

Instability, Computational Efficiency and Statistical Accuracy

Nhat Ho*

*Department of Statistics and Data Sciences
University of Texas, Austin*

MINHNHAT@UTEXAS.EDU

Koulik Khamaru*

*Department of Statistics
Rutgers University*

KK1241@STAT.RUTGERS.EDU

Raaz Dwivedi*

*Department of Operations Research & Information Engineering
Cornell*

RD597@CORNELL.EDU

Martin J. Wainwright

*Department of EECS, Department of Mathematics
Massachusetts Institute of Technology*

WAINWRIGWORK@GMAIL.COM

Michael I. Jordan

*Department of EECS, Department of Statistics
University of California, Berkeley
INRIA, Paris*

JORDAN@CS.BERKELEY.EDU

Bin Yu

*Department of EECS, Department of Statistics
University of California, Berkeley*

BINYU@BERKELEY.EDU

Editor: Prateek Jain

Abstract

Many statistical estimators are defined as the fixed point of a data-dependent operator, with estimators based on minimizing a cost function being an important special case. The limiting performance of such estimators depends on the properties of the population-level operator in the idealized limit of infinitely many samples. We develop a general framework that yields bounds on statistical accuracy based on the interplay between the deterministic convergence rate of the algorithm at the population level, and its degree of (in)stability when applied to an empirical object based on n samples. Using this framework, we analyze both stable forms of gradient descent and some higher-order and unstable algorithms, including Newton’s method and its cubic-regularized variant, as well as the EM algorithm. We provide applications of our general results to several concrete classes of models, including Gaussian mixture estimation, non-linear regression models, and informative non-response models. We exhibit cases in which an unstable algorithm can achieve the same statistical accuracy as a stable algorithm in exponentially fewer steps—namely, with the number of iterations being reduced from polynomial to logarithmic in sample size n .

. * Raaz Dwivedi, Nhat Ho, and Koulik Khamaru contributed equally to this work.

1. Introduction

The interplay between the stability and computational efficiency of optimization algorithms has long been a fundamental problem in statistics and machine learning. The stability of the algorithm, a classical desideratum, is often believed to be a necessity for obtaining efficient statistical estimators. Such a belief rules out the use of a variety of faster algorithms due to their instability. This paper shows that this popular belief can be misleading: the situation is more subtle in that there are various settings in which unstable algorithms may be preferable to their stable counterparts.

Recent years have seen a significant body of work involving performance of various machine-learning algorithms when applied to statistical estimation problems. Examples include sparse signal recovery (Hale et al., 2008; Garg and Khandekar, 2009; Beck and Teboulle, 2009; Becker et al., 2011), more general forms of M-estimation (Agarwal et al., 2012; Zhang and Zhang, 2012; Loh and Wainwright, 2015), principal component analysis (Amini and Wainwright, 2008; Ma, 2013; Yuan and Zhang, 2013), regression with concave penalties (Loh and Wainwright, 2015; Wang et al., 2014), phase retrieval problems retrieval (e.g., (Candès et al., 2012, 2015; Chen and Wainwright, 2015; Zhang et al., 2017; Chen et al., 2018b)), and mixture model estimation (Balakrishnan et al., 2017; Yang et al., 2017; Cai et al., To Appear; Yi and Caramanis, 2015).

A unifying theme in these works is to study, in a finite-sample setting, the computational efficiency of different algorithms and the statistical accuracy of the resulting estimates. For estimators based on solving optimization problems that are convex, standard algorithms and theory can be applied. However, many modern estimators arise from non-convex optimization problems, in which case the associated algorithms become more complex to understand. But evidence is accumulating for the practical and theoretical advantages of such algorithms. For instance, the paper (Agarwal et al., 2012) established the fast convergence of projected gradient descent (GD) for high-dimensional signal recovery in a weakly convex setting, whereas the papers (Loh and Wainwright, 2015; Wang et al., 2014) provided similar guarantees for a class of non-convex learning problems. Other work has demonstrated fast convergence of the truncated power method for PCA (Yuan and Zhang, 2013), analyzed the behavior of projected gradient methods for low-rank matrix recovery (Chen and Wainwright, 2015), and characterized the behavior of gradient descent for phase-retrieval problems (Chen et al., 2018b). Additionally, there is also a recent line on work on the fast convergence of EM for various types of mixture models (Balakrishnan et al., 2017; Yang et al., 2017; Cai et al., To Appear). Finally, there is a line of work (Hardt et al., 2016; Chen et al., 2018a; Kuzborskij and Lampert, 2018; Charles and Papailiopoulos, 2018) that provides statistical error bounds for generic machine learning problems (with certain assumptions on loss functions) in terms of estimators obtained via iterative optimization algorithms (e.g., stochastic gradient methods).

1.1 Population-to-sample or stability-based analysis

The analysis in these works falls into two distinct categories. The first is a *direct analysis*, in which one directly characterizes the behavior of the iterates of the algorithm on the finite-sample objective. A long line of papers has used the direct approach (e.g., (Agarwal et al., 2012; Loh and Wainwright, 2015; Wang et al., 2014; Zhang and Zhang, 2012; Yuan

and Zhang, 2013)) to demonstrate that certain optimization algorithms converge at geometric rates to a local neighborhood of the true parameter, with the radius proportional to the statistical minimax risk. The second kind of analysis is more indirect and can be referred to as *population-to-sample analysis* or *stability-based analysis* where one analyzes the algorithmic convergence of population-level iterates, and derives statistical errors for the sample-level updates via uniform laws for stability/perturbation bounds. These approaches have been used to analyze the performance of EM and its variants in several statistical settings, see the papers (Balakrishnan et al., 2017; Cai et al., To Appear; Yang et al., 2017; Yi and Caramanis, 2015; Dwivedi et al., 2020a,b) and the references therein. In general settings, it has been used to derive statistical errors for iterates from stochastic optimization methods (Hardt et al., 2016; Chen et al., 2018a; Kuzborskij and Lampert, 2018; Charles and Papailiopoulos, 2018).

The contributions of this paper build upon the stability-based analysis, so let us discuss it in a little more detail. Let F and F_n denote the operators that define the iterates at the population level, corresponding to the idealized limit of an infinite sample size, and sample-level based on a dataset of size n . Suppose θ^* denotes the parameter of interest, such that the population-level iterates defined as $\theta^t = F(\theta^{t-1})$ for $t = 1, 2, \dots$ with initialization θ^0 , i.e., $\theta^t = F^t(\theta^0)$, converge to θ^* as $t \rightarrow \infty$. Of interest is to characterize the best possible estimate of θ^* obtained from the sample-based (noisy) iterates, defined as $\theta_n^t = F_n^t(\theta^0)$ (with initialization θ^0), and possibly characterize the change in the error $\|F_n^t(\theta^0) - \theta^*\|$ as a function of the iteration t and the sample size n . The population-to-sample or the stability analysis proceeds by using the following decomposition:

$$F_n^t(\theta^0) - \theta^* = \underbrace{F^t(\theta^0) - \theta^*}_{=:\varepsilon_{\text{opt}}^t} + \underbrace{F_n^t(\theta^0) - F^t(\theta^0)}_{=:\varepsilon_{\text{stab}}^t}. \quad (1)$$

Given this decomposition, the analysis proceeds in two steps:

- The first step is a deterministic convergence analysis of the algorithm to the true parameter at the population-level, namely, obtain a control on the *optimization error* $\varepsilon_{\text{opt}}^t$ as a function of t .
- The second step is to perform a stability analysis of the difference between the population and the sample-based iterates, namely, obtain a control on the *perturbation/stability error* $\varepsilon_{\text{stab}}^t$ as a function of t .

The ultimate convergence guarantee—what statistical error can be achieved with the sample-based operator F_n , and in how many iterations—is then derived based on the interplay between the two errors in equation (1), namely, $\varepsilon_{\text{opt}}^t$ and $\varepsilon_{\text{stab}}^t$.

The ERM-based approach: We remark that the decomposition in equation (1) is different from that used when invoking the uniform laws for the empirical risk minimizer (ERM). Assuming the sample-based iterates converges to the ERM, i.e., $\lim_{t \rightarrow \infty} F_n^t(\theta^0) = \hat{\theta}_{\text{ERM}}$,

the typical decomposition in the ERM-based approach is given by

$$F_n^t(\theta^0) - \theta^* = \underbrace{F_n^t(\theta^0) - \hat{\theta}_{\text{ERM}}}_{=:\varepsilon_{\text{opt-sample}}^t} + \underbrace{\hat{\theta}_{\text{ERM}} - \theta^*}_{=:\varepsilon_{\text{unif-gen}}}.$$

Here the first term in the RHS corresponds to the *optimization error at the sample-level* at iteration t and the second term corresponds to the (iteration-independent) *uniform generalization bound*. Depending on the application, a precise characterization of either of these terms can be non-trivial; moreover, applying uniform bounds to control the term $\varepsilon_{\text{unif-gen}}$ may lead to bounds that are overly loose. In such settings, the population-to-sample or stability-based analysis can prove to be a useful alternative.

