)} \in \mathcal{B}_{R_0}(\hat{\theta})$ by Lemma 27.

If instead $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 < \check{r}_{priv}$, we use an alternative argument to conclude that $\theta^{(k+1)} \in \mathcal{B}_{R_0}(\hat{\theta})$: Note that $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \ge 2\tau_{1,R_0}\|\theta^{(k)} - \hat{\theta}\|_2$, so the triangle inequality and a sufficiently large n (to be specified below) show that

$$\begin{aligned}
\|\theta^{(k+1)} - \hat{\theta}\|_2 &\le \|\theta^{(k)} - \hat{\theta}\|_2 + \|\theta^{(k+1)} - \theta^{(k)}\|_2 \\
&\le \frac{1}{2\tau_{1,R_0}}\check{r}_{priv} + \eta\|\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\}^{-1}\nabla \mathcal{L}_n(\theta^{(k)}) - \tilde{N}_k\|_2 \\
&\le \frac{1}{2\tau_{1,R_0}}\check{r}_{priv} + \frac{\eta\check{r}_{priv}}{2\tau_{1,R_0}} + \eta\tilde{r}_{priv} \\
&\le \min\left\{\sqrt{\frac{\Delta_0}{\tau_2}}, R_0\right\}.
\end{aligned} \tag{68}$$

Therefore, we conclude that if $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 < \check{r}_{priv}$, then $\theta^{(k+1)} \in \mathcal{B}_{R_0}(\hat{\theta})$, as well. Furthermore, by smoothness, inequality (68) also gives

$$\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) \le \tau_2\|\theta^{(k+1)} - \hat{\theta}\|_2^2 \le \tau_2 \cdot \frac{\Delta_0}{\tau_2} \le \Delta_0.$$

From the definitions of $\tilde{r}_{priv}$ and $\check{r}_{priv}$ in inequalities (65) and (67), we can check that the last inequality in the chain (68) can be ensured via the assumed minimum sample size condition $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$. This completes the inductive step. □

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pages 308–318, 2016.

J. Acharya, Z. Sun, and H. Zhang. Differentially private testing of identity and closeness of discrete distributions. Advances in Neural Information Processing Systems, 31:6878–6891, 2018.

M. Avella Medina. Privacy-preserving parametric inference: A case for robust statistics. Journal of the American Statistical Association, (116):969–983, 2021.

J. Awan and A. Slavković. Differentially private uniformly most powerful tests for binomial data. Advances in Neural Information Processing Systems, 2018:4208–4218, 2018.

F. Bach. Self-concordant analysis for logistic regression. Electronic Journal of Statistics, 4:384–414, 2010.

B. Balle, P. Kairouz, B. McMahan, O. D. Thakkar, and A. Thakurta. Privacy amplification via random check-ins. Advances in Neural Information Processing Systems, 33, 2020.

R. F. Barber and J. Duchi. Privacy: A few definitional aspects and consequences for minimax mean-squared error. In 53rd IEEE Conference on Decision and Control, pages 1365–1369. IEEE, 2014.

A. F. Barrientos, J. P. Reiter, A. Machanavajjhala, and Y. Chen. Differentially private significance tests for regression coefficients. Journal of Computational and Graphical Statistics, 28(2):440–453, 2019.

R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, pages 464–473. IEEE, 2014.

R. Bassily, V. Feldman, K. Talwar, and A. Thakurta. Private stochastic convex optimization with optimal rates. arXiv preprint arXiv:1908.09970, 2019.

S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, 2004.

Z. Bu, J. Dong, Q. Long, and W. J. Su. Deep learning with Gaussian differential privacy. Harvard Data Science Review, 2020(23), 2020.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. Foundations and Trends® in Machine Learning, 8(3-4):231–357, 2015.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. Annals of Statistics, 49(5):2825–2850, 2019.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. arXiv preprint arXiv:2011.03900, 2020.

E. Cantoni and E. Ronchetti. Robust inference for generalized linear models. Journal of the American Statistical Association, 96(455):1022–1030, 2001.

Karan Chadha, John Duchi, and Rohith Kuditipudi. Private confidence sets. In NeurIPS 2021 Workshop Privacy in Machine Learning, 2021.

K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In Advances in Neural Information Processing Systems, volume 22, pages 289–296. Citeseer, 2008.

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. Journal of Machine Learning Research, 12(3), 2011.

Christian Covington, Xi He, James Honaker, and Gautam Kamath. Unbiased statistical estimation and valid confidence intervals under differential privacy. arXiv preprint arXiv:2110.14465, 2021.

A. d'Aspremont. Smooth optimization with approximate gradient. SIAM Journal on Optimization, 19(3):1171–1183, 2008.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming, 146(1):37–75, 2014.

B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 3574–3583, 2017.

J. Dong, A. Roth, and W. J. Su. Gaussian differential privacy. Journal of the Royal Statistical Society: Series B (to appear), 2021.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation. Journal of the American Statistical Association, 113(521):182–201, 2018.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407, 2014.

C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing, pages 11–20, 2014.

Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pages 1054–1067, 2014.

V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pages 439–449, 2020.

R. Fraiman, V. J. Yohai, and R. H. Zamar. Optimal robust M-estimates of location. Annals of Statistics, pages 194–223, 2001.