1.2 Past works focus on stable methods

Most of the past work with the population-to-sample analysis has focused on algorithms whose updates are *stable*, meaning that the perturbation error between sample-level and population-level iterates decays to zero as the iterates approach the true parameter. For example, the papers (Balakrishnan et al., 2017; Cai et al., To Appear; Yang et al., 2017; Yi and Caramanis, 2015) used this framework for problems where the population updates converge at a geometric rate to the true parameter, and iterates based on n samples yield an estimate within $n^{-1/2}$ of the true parameter. On the other hand, other papers (Dwivedi et al., 2020a,b) have shown that with over-specified Gaussian mixtures, the EM algorithm, which is a stable algorithm, takes a large number of steps to find an estimate whose statistical error is of order $n^{-1/4}$ or $n^{-1/8}$. Although for those problems the larger final statistical error of EM is minimax optimal, several natural questions remained unanswered: Can an algorithm converge to a statistically optimal estimate in significantly fewer steps than EM for over-specified mixtures? Moreover, will the faster algorithm continue to be stable? Besides the analysis in recent works (Dwivedi et al., 2020a,b) relied heavily on the facts that the EM updates had closed-form analytical expressions. To our best knowledge, general statistical guarantees for a generic stable or unstable algorithm (without a closed-form expression) when the algorithmic convergence is slow, are not present in the literature.

In past work, Chen et al. (2018b) provided a trade-off between stability and number of iterations to converge. In particular, they showed that the minimax error of a problem class forces a trade-off between the two errors in equation (1), $\varepsilon_{\text{opt}}^t$ and $\varepsilon_{\text{stab}}^t$, for any iterative algorithm used for solving it. In simple words, given the minimax error, an algorithm that converges quickly is necessarily unstable, and conversely, a stable algorithm cannot converge quickly. Their work, however, did not address the following converse questions: Under what conditions does an algorithm, either stable or unstable, achieve a statistically optimal rate? When is an unstable algorithm to be preferred to a stable counterpart?

Such questions about the trade-off between stability, computational efficiency and the statistical error upon convergence are of special interest for singular problems in which the

. There is a subtle difference in the definition of (in)stability used in Chen et al.’s work (Chen et al., 2018a) compared to ours. In their work, stability refers to a *slow* growth in the error $\|F^t(\theta) - F_n^t(\theta)\|$ with number of iterations t , where slow is defined in a relative sense with other methods. In our case, we use stability for the settings when $\|F(\theta) - F_n(\theta)\|$ decreases with $\|\theta - \theta^*\|$ as $\theta \rightarrow \theta^*$.

Fisher information matrices are degenerate. Singular problems appear in a wide range of statistical settings, including mode estimation (Chernoff, 1964), robust regression (Rousseeuw, 1984), stochastic utility models (Manski, 1975), informative non-response in missing data (Heckman, 1976; Diggle and Kenward, 1994), high-dimensional linear regression (Hastie et al., 2015), and over-specified mixture models (Chen, 1995; Rousseau and Mengersen, 2011; Nguyen, 2013). Several papers have shown that maximum likelihood estimates for singular problems have much lower accuracy than the classical parametric rate $n^{-1/2}$; problems that exhibit slow rates of this type include stochastic frontier models (Lee and Chesher, 1986; Lee, 1993), certain classes of parametric models (Rotnitzky et al., 2000), and in strongly or weakly identifiable mixture models (Chen, 1995; Nguyen, 2013; Ho and Nguyen, 2016). Nevertheless, the computational aspects of parameter estimation and the trade-offs with stability in such models are not well understood at the current time.

1.3 Our contributions

This paper lays out a general framework to address the questions raised above. Making use of the population-to-sample approach and a generalization of the localization argument from our previous works (Dwivedi et al., 2020a,b), we derive tight bounds on the statistical error of the final iterate produced by an algorithm. The final error and the number of steps taken depend on two things: (i) the rate of convergence of the corresponding population-level iterates, and (ii) the (in)stability of the sample-level iterates with respect to that at the population-level. As a first contribution, our statistical guarantees for slowly converging stable algorithms and (fast/slow converging) unstable algorithms complement the findings of Balakrishnan et al. (2017) for fast converging stable algorithms (Theorems 5 and 6). We provide an overview of these general results in Table 1.

The second contribution extends the work of Chen et al. (2018a) by showing how the final statistical errors achieved by stable and unstable algorithms can be used to directly compare and contrast the (dis)advantages between the two (Section 4). Our third contribution is an explicit demonstration of the fact that unstable methods can converge in significantly fewer steps when compared to stable methods, while still yielding statistically optimal estimates (Corollaries 7, 8 and 9). In particular, applying our framework to three estimation problems—single index models with known link, informative non-response models, and Gaussian mixture models—we show that while the (unstable) Newton method converges after on the order of $\log n$ steps, there is some $q > 0$ such that gradient descent—which we show to be a stable method—takes on the order of n^q steps. Finally, we also establish that our guarantees are unimprovable in general on both statistical accuracy, and the iteration complexity.

Organization: The remainder of our paper is organized as follows. We begin in Section 2 with simulations that illustrate the phenomena to be investigated in this paper. We then introduce some definitions and discuss different properties of the sample and population operators. Section 3 is devoted to statements of our general computational and statistical guarantees with detailed proofs presented in Appendix A. In Section 4, we apply our general results to demonstrate the trade-off between stable and unstable methods for several examples. We conclude with a discussion of potential future work in Section 5. Proofs of supporting lemmas and technical results are provided in the appendices.

Operator Properties	Optimization Rate	Stability	Iterations for convergence	Statistical error on convergence
General expressions				
Fast, stable (Balakrishnan et al., 2017)	FAST(κ)	STA(0)	$\log(1/\varepsilon(n, \delta))$	$\varepsilon(n, \delta)$
Slow, stable (Thm. 5)	SLOW(β)	STA(γ)	$\varepsilon(n, \delta^*)^{-\frac{1}{1+\beta-\gamma\beta}}$	$[\varepsilon(n, \delta^*)]^{\frac{\beta}{1+\beta-\gamma\beta}}$
Fast, unstable (Thm. 6)	FAST(κ)	UNS(γ)	$\log(1/\varepsilon(n, \delta))$	$[\varepsilon(n, \delta)]^{\frac{1}{1+ \gamma }}$
Slow, unstable (Thm. 6)	SLOW(β)	UNS(γ)	$[\varepsilon(n, \delta)]^{-\frac{1}{1+\beta}}$	$[\varepsilon(n, \delta)]^{\frac{\beta}{1+\beta+ \gamma \beta}}$
Specific examples $\varepsilon(n, \delta) = \log(1/\delta)/\sqrt{n}$				
FAST(κ), STA(0)	$e^{-\kappa t}$	$\frac{1}{\sqrt{n}}$	$\log n$	$n^{-1/2}$
SLOW($\frac{1}{2}$), STA(1)	$\frac{1}{\sqrt{t}}$	$\frac{r}{\sqrt{n}}$	$n^{1/2}$	$n^{-1/4}$
FAST(κ), UNS(-1)	$e^{-\kappa t}$	$\frac{1}{r\sqrt{n}}$	$\log n$	$n^{-1/4}$
SLOW($\frac{1}{2}$), UNS(-1)	$\frac{1}{\sqrt{t}}$	$\frac{1}{r\sqrt{n}}$	$n^{1/3}$	$n^{-1/8}$

Table 1. A high-level overview of our results. The notation in the problem set-up (columns 2 and 3) is formalized in Section 2.2, and the formal results (columns 4 and 5) are discussed in Section 3. In the top panel, we provide general expressions from our results, and in the bottom panel, we provide some explicit expressions for few specific settings. The second and third columns respectively denote the optimization and stability properties of the operator, and the last two columns provide the expressions for iterations for convergence, and the final statistical errors of the estimate returned the sample-based (noisy) operator (see equations (14) for the definition of δ^*). For the bottom panel, we use $\beta = \frac{1}{2}, \gamma = 0, -1$ with the noise function $\varepsilon(n, \delta) = \log(1/\delta)/\sqrt{n}$. For brevity, we omit log-factors (in n, δ) and universal constants for the expressions in the bottom panel.

Notation: A few remarks on notation: for a pair of sequences $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, we write $a_n \gtrsim b_n$ or $a_n = \Omega(b_n)$ to mean that there is a universal constant c such that $a_n \geq cb_n$ for all $n \geq 1$. We write $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. We use $\lceil x \rceil$ to denote the smallest integer greater than or equal to x for any $x \in \mathbb{R}$. In the paper, we use c, c', c_i, c'_i when $i \geq 1$ to denote the universal constants. Note that the values of universal constants may change from line to line. Finally, for our operator notation, we use the subscript n to distinguish a sample-based operator (e.g., $F_n, G_n^{\text{NM}}, M_n^{\text{GA}}$) from its corresponding population-based analog (respectively $F, G^{\text{NM}}, M^{\text{GA}}$).

2. Motivation and problem set-up

We begin in Section 2.1 by motivating the analysis to follow by showing and discussing the results of some computational studies for the class of non-linear regression models. These results demonstrate a wide range of possible convergence rates, and associated stability (or

instability) of the operator to perturbations. With this intuition in hand, we then turn to Section 2.2, in which we set up the definitions that underlie our analysis. In particular, we state the (i) local Lipschitz condition, and (ii) local convergence behavior for the population-level operator F , and (iii) the stability and instability condition of the sample-level operator F_n with respect to F .

2.1 A vignette on non-linear regression

We first consider a certain class of statistical estimation problems in which there are interesting differences between algorithms. Here we keep the discussion very brief; see Section 4.3 for a more detailed discussion. We consider a simple type of non-linear regression model, one based on a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ that can be written in the form $f(x) = g(\langle x, \theta \rangle)$ for some parameter vector $\theta \in \mathbb{R}^d$, and some univariate function $g : \mathbb{R} \rightarrow \mathbb{R}$. In the simplest setting considered here, the univariate function g is known, and we have a parametric family of functions as θ ranges over \mathbb{R}^d ; when g is unknown, we have a semi-parametric family. Now suppose that we are given a collection of pairs $\{(X_i, Y_i)\}_{i=1}^n$, generated from a noisy regression model of the form

$$Y_i = g(\langle X_i, \theta^* \rangle) + \xi_i, \quad \text{for } i = 1, \dots, n. \quad (2)$$

Here ξ_i is a zero-mean noise variable with variance σ^2 , which we assume to be independent of X_i . The single index regression model (2) has been studied extensively in the literature (e.g., (Carroll et al., 1997; Ichimura, 1993)).

When g is known, a natural procedure for estimating θ is based on minimizing the least squares objective function

$$\mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^n \{Y_i - g(\langle X_i, \theta \rangle)\}^2. \quad (3)$$

When the variables ξ_i are Gaussian, then this objective coincides (up to scaling and constant factors) with the negative log-likelihood function, so that minimizing it yields the maximum likelihood estimate.

Under suitable regularity conditions on g in a neighborhood of θ^* , it is known that it is possible to estimate θ^* at the usual parametric rate of $n^{-1/2}$. However, problems can arise when the signal-to-noise ratio (SNR), as measured by the ratio $\|\theta^*\|_2/\sigma$, tends to zero. In particular, consider a function g whose derivative vanishes at zero—that is, $g'(0) = 0$. For instance, the function $g(t) = t^2$, which arises in the application of the non-linear regression framework to the problem of phase retrieval, has this property. Taking the limit of low SNR amounts to trying to estimate the vector $\theta^* = 0$ based on observations from the model (2). For this type of singular statistical model, we see many interesting differences between algorithms that might be used to minimize the least-squares criterion (3).

More concretely, let us consider three standard optimization algorithms that might be applied to the objective (3): (i) gradient descent; (ii) Newton’s method, and; (iii) cubic-regularized Newton’s method. See Appendix D.3 for a precise description of these algorithms and the associated updates in application to this model.

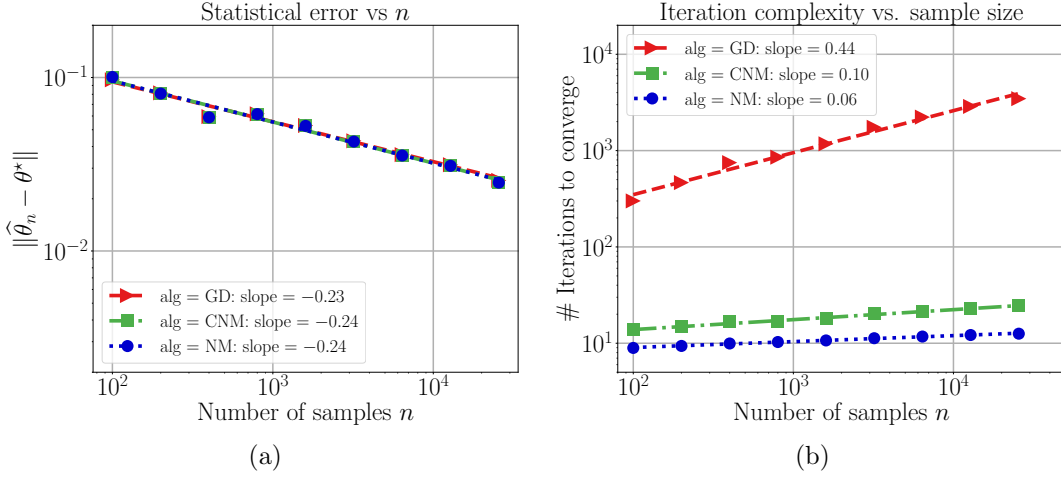


Figure 1. Plots characterizing the behavior of different algorithms, namely gradient descent (GD), cubic-regularized Newton’s method (CNM), and the vanilla Newton’s method (NM) for the non-linear regression model when $\theta^* = 0$. (a) Log-log plots of the Euclidean distance $\|\hat{\theta}_n - \theta^*\|_2$ versus the sample size. It shows that all the algorithms converge to an estimate at Euclidean distance of the order $n^{-1/4}$ from the true parameter θ^* . (b) Log-log plots for the number of iterations taken by different algorithms to converge to the final estimate. Newton’s method takes the least number of steps. On the other hand, gradient descent takes significantly larger number of steps, with an empirical scaling close to \sqrt{n} .

Statistical and iteration complexity of optimization algorithms: For each procedure, we are interested both in the associated statistical error—that is, the Euclidean distance between their output and the true parameter θ^* —and their iteration complexity, meaning the number of iterations required to converge. In order to gain some understanding, we performed some simulations for non-linear regression based on the function $g(t) = t^2$ in dimension $d = 1$, over a range of sample sizes n . Figure 1 provides some plots that summarize some results from these simulations. Panel (a) plots the Euclidean error associated with the estimate versus the sample size n on a log-log plot, along with associated least-squares fits to these data. As can be seen, all three methods lie upon a line with slope $-1/4$ on the log-log scale, showing that the statistical error decays at the rate $n^{-1/4}$. This “slow rate”—to be contrasted with the usual $n^{-1/2}$ parametric rate—is a consequence of the singularity in the model. Panel (b) plots the iteration complexity of the three algorithms versus the sample sizes, again on a log-log plot. For a given problem based on n samples, the iteration complexity is the number of iterations required for the distance between the iterate and θ^* to drop below $n^{-1/4}$. Here we see some interesting differences, with the gradient method having an empirical iteration complexity that grows as $\approx n^{0.44}$, based on our fits, with the two forms of Newton’s method having much milder growth in iteration complexity. In the theory to follow, we will prove that iteration complexity for the gradient method scales at most like \sqrt{n} , that of the cubic-regularized Newton method scales as $n^{1/6}$, whereas that of Newton’s method scales only as $\log n$. (See Corollary 9 for a precise statement.)

Behavior of optimization operators: The plots in Figure 1 all concern the behavior of algorithms in practice, as applied to the empirical objective function, and our ultimate goal is to provide a theoretical explanation of phenomena of these types. In order to do so, our

analysis makes use of the population-level algorithms obtained in the limit of infinite sample size; i.e., $n \rightarrow \infty$. In the special case of the non-linear regression model considered here, we refer the Appendix D.3 for the precise forms of these operators (cf. equations (107a)–(107c)). The plots in Figure 2 illustrate the two properties of the operators that underlie our theoretical analysis: convergence rate of the population operators (panel (a)), and the stability of the empirical operators relative to the population version (panel (b)).

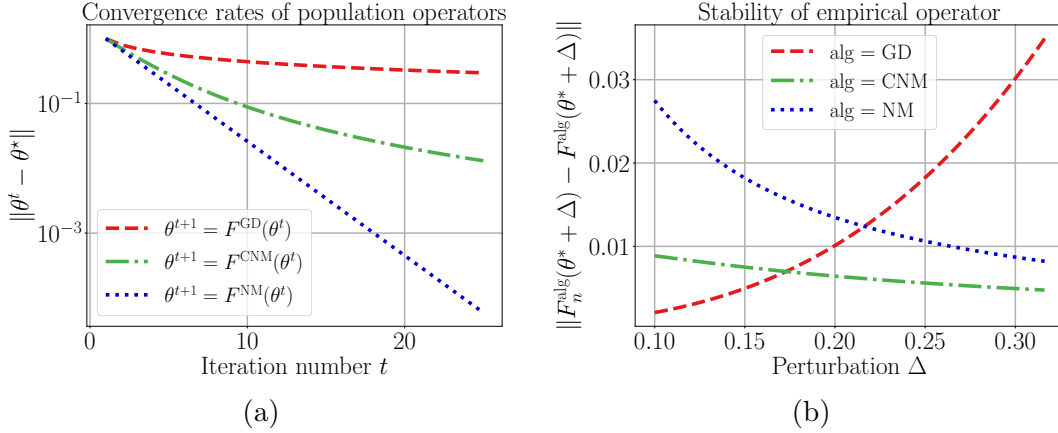


Figure 2. Exploration of the population level updates, and their connection to the empirical updates for the non-linear regression problem. (a) Plots showing the convergence rate of the error $\|\theta^t - \theta^*\|$ for different algorithms—namely gradient descent (GD), standard Newton’s method (NM), and cubic-regularized Newton’s method (CNM)—applied at the population level (limit of infinite sample size). Notice the log-scale on the y -axis. The sequence from the Newton’s method converges a geometric rate to θ^* , whereas the gradient method converges at a sub-linear rate. (b) Plots showing the scaling of the perturbation error $\|F_n(\theta^* + \Delta) - F(\theta^* + \Delta)\|$ versus the perturbation Δ . For an unstable operator, the perturbation error can increase as $\|\Delta\| \rightarrow 0$, with Newton’s method showing a strong version of such instability. In contrast, the gradient descent method is a stable procedure in this setting.