M. Gaboardi, H. Lim, R. Rogers, and S. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In International Conference on Machine Learning, pages 2111–2120. PMLR, 2016.

S. Garfinkel, J. M. Abowd, and C. Martindale. Understanding database reconstruction attacks on public data. Communications of the ACM, 62(3):46–53, 2019.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. SIAM Journal on Optimization, 22 (4):1469–1492, 2012.

F. Hampel, C. Hennig, and E. Ronchetti. A smoothing principle for the Huber and other location M-estimators. Computational Statistics & Data Analysis, 55(1):324–337, 2011.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. Robust Statistics: The Approach Based on Influence Functions, volume 196. John Wiley & Sons, 1986.

P. J. Huber. Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1):73–101, 1964.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 221–233. University of California Press, 1967.

P. J. Huber and E. Ronchetti. Robust Statistics. Wiley, New York, second edition, 2009.

R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang. Towards practical differentially private convex optimization. In 2019 IEEE Symposium on Security and Privacy (SP), pages 299–316. IEEE, 2019.

P. Jain and A. Guha. Thakurta. (near) dimension independent risk bounds for differentially private learning. In International Conference on Machine Learning, pages 476–484. PMLR, 2014.

P. Jain, P. Kothari, and A. Thakurta. Differentially private online learning. In Conference on Learning Theory, pages 24–1. JMLR Workshop and Conference Proceedings, 2012.

S. P. Karimireddy, S. U. Stich, and M. Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. arXiv preprint arXiv:1806.00413, 2018.

V. Karwa and S. Vadhan. Finite sample differentially private confidence intervals. arXiv preprint arXiv:1711.03908, 2017.

D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In Conference on Learning Theory, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

Tejas Kulkarni, Joonas Jälkö, Antti Koskela, Samuel Kaski, and Antti Honkela. Differentially private Bayesian inference for generalized linear models. In International Conference on Machine Learning, pages 5838–5849. PMLR, 2021.

H. R. Künsch, L. A. Stefanski, and R. J. Carroll. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. Journal of the American Statistical Association, 84(406):460–466, 1989.

J. Lee and D. Kifer. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1656–1665, 2018.

J. Lei. Differentially private m-estimators. Advances in Neural Information Processing Systems, 2011:361–369, 2011.

P. Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. The Annals of Statistics, 45(2):866–896, 2017.

P. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. The Journal of Machine Learning Research, 16(1):559–616, 2015.

P. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62:22–31, 2014.

Yurii Nesterov. Lectures on convex optimization, volume 137. Springer, second edition, 2018.

D. M. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. Electronic Journal of Statistics, 15(1):326–391, 2021.

Víctor Peña and Andrés F Barrientos. Differentially private methods for managing model uncertainty in linear regression models. arXiv preprint arXiv:2109.03949, 2021.

A. Rajkumar and S. Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In Artificial Intelligence and Statistics, pages 933–941. PMLR, 2012.

P. Rigollet and J.-C. Hütter. High-dimensional statistics. Lecture Notes for Course 18S997, 2017.

R. Rogers and D. Kifer. A new class of private chi-square hypothesis tests. In Artificial Intelligence and Statistics, pages 991–1000. PMLR, 2017.

Terrance D Savitsky, Matthew R Williams, and Jingchen Hu. Bayesian pseudo posterior mechanism under differential privacy. arXiv preprint arXiv:1909.11796, 2019.

O. Sheffet. Differentially private ordinary least squares. In International Conference on Machine Learning, pages 3105–3114. PMLR, 2017.

Aleksandra Slavkovic and Roberto Molinari. Perturbed M-estimation: A further investigation of robust statistics for differential privacy. arXiv preprint arXiv:2108.08266, 2021.

S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In 2013 IEEE Global Conference on Signal and Information Processing, pages 245–248. IEEE, 2013.

T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: A recipe for Newton-type methods. Mathematical Programming, 178(1-2):145–213, 2019.

T. Sun, I. Necoara, and Q. Tran-Dinh. Composite convex optimization with global and local inexact oracles. Computational Optimization and Applications, pages 1–56, 2020.

K. Talwar, A. Thakurta, and L. Zhang. Nearly-optimal private LASSO. In Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2, pages 3025–3033, 2015.

J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. arXiv preprint arXiv:1709.02753, 2017.

J. A. Tropp. An introduction to matrix concentration inequalities. Foundations and Trends® in Machine Learning, 8(1-2):1–230, 2015.

C. Uhler, A. Slavković, and S. E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. The Journal of Privacy and Confidentiality, 5(1):137, 2013.

D. Wang, M. Ye, and J. Xu. Differentially private empirical risk minimization revisited: Faster and more general. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 2719–2728, 2017a.

X. Wang, S. Ma, D. Goldfarb, and W. Liu. Stochastic quasi-Newton methods for nonconvex stochastic optimization. SIAM Journal on Optimization, 27(2):927–956, 2017b.

Y. Wang, S. Fienberg, and A. Smola. Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In International Conference on Machine Learning, pages 2493–2502. PMLR, 2015.

Y. Wang, D. Kifer, and J. Lee. Differentially private confidence intervals for empirical risk minimization. Journal of Privacy and Confidentiality, 9(1), 2019.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. Journal of the American Statistical Association, 105(489):375–389, 2010.

F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg. Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In International Conference on Privacy in Statistical Databases, pages 170–184. Springer, 2014.