The plots in panel (a) reveal that the three algorithms differ dramatically in their convergence rate at the population level. The ordinary Newton updates converge at a geometric rate, with the distance to the optimum θ^* decreasing as κ^t with the number of iterations t , where $\kappa \in (0, 1)$ is a contraction coefficient. In contrast, the other two algorithms exhibit an inverse polynomial rate of convergence, with the distance to optimality decreasing at the rate $1/t^\beta$ for some exponent $\beta > 0$. In the analysis to follow, we prove that gradient descent has inverse polynomial decay with exponent $\beta = 1/2$, whereas the cubic-regularized Newton updates exhibit inverse polynomial decay with exponent $\beta = 2$.

In Corollary 9 and its proof, we characterize the optimization rate (algorithmic rate of convergence), the stability and the final statistical error obtained by these three methods. For reader’s convenience, we summarize these results in Table 2.

2.2 Problem set-up

Having provided a high-level overview of the phenomena that motivate our analysis, let us now set up the problem more abstractly, and introduce some key definitions. Consider an operator F that maps a space Θ to itself; typical examples of the space Θ that we consider

Algorithm	Optimization Rate	Stability	Iterations for convergence	Statistical error on convergence
Gradient descent	$\frac{1}{\sqrt{t}}$	$\frac{r}{\sqrt{n}}$	$n^{1/2}$	$n^{-1/4}$
Newton's method	$e^{-\kappa t}$	$\frac{1}{r\sqrt{n}}$	$\log n$	$n^{-1/4}$
Cubic-regularized Newton's method	$\frac{1}{t^2}$	$\frac{1}{\sqrt{r}\sqrt{n}}$	$n^{1/6}$	$n^{-1/4}$

Table 2. Overview of results illustrated in Figures 1 and 2 for non-linear regression model with the link function $g(t) = t^2$ and $\theta^* = 0$. By characterizing the optimization rate and stability precisely, and invoking our general theory (summarized in Table 1), we establish that while the three methods differ significantly in terms of their optimization rate and stability, they achieve the same statistical error upon convergence, albeit by taking different number of iterations to converge. We omit logarithmic factors and universal constants for brevity. See Corollary 9 and its proof for precise details.

are subsets of the Euclidean space \mathbb{R}^d , and subsets of symmetric matrices. Let θ^* be a fixed point of the operator—i.e., an element $\theta^* \in \Theta$ such that $F(\theta^*) = \theta^*$. The challenge is that we do not have access to the operator F directly, but rather can observe only a random operator F_n that can be understood as a noisy estimate of F . Throughout, we call F the *population operator* and F_n the *empirical operator*. Using the empirical operator, we generate a sequence of iterates via the fixed-point updates

$$\theta_n^{t+1} = F_n(\theta_n^t) \quad \text{for } t = 1, 2, \dots, \quad (4)$$

with a suitable initialization $\theta_n^0 \in \Theta$. Our goal is to determine conditions under which the sequence $\{\theta_n^t\}_{t \geq 0}$ approaches a suitably defined neighborhood of θ^* . More precisely, for any given triple (F, F_n, t) we provide a sharp characterization of the optimality gap $\|\theta_n^t - \theta^*\|_2$ as a function of the iteration count t and the error $\|F - F_n\|_2$ of the empirical operator F_n .

One interesting class of problems where the operators F and F_n arise naturally is estimation problems in statistics and machine learning. More concretely, consider the problem of finding the unique minimizer θ^* of an objective function $\mathcal{L} : \Theta \rightarrow \mathbb{R}$. In practice, we do not know the true objective function \mathcal{L} , instead we have access to an approximate (random) objective function \mathcal{L}_n , which is an unbiased estimate of the true objective function \mathcal{L} . Given the pair $(\mathcal{L}, \mathcal{L}_n)$, we can obtain different operators F by applying various optimization algorithms to minimize \mathcal{L} , including gradient methods, proximal methods, the EM algorithm and related majorization-minimization algorithms, as well as Newton and other higher-order methods. The noisy operators F_n are obtained by applying the same optimization algorithms to the approximate objective function \mathcal{L}_n .

. For ease of exposition, going forward the index n is synonymous with the sample size that defines the operator F_n ; while our general results, namely, Theorems 5 and 6 do not rely on use of this simplification.

2.2.1 PROPERTIES OF THE OPERATOR F

We begin by formalizing some properties of the operator F . We assume that the operator F has a unique fixed point θ^* in the local neighborhood of the Euclidean ball

$$\mathbb{B}(\theta^*, \rho) := \left\{ \theta \in \Theta \mid \|\theta - \theta^*\|_2 \leq \rho \right\} \quad (5)$$

centered at θ^* and we study its behavior in that local neighborhood. Our first condition is a standard Lipschitz condition on the operator F . In particular, we say that the operator F is 1-Lipschitz in $\|\cdot\|$ norm over the ball $\mathbb{B}(\theta^*, \rho)$ if

$$\|F(\theta_1) - F(\theta_2)\| \leq \|\theta_1 - \theta_2\| \quad \text{for all } \theta_1, \theta_2 \in \mathbb{B}(\theta^*, \rho). \quad (6)$$

In words, the 1-Lipschitz condition guarantees that the operator F is non-expansive with respect to perturbations of its argument.

Our next two definitions distinguish between fast and slow rates of convergence. The first definition captures an especially favorable property of operator F ; namely, it is locally contractive around the fixed point θ^* . The second definition considers a substantially slower (sub-linear) rate of convergence of the operator F .

Definition 1 (Fast convergence) *For a contraction coefficient $\kappa \in (0, 1)$, the operator F is **FAST**(κ)-convergent on the ball $\mathbb{B}(\theta^*, \rho)$ if*

$$\|F^t(\theta_0) - \theta^*\| \leq \kappa^t \|\theta_0 - \theta^*\| \quad \text{for all iterations } t = 1, 2, \dots, \quad (7)$$

and for all $\theta_0 \in \mathbb{B}(\theta^*, \rho)$.

Definition 2 (Slow convergence) *Given an exponent $\beta > 0$, the operator F is **SLOW**(β)-convergent over the ball $\mathbb{B}(\theta^*, \rho)$ means that*

$$\|F^t(\theta_0) - \theta^*\| \leq \frac{c}{t^\beta} \quad \text{for all iterations } t = 1, 2, \dots, \quad (8)$$

and for all $\theta_0 \in \mathbb{B}(\theta^*, \rho)$, where c is a universal constant.

These notions of fast and slow convergence are ubiquitous in analysis of iterative methods, especially in the optimization literature. For example, when the operator F corresponds to gradient descent for some objective \mathcal{L} , a sufficient condition for fast convergence is local strong convexity of the objective \mathcal{L} , and if \mathcal{L} is just convex, F satisfies slow convergence. Let us now illustrate these definitions with a simple example.

Example 1 (Fast versus slow convergence) *Consider the function $\mathcal{L}(\theta) = \frac{\theta^{2p}}{2p}$ for some positive integer $p \geq 1$. Note that for any $p \geq 1$, the function $\mathcal{L}(\cdot)$ has a unique global minimum at $\theta^* = 0$. The first two derivatives of $\mathcal{L}(\cdot)$ are given by*

$$\mathcal{L}'(\theta) = \theta^{2p-1}, \quad \text{and} \quad \mathcal{L}''(\theta) = (2p-1)\theta^{2p-2}.$$

Consequently, a gradient descent update with a constant stepsize $h > 0$ takes the form

$$F^{GRD}(\theta) = \theta - h\mathcal{L}'(\theta) = \theta(1 - h\theta^{2p-2}). \quad (9)$$

Thus, when $p = 1$, for any $h \in (0, 1)$, this gradient descent update is a $\text{FAST}(\kappa)$ -convergent algorithm with $\kappa = 1 - h$. On the other hand, for any $p \geq 2$, it can be shown that gradient descent is $\text{SLOW}(\beta)$ -convergent with parameter $\beta = \frac{1}{2p-2}$ in the ball $\mathbb{B}(\theta^*, \rho)$ with $\theta^* = 0$ and $\rho = h^{-\frac{1}{2p-2}}$.

Now, let us consider Newton's method with step size one, namely the update

$$F^{NWT}(\theta) = \theta - (\mathcal{L}''(\theta))^{-1} \mathcal{L}'(\theta) = \theta - \frac{\theta^{2p-1}}{(2p-1)\theta^{2p-2}} = \theta \left(1 - \frac{1}{2p-1}\right). \quad (10)$$

For $p = 1$, this update converges in a single step (simply because the quadratic approximation that underlies Newton's method is exact in this special case). For $p \geq 2$, the pure Newton update is $\text{FAST}(\kappa)$ -convergent with $\kappa = 1 - \frac{1}{2p-1}$ for all $\theta \in \mathbb{R}$.

2.2.2 FROM THE EMPIRICAL OPERATOR F_n TO THE POPULATION OPERATOR F

In this section, we introduce some key concepts that characterize the (in)-stability of the sample operator F_n with respect to the population operator F . Given a pair of operators (F_n, F) and a tolerance parameter $\delta \in (0, 1)$, our definitions involve a *perturbation function* $\varepsilon(\cdot)$ that maps the triple (F_n, F, δ) to a positive (deterministic) scalar $\varepsilon(n, \delta)$. In general, we impose the following conditions on the perturbation function $\varepsilon(\cdot)$:

- It is decreasing in n for any fixed δ , and is monotonically increasing in δ for any fixed n .
- For any fixed $\delta \in (0, 1)$, we have $\varepsilon(n, \delta) \rightarrow 0$ as $n \rightarrow \infty$, and similarly, for any fixed $n > 0$, we have $\varepsilon(n, \delta) \rightarrow \infty$ as $\delta \rightarrow 0$.

Note that $\varepsilon(n, \delta) = \sqrt{\log(1/\delta)/n}$ would satisfy these requirements. Given some choices of perturbation function, we can define our first stability condition as follows:

Definition 3 (STA(γ)-Stability) For a given parameter $\gamma \geq 0$, the operator F_n is $\text{STA}(\gamma)$ -stable over $\mathbb{B}(\theta^*, \rho)$ with noise function $\varepsilon(\cdot)$ means that, for any radius $r \in (0, \rho)$ and tolerance $\delta \in (0, 1)$, we have

$$\mathbb{P} \left[\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|F_n(\theta) - F(\theta)\| \leq c_2 \min \left\{ r^\gamma \varepsilon(n, \delta), r \right\} \right] \geq 1 - \delta, \quad (11)$$

for some positive universal constant c_2 .

Informally, given the randomness of the data associated with the operator F_n the stability condition (11) guarantees that with high probability, the error $\|F_n(\theta) - F(\theta)\|$ is upper bounded by $c_2 \min\{r^\gamma \varepsilon(n, \delta), r\}$ uniformly over a disk of radius r . Note moreover that the upper bound decays to 0 as the radius $r \rightarrow 0^+$.

Next we consider the case when $\gamma < 0$, i.e., the perturbation error $\|F_n(\theta) - F(\theta)\|$ blows up as θ gets close to θ^* . We refer to such operators as unstable operators. Given radii r_1, r_2 such that $r_2 > r_1 \geq 0$, let $\mathbb{A}(\theta^*, r_1, r_2) = \mathbb{B}(\theta^*, r_2) \setminus \mathbb{B}(\theta^*, r_1)$ denote the annulus around θ^* with inner and outer radii r_1 and r_2 respectively.

Definition 4 (UNS(γ)-Instability) For a given parameter $\gamma < 0$ and radii $0 < \rho_{in} < \rho_{out}$, we say that the operator F_n is UNS(γ)-unstable over the annulus $\mathbb{A}(\theta^*, \rho_{in}, \rho_{out})$ with noise function $\varepsilon(\cdot)$ if

$$\mathbb{P} \left[\sup_{\theta \in \mathbb{A}(\theta^*, r, \rho_{out})} \|F_n(\theta) - F(\theta)\| \leq \varepsilon(n, \delta) \max \left\{ \frac{1}{r^{|\gamma|}}, \rho_{out} \right\} \right] \geq 1 - \delta, \quad (12)$$

for any radius $r \in [\rho_{in}, \rho_{out}]$ and any tolerance $\delta \in (0, 1)$.

Two remarks are in order: First, note that the main difference between STA(γ) and UNS(γ) is how the error scales with the radius r as it gets smaller. For stable operators, the error decreases with scaling r^γ , while for unstable operators the error blows up as $r^{-|\gamma|}$ (where we use $|\gamma|$ for clarity). There is another subtle difference: the condition (12) defines the instability of the perturbation error $\|F_n(\theta) - F(\theta)\|$ in an annulus with the inner radius bounded below by ρ_{in} , and does not characterize the behavior as the distance $\|\theta - \theta^*\| \rightarrow 0$. Let us now illustrate these definitions by following up on Example 1.

Example 2 (Stable versus unstable updates) Consider an empirical function of the form

$$\mathcal{L}_n(\theta) = \frac{1}{2p} \theta^{2p} + \frac{\sigma w}{2\sqrt{n}} \theta^2, \quad \text{where } w \sim N(0, 1). \quad (13)$$

Here $p \geq 2$ is a positive integer. Note that $\mathbb{E}[\mathcal{L}_n(\theta)] = \frac{1}{2p} \theta^{2p}$, which is equivalent to the population likelihood function considered in Example 1.

A gradient update with stepsize $h > 0$ on the empirical objective leads to the empirical gradient operator

$$F_n^{GRD}(\theta) = \theta \left\{ 1 - h\theta^{2p-2} - h \frac{\sigma w}{\sqrt{n}} \right\}.$$

Comparing with equation (9), we obtain that $|F_n^{GRD}(\theta) - F^{GRD}(\theta)| = \frac{\sigma}{\sqrt{n}} |w| |\theta|$. Since $|w| \leq 4\sqrt{\log(1/\delta)}$ with probability at least $1 - \delta$, we see for any $\rho > 0$ and $n \geq 16\sigma^2 \log(1/\delta)$, the operator F_n^{GRD} is STA(γ)-stable with parameter $\gamma = 1$, with respect to the noise function

$$\varepsilon(n, \delta) = 4\sigma \sqrt{\frac{\log(1/\delta)}{n}}.$$

As for the Newton update for the problem (13), we have

$$F_n^{NWT}(\theta) = \theta - \frac{\theta^{2p-1} + \sigma w \theta / \sqrt{n}}{(2p-1)\theta^{2p-2} + \sigma w / \sqrt{n}},$$

and hence

$$|F_n^{NWT}(\theta) - F^{NWT}(\theta)| = \frac{(2p-2)}{(2p-1)} \cdot \frac{\sigma |w| |\theta| / \sqrt{n}}{(2p-1)\theta^{2p-2} + \sigma w / \sqrt{n}}.$$

Recall that $|w| \leq 4\sqrt{\log(1/\delta)}$ with probability at least $1 - \delta$. Plugging in $w > -4\sqrt{\log(1/\delta)}$ in the denominator and $w < 4\sqrt{\log 1/\delta}$ of the RHS, and doing some algebra yields that

$$|F^{NWT}(\theta) - F_n^{NWT}(\theta)| \leq \frac{c_p}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}} \quad \text{for } |\theta| > \left(c'_p \sigma \sqrt{\frac{\log(1/\delta)}{n}}\right)^{\frac{1}{2p-2}},$$

where $c_p = \frac{16(p-1)}{2p-1}$ and $c'_p = \frac{8}{2p-1}$. Thus, we conclude that the operator F_n^{NWT} is $\text{UNS}(\gamma)$ -unstable with parameter $\gamma = -1$ over the annulus $\mathbb{A}(\theta^*, \rho_{in}, \rho_{out})$ with noise function ε where

$$\rho_{in} = \left(c'_p \sigma \sqrt{\frac{\log(1/\delta)}{n}}\right)^{\frac{1}{2p-2}}, \quad \rho_{out} = \infty, \quad \text{and} \quad \varepsilon(n, \delta) = c_p \sigma \sqrt{\frac{\log(1/\delta)}{n}}.$$

2.2.3 COMPARISON OF OUR ASSUMPTIONS WITH EMPIRICAL PROCESS LITERATURE

We note that while the definition of $\text{STA}(\gamma)$ is reminiscent of the typical assumptions in empirical process literature, there is a subtle difference in our set-up. In a typical statistical learning problem, the following assumptions are commonly made on: (a) the local curvature of the population objective function (e.g., the expected negative log-likelihood \mathcal{L}), and (b) bounds on the perturbation error between the population and sample objective functions (e.g., $\sup_{\theta \in \mathbb{B}(\theta^*, r)} |\mathcal{L}(\theta) - \mathcal{L}_n(\theta)|$). With these assumptions, the statistical guarantees for the critical points (e.g., the maximum likelihood estimate (MLE)) are then established. See, e.g., Theorem 3.2.5 (van der Vaart, 1998).

Such a framework is oblivious about any computational aspect of the problem (e.g., how the MLE is computed), which is one of the key focus in our work. Our goal is to study the interplay of computational-statistical tradeoffs between various algorithms that are used to solve these learning problems. In particular, our aim is to identify the number of iterations taken by an algorithm, and the final statistical accuracy of the estimate returned by it. Consequently, our conditions are defined in terms of operators that correspond to the algorithm employed by the user to solve the problem at hand, rather than the landscape of the objective itself. In particular, in place of the curvature condition (a) on \mathcal{L} , we make assumptions on the convergence rate of the population operator F ($\text{SLOW}(\beta)/\text{FAST}(\kappa)$). And, in place of the perturbation bounds on the objective functions $(\mathcal{L}, \mathcal{L}_n)$, we make assumptions on the operator perturbation errors between F and F_n as in equations (11) and (12) ($\text{STA}(\gamma)/\text{UNS}(\gamma)$).

2.2.4 FURTHER DISCUSSION ON OUR DEFINITIONS

In several cases (also applicable to all examples in this paper), the user often knows (by design) the explicit relationship between the operators F and F_n and the corresponding objectives \mathcal{L} and \mathcal{L}_n , e.g., when F and F_n correspond to gradient ascent (GA) or Newton's method (NM). In these situations, often it is possible to derive whether $\text{STA}(\gamma)$ or $\text{UNS}(\gamma)$ conditions are satisfied given the assumptions on the curvature of \mathcal{L} and the perturbation error between \mathcal{L} and \mathcal{L}_n as in the empirical process literature. Our framework allows the user to simultaneously study the tradeoffs between the final statistical error and the computational budget needed between several algorithms at once. For example, we show in several settings that NM while being unstable provides computational benefits compared

to its stable counterpart GA, since both NM and GA yield an estimate with comparable statistical error upon convergence while the former takes very few steps (although such a condition is not guaranteed always).

We remark that the property $\text{UNS}(\gamma)$ of the operators F and F_n as introduced above has not been commonly used in prior work, while $\text{STA}(\gamma)$ has appeared often in prior works (albeit in slightly different forms, (Balakrishnan et al., 2017; Chen et al., 2018a; Dwivedi et al., 2020a)). The condition $\text{STA}(0)$ is perhaps the most common, which holds for most well-conditioned problems (when the Fisher information matrix is invertible). In such settings, the commonly used methods like GA and NM are also $\text{FAST}(\kappa)$ operators so that the final statistical error is of order $\varepsilon(n, \delta)$ which is obtained in roughly $\log(1/\varepsilon(n, \delta))$ steps (Balakrishnan et al., 2017). In simple words, the statistical-computational tradeoffs across several algorithms are fairly similar for such cases.

On the other hand, operators with $\text{SLOW}(\beta)$, and $\text{STA}(\gamma)$ with $\gamma \geq 1$, would typically arise when the log-likelihood is not well-conditioned and one uses methods like GA, and $\text{UNS}(\gamma)$ with $\gamma < 0$ would appear in such a setting when one uses a higher-order optimization scheme like NM to solve these ill-conditioned problems. So far, there is a limited understanding of the statistical-computational tradeoff for slowly converging stable algorithms as well as any unstable algorithm that can arise in such settings. When the population operator has a unique fixed point, our main results provide a fairly comprehensive understanding towards this end.

Let us revisit Examples 1 and 2 which serve as motivating examples for the various ill conditioned settings: Suppose that the population negative log-likelihood is given by $\mathcal{L}(\theta) = \theta^{2p}/(2p)$ (and the true parameter is $\theta^* = 0$), and the sample negative log-likelihood is given by $\mathcal{L}_n(\theta) = \theta^{2p}/(2p) + \sigma w \theta^2/(2\sqrt{n})$. For this setting, the population Fisher information (the second derivative of \mathcal{L}) is given by $(2p - 1)\theta^{2p-2}$, and the perturbation term $\sigma w \theta^2/(2\sqrt{n})$ between the two objectives mimics the intuition that the finite sample Fisher information would typically have $1/\sqrt{n}$ fluctuations around its population-level objective with n samples. With $p = 1$, it is easy to establish that the operators F and F_n corresponding to both GD and NM are $\text{FAST}(\kappa)$ and $\text{STA}(0)$ operators. However, for $p \geq 2$, as our earlier computations illustrated, we observe a more interesting set of behaviors with GD and NM. In particular, the operators corresponding to GA are $\text{SLOW}(\beta)$ and $\text{STA}(\gamma)$ with $\gamma > 1$, and the NM operators exhibit a $\text{FAST}(\kappa)$ and $\text{UNS}(\gamma)$ with $\gamma < 0$ behavior. The theory to follow provides a precise characterization of the statistical error achievable and the computational budget needed for convergence in such settings.

3. General convergence results

With the definitions from the previous section in place, we are now ready to state our main results. In Section 3.1, we consider the case when F_n is a stable perturbation of F , and in Section 3.2, we consider the case when it is an unstable perturbation of F . We summarize our findings in Table 1.

3.1 Results for slowly converging but stable operators

We first consider the setting in which the sample-based operator F_n is a stable perturbation of the population-level operator F . If, in addition, we assume that the operator F has

fast convergence (cf. equation (7)), then past work is applicable. In particular, Theorem 2 of Balakrishnan et al. (2017) provides a precise characterization of the convergence behavior of iterates from the empirical operator F_n . Here we instead consider the more challenging setting in which the operator F exhibits slow convergence to θ^* . Analysis of this slow convergence case requires rather different techniques than those used to analyze the fast-convergent case.

Let us collect the assumptions needed to state our first result. The first two assumptions involve the Euclidean ball $\mathbb{B}(\theta^*, \rho)$ centered at θ^* of some fixed radius $\rho > 0$.

- (A) The population operator F is 1-Lipschitz (6) and is $\text{SLOW}(\beta)$ -convergent (8) over the ball $\mathbb{B}(\theta^*, \rho)$.
- (B) There is some $\gamma \in [0, \beta^{-1}]$ such that the empirical operator F_n is $\text{STA}(\gamma)$ -stable (11) over $\mathbb{B}(\theta^*, \rho)$.
- (C) The tolerance parameters $\delta \in (0, 1)$ and $\alpha \in (0, \frac{\beta}{1+\beta-\gamma\beta})$ are fixed and the sample size is large enough such that

$$\varepsilon(n, \delta^*) \leq c \quad \text{where} \quad \delta^* := \delta \cdot \frac{\log(\frac{1+\beta}{\beta\gamma})}{8 \log(\frac{\beta}{\alpha(1+\beta-\gamma\beta)})}, \quad (14)$$

and $c \in (0, 1)$ is a sufficiently small constant.

Assumptions (A) and (B) quantify, respectively, the convergence behavior of the operator F and the stability of the operator F_n ; Assumption (C) is a book-keeping device needed to state our results cleanly. Given the above conditions, we now state our first main result.

Theorem 5 *Under Assumptions (A), (B), and (C), consider the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ generated from an initialization $\theta_n^0 \in \mathbb{B}(\theta^*, \rho/2)$. Then there is a universal constant c' such that for any fixed $\alpha \in (0, \frac{\beta}{1+\beta-\gamma\beta})$ and uniformly for all iterations $t \geq c'(1/\varepsilon(n, \delta^*))^{\frac{1}{1+\beta-\gamma\beta}} \log \frac{1}{\alpha}$, we have*

$$\|\theta_n^t - \theta^*\| \leq 2[\varepsilon(n, \delta^*)]^{\frac{\beta}{1+\beta-\gamma\beta}-\alpha} \quad \text{with probability at least } 1 - \delta. \quad (15)$$

Let us make some comments on this result (see Appendix A.1 for a detailed proof).

Tightness of Theorem 5: Disregarding the term α and constants, the bound (15) guarantees that the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ converges to a statistical tolerance of order $[\varepsilon(n, \delta^*)]^{\frac{\beta}{1+\beta-\gamma\beta}}$ with respect to θ^* in order $[\varepsilon(n, \delta^*)]^{-\frac{1}{1+\beta-\gamma\beta}}$ step. This guarantee turns out to be unimprovable under the given assumptions. In Appendix B (see Proposition 12), we construct a family of examples with the operators F , F_n and noise functions ε_n (constant with respect to δ) satisfying the assumptions for Theorem 5, such that the following additional results hold:

$$\|\theta_n^t - \theta^*\| \begin{cases} \geq \varepsilon_n^{\frac{\beta}{1+\beta-\gamma\beta}} & \text{for all } t \geq 1, \\ \geq 2\varepsilon_n^{\frac{\beta}{1+\beta-\gamma\beta}} & \text{for all } t \leq c'\varepsilon_n^{-\frac{1}{1+\beta-\gamma\beta}}. \end{cases}$$

As a result, we conclude that the results of Theorem 5 are tight for both statistical accuracy and the number of iterations needed for convergence.

Relation to prior work: As noted earlier, prior work (Balakrishnan et al., 2017) shows that when the operator is F is $\text{FAST}(\kappa)$ -convergent, and F_n is a $\text{STA}(0)$ perturbation of F_n with error $\varepsilon(n, \delta)$, we have $\|\theta_n^T - \theta^*\| \lesssim \varepsilon(n, \delta)$ for $T \gtrsim \log(1/\varepsilon(n, \delta))$. On the other hand, Chen et al. (2018a) argued that given the minimax error of a problem class, a fast converging algorithm can not be too stable. Neither of these works provided a precise characterization of what statistical errors are achievable when dealing with a slow converging stable operator, which is the focus of our Theorem 5.

A direct sub-optimal proof argument: Let us first illustrate how a naive argument that tries to directly tradeoff the perturbation error of F_n with the convergence rate of F leads to a sub-optimal guarantee. Let the assumptions in Theorem 5 remain in force. Roughly speaking, one can show that (cf. Lemma 10), the operator F_n^t is also $\text{STA}(\gamma)$ perturbation of F^t with the noise function $t \cdot \varepsilon(n, \delta)$, so that we can bound the error at iteration t as follows:

$$\|\theta_n^t - \theta^*\| = \|F_n^t(\theta_n^0) - \theta^*\| \leq \|F_n^t(\theta_n^0) - F^t(\theta_n^0)\| + \|F^t(\theta_n^0) - \theta^*\| \leq tC_\rho \cdot \varepsilon(n, \delta) + \frac{1}{t^\beta}, \quad (16)$$

where $C_\rho = \rho^\gamma$ denotes a constant corresponding to the radius ρ of initialization. Minimizing the last bound in the display above over the iteration index t , we find that the best possible error is of order $(\varepsilon(n, \delta))^{\beta/(1+\beta)}$. This rate is clearly sub-optimal when compared to the statistical error of order $(\varepsilon(n, \delta))^{\beta/(1+\beta-\gamma\beta)}$ (unless $\gamma = 0$) guaranteed by display (15) from Theorem 5. The reason for sub-optimality of this bound is our failure to localize the argument with the perturbation error as the iterates θ_n^t converge closer to θ^* .

Outline of proof: In order to derive the sharp guarantee, we need to establish a more refined tradeoff than that in equation (16). To this end, we generalize and refine the annulus-based localization argument introduced in our prior work on the EM algorithm (Dwivedi et al., 2020a,b). In the past work (Dwivedi et al., 2020a,b), we studied particular instantiations of the EM algorithm, for which the operators F and F_n had closed-form solutions. Here in the absence of closed-form expressions, the argument is necessarily more abstract to establish a sharp guarantee under the more general Assumptions (A), (B), and (C), which also handles the previous analysis as a special case (as illustrated in Section 4.2).

At a high-level, the proof proceeds by decomposing the total collection of iterations $\{1, 2, \dots, t\}$ into a disjoint partition of subsets $\{T_\ell\}_{\ell \geq 0}$, referred to as epochs, where the nonnegative integers ℓ and T_ℓ respectively denote the index of a given epoch and the number of iterations in that epoch. We use $S_\ell := \sum_{i=0}^\ell T_i$ to denote the total number of iterations up to epoch ℓ . By carefully choosing the sequence $\{T_\ell\}_{\ell \geq 0}$, we ensure that at the end of a given epoch ℓ , the error $\|\theta_n^{S_\ell} - \theta^*\|$ has decreased to a prescribed threshold. More precisely,

. In several statistical settings, the error function satisfies $\varepsilon(n, \delta) = c\sqrt{\log(1/\delta)/n}$. When the problems are well-conditioned, e.g., if the Fisher information matrix is invertible while estimating MLE, we typically have $\text{FAST}(\kappa)$ and $\text{STA}(0)$ condition for commonly used algorithms like gradient ascent, and Newton's method. In general, we do not expect a setting where the operators are $\text{FAST}(\kappa)$ and $\text{STA}(\gamma)$ with $\gamma \geq 1$. Although, one can construct pathological examples, in which case, our localization argument augmented with the earlier proofs by Balakrishnan et al. (2017) would yield that $\theta_n^t \xrightarrow{t \rightarrow \infty} \theta^*$, i.e., the statistical error converges to zero as the number of iterations goes to ∞ (even with finite samples).

using an inductive argument on ℓ , we show that

$$\|\theta_n^{S_\ell} - \theta^*\| \leq \varepsilon(n, \delta^*)^{\lambda_\ell} \quad \text{for all epoch } \ell \geq 1, \quad (17)$$

where the sequence $\{\lambda_\ell\}_{\ell \geq 0}$ is defined via the recursion

$$\lambda_0 = 0 \quad \text{and} \quad \lambda_{\ell+1} = \nu \lambda_\ell + \nu', \quad \text{for all } \ell \geq 1, \quad (18)$$

with the scalars $\nu \in (0, 1)$ and $\nu' > 0$ determined by the problem parameters β and γ . We show that the sequence $\{\lambda_\ell\}_{\ell \geq 0}$ converges to $\nu_* := \frac{\beta}{1+\beta-\gamma\beta}$ fast enough and we have $|\lambda_\ell - \nu_*| \leq \alpha$ for all $\ell \geq \mathcal{O}(\log(1/\alpha))$. Deriving a suitable upper bound on T_{\max} on the epoch size T_i , we then put the pieces together to (roughly) conclude that

$$\|\theta_n^t - \theta^*\| \leq c\varepsilon(n, \delta^*)^{\nu_* - \alpha} \quad \text{for } t \geq c'T_{\max} \cdot \log \frac{1}{\alpha}.$$

As expected, much of the technical work is required to establish the inductive step. The full proof of the theorem is given in Appendix A.1. We also illustrate the high-level ideas of the epoch-based localization argument in Figure 3.

3.2 Results for unstable operators

We now turn to our next main result which characterizes the convergence when the operator F_n is an unstable perturbation of the operator F . We consider two distinct cases depending on whether the operator F is (a) **FAST**(κ)-convergent or (b) **SLOW**(β)-convergent. In order to obtain sharp upper bounds—ones that depend purely on the noise function ε —the inner radius for the instability of the operator F_n around the operator F , which we refer to as $\tilde{\rho}_n$, must be sample size-dependent and chosen suitably.

Theorem 6 *For a given parameter $\delta \in (0, 1)$, consider the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ for some initial point θ_n^0 in the ball $\mathbb{B}(\theta^*, \rho/2)$. Suppose that for some $\gamma < 0$, the empirical operator F_n is **UNS**(γ)-unstable over the annulus $\mathbb{A}(\theta^*, \tilde{\rho}_n, \rho)$ with respect to the noise function ε .*

- (a) *Suppose that the operator F is **FAST**(κ)-convergent over the ball $\mathbb{B}(\theta^*, \rho)$, and the sample size n is sufficiently large so as to ensure that*

$$[\varepsilon(n, \delta)]^{\frac{1}{1+|\gamma|}} \leq (1 - \kappa)\rho. \quad (19a)$$

Then with probability at least $1 - \delta$, for any iteration $t \geq \frac{\log(\frac{\rho}{\varepsilon(n, \delta)})}{(1+|\gamma|) \log \frac{1}{\kappa}}$, we have

$$\min_{k \in \{0, 1, \dots, t\}} \|\theta_n^k - \theta^*\| \leq \max \left\{ \frac{(2 - \kappa)}{(1 - \kappa)} \cdot [\varepsilon(n, \delta)]^{\frac{1}{1+|\gamma|}}, \tilde{\rho}_n \right\}. \quad (19b)$$

- (b) *Suppose that the operator F is 1-Lipschitz and **SLOW**(β)-convergent for some $\beta > 0$, and that the sample size n is large enough to ensure that*

$$[\varepsilon(n, \delta)]^{\frac{\beta}{1+\beta-\gamma\beta}} \leq \rho. \quad (20a)$$

Then with probability at least $1 - \delta$, for any iteration $t \geq \frac{1}{[\varepsilon(n, \delta)]^{\frac{1}{1+\beta}}}$, we have

$$\min_{k \in \{0, 1, \dots, t\}} \|\theta_n^k - \theta^*\| \leq \max \left\{ [\varepsilon(n, \delta)]^{\frac{\beta}{1+\beta-\gamma\beta}}, \tilde{\rho}_n \right\}. \quad (20b)$$

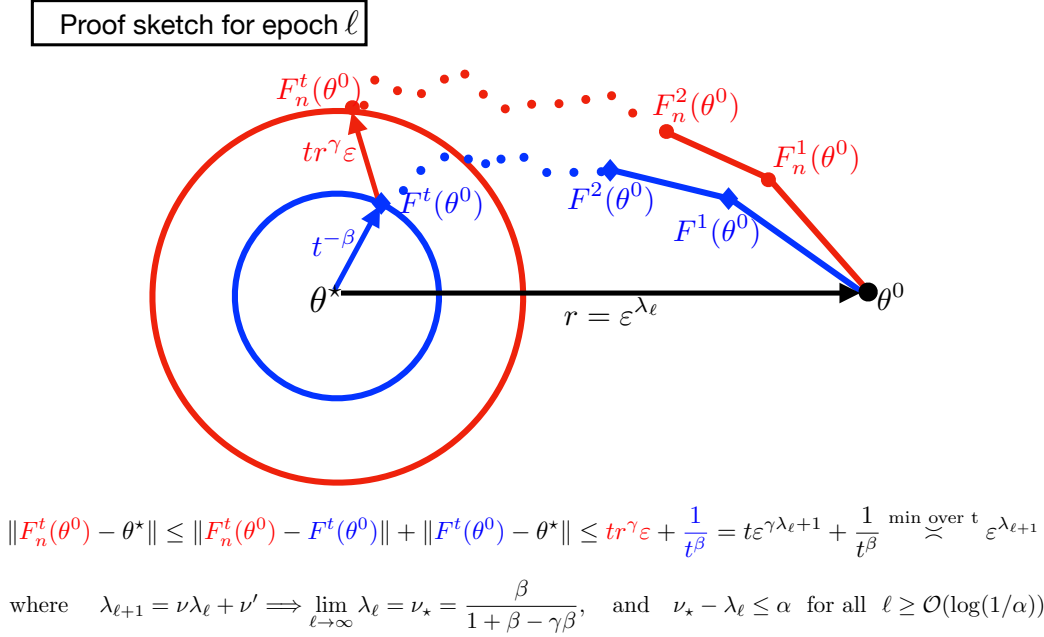


Figure 3. An illustration of the epoch-based argument when the population operator F is $\text{SLOW}(\beta)$ -convergent, and the noisy operator is $\text{STA}(\gamma)$ -stable (Theorem 5). In order to simplify the visualization, we use the shorthand $\varepsilon = \varepsilon(n, \delta^*)$. Moreover, here θ^0 denotes the starting point for a given epoch ℓ (assumed to be at distance $r = \varepsilon^{\lambda_\ell}$ from θ^*), and the iterations $1, 2, \dots, t$ denote the iteration count in that epoch. The population iterates $F^1(\theta^0), F^2(\theta^0), \dots$ converge towards to θ^* at the rate $t^{-\beta}$ (shown in blue), and their distance from the noisy iterates $F_n^1(\theta^0), F_n^2(\theta^0), \dots$ grows at the rate at a distance of $tr^\gamma \varepsilon$. Trading-off the two errors, we can show that at the end of epoch ℓ (denoted by a suitable choice of t), the distance $\|F_n^t(\theta^0) - \theta^*\| \lesssim \varepsilon^{\lambda_{\ell+1}}$. By establishing that λ_ℓ converges to ν_* exponentially fast, and that similar arguments can be made for sufficiently many epochs, we obtain the result in Theorem 5. See Appendix A.1 for a formal argument.

Let us make a few comments about these bounds. (See Appendix A.2 for a detailed proof.)

Choice of the inner radius $\tilde{\rho}_n$: Focusing on part (a), if we ensure that $\tilde{\rho}_n \leq [\varepsilon(n, \delta)]^{\frac{1}{1+\gamma}}$, then we obtain an upper bound on the error that involves only the noise function. We show how to make such choices in our applications of this general theorem. A similar statement applies to part (b) of the theorem.

Tightness of Theorem 6: In Appendix B, we construct examples of the operators F and F_n which satisfy the assumptions of Theorem 6, and with the inner radius satisfying the bound $\tilde{\rho}_n \leq [\varepsilon(n, \delta)]^\tau$, $\tau = \frac{1}{1+\gamma}$ for part (a) or $\tau = \frac{\beta}{1+\beta-\gamma\beta}$ for part (b). For each of these examples, we show that the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ satisfies the lower bound

$$\|\theta_n^t - \theta^*\| \geq [\varepsilon(n, \delta)]^\tau \quad \text{for all } t \geq 0,$$

with constant probability. Thus, we conclude that the results of Theorem 6 are tight and not improvable in general.

Necessity of the minimum: Note that both of the bounds (19b) and (20b) apply to the minimum over all iterates $k \in \{1, 2, \dots, t\}$, as opposed to the final iterate t . For this reason, our results only guarantee that the iterates produced by an unstable operator F_n converge at least once to a vicinity of the parameter θ^* , but *not* that they necessarily stay there for all the future iterations. In fact, such “escape” behavior for an unstable algorithm is unavoidable in the absence of any additional regularity assumptions. In particular, we provide a simple example in Appendix B.4 that illustrates this unavoidability.

Additional regularity condition: If we impose an additional regularity condition, then we can remove the minimum from the guarantee. In particular, consider the condition:

- (D) There exists a universal constant C such that for a given initialization θ_n^0 , the sequence $\theta_n^t = F_n^t(\theta_n^0)$ has the following property:

$$\|\theta_n^{t+1} - \theta^*\| \leq C\tilde{\rho} \quad \text{whenever} \quad \|\theta_n^t - \theta^*\| \leq \tilde{\rho}, \quad (21)$$

where the radius $\tilde{\rho}$ corresponds to equation (19b) or (20b) as relevant to F .

Under this condition, it is straightforward to modify the proof of Theorem 6 to show that the bounds in both parts (a) and (b) can be sharpened by replacing the term $\min_{k \in \{0, 1, \dots, t\}} \|\theta_n^k - \theta^*\|$ with $\|\theta_n^t - \theta^*\|$. In Section 4 to follow, we provide a number of examples for which Assumption (D) is satisfied.

4. Some concrete results for specific models

In this section, we study three interesting classes of statistical problems that fall within the framework of the paper. We also discuss various consequences of Theorems 5 and Theorem 6 when applied to these problems.

4.1 Informative non-response model

In our first example, let us consider the problem of biased or informative non-response in sample surveys. In certain settings, the chance of a response to not be observed depends on the value of the response. This form of non-response introduces systematic biases in the survey and associated conclusions (Heckman, 1976). Some examples where this issue arises include longitudinal data (Diggle and Kenward, 1994), housing surveys and election polls (Shaiko et al., 1991). In such settings, it is common practice to estimate the non-responsive behavior in order to correct for the bias. We now describe one simple formulation of such a setting.

Suppose that we have n i.i.d. values Y_1, \dots, Y_n for the response variable $Y \sim \mathcal{N}(\mu, \sigma^2)$, where for each Y_i there is a chance that the value is not observed. To account for such a possibility, we define $\{0, 1\}$ -valued random variables R_i for $i = 1, \dots, n$ as follows:

$$R_i = 1 \quad \text{if } Y_i \text{ is observed,} \quad \text{and} \quad R_i = 0 \quad \text{otherwise.} \quad (22a)$$

We assume that the conditional distribution $R_i|Y_i$ takes the form

$$\mathbb{P}_\theta(R_i = 1|Y_i = y) = \exp\left(H\left(\frac{\theta(y - \mu)}{\sigma}\right)\right), \quad (22b)$$

where H is a known function and θ is an unknown parameter which controls the dependence of the probability of non-response on the observation $Y = y$. In a general setting, all the parameters μ, σ and θ are unknown and are estimated jointly from the data. However, to simplify our presentation, we assume that the parameters (μ, σ) are known and only θ needs to be estimated. In particular, we consider the case when the response variable $Y \sim \mathcal{N}(\mu, \sigma^2) \equiv \mathcal{N}(0, 1)$ and $H(x) = -x^2 - \log 2$. Under these assumptions, simple algebra yields that

$$\mathbb{P}_\theta(R_i = 1 | Y_i = y) = \exp\left(-\frac{\theta^2 y^2}{2} - \log 2\right) \quad \text{and} \quad \mathbb{P}_\theta(R_i = 1) = \frac{1}{2\sqrt{\theta^2 + 1}}. \quad (22c)$$

Given n i.i.d. samples $\{R_i, Y_i\}_{i=1}^n$, where we note that Y_i is not observed when $R_i = 0$, the log-likelihood is given by

$$\bar{\mathcal{L}}_n(\theta) := \frac{1}{n} \sum_{i=1}^n -\frac{R_i (Y_i^2(\theta^2 + 1) + 2 \log 2)}{2} + (1 - R_i) \log\left(1 - \frac{1}{2\sqrt{\theta^2 + 1}}\right). \quad (23)$$

Note that the likelihood above does not depend on the unobserved Y_i since $R_i = 0$ makes the contribution of the corresponding term 0.

In the remainder of this section, we focus on the singular regime, i.e., when the true parameter $\theta^* = 0$ and consequently the probability of observing any sample $Y_i = y$ is always $1/2$ (independent of the value y). For such a setting, the results of Rotnitzky et al. (2000) imply that the statistical error of the MLE is larger than the parametric rate $n^{-\frac{1}{2}}$. In particular, they showed that $|\hat{\theta}_{n, \text{MLE}} - \theta^*| = \mathcal{O}(n^{-\frac{1}{4}})$. However, with high probability, the log-likelihood $\bar{\mathcal{L}}_n$ is non-concave and thereby a closed-form for the maximum-likelihood estimate is not available. Thus a theoretical analysis of the estimates obtained via different optimization algorithms (that can be used to maximize the log-likelihood $\bar{\mathcal{L}}_n$) can be of significant interest. We now apply our general theory to analyze two optimization methods: (i) gradient ascent, and (ii) Newton's method.

4.1.1 THEORETICAL GUARANTEES

We now state a theoretical guarantee on the behavior of the optimization algorithms in practice with the informative non-response model (22)—that is, when applied to the sample log likelihood (23). We analyze the gradient ascent updates for a step-size $\eta \in (0, \frac{8}{3})$, and the pure Newton updates. We use M_n^{GA} and M_n^{NM} respectively to denote the sample-based operators for gradient ascent and Newton's method (see Appendix D.1 for the precise form of these operators). The following statement also involves other universal constants c, c_i, c'_i, c''_i etc.

Corollary 7 *For the singular setting of informative non-response model ($\theta^* = 0$) and given some $\delta \in (0, 1)$, the following properties hold with probability at least $1 - \delta$:*

. For instance, when $\sum_{i=1}^n R_i(Y_i^2 + 1) < n$, the sample log-likelihood function is bimodal and symmetric around 0.

- (a) For any fixed $\alpha \in (0, 1/4)$ and initialization $\theta^0 \in \mathbb{B}(\theta^*, 1/2)$, the sequence $\theta^t := (\mathbf{M}_n^{\text{GA}})^t(\theta^0)$ of gradient iterates satisfies the bound

$$|\theta^t - \theta^*| \leq c_1 \left(\frac{\log(\frac{\log(1/\alpha)}{\delta})}{n} \right)^{\frac{1}{4}-\alpha} \quad \text{for all iterates } t \geq c'_1 \sqrt{n} \log \frac{1}{\alpha}, \quad (24a)$$

as long as $n \geq c''_1 \log \frac{\log(1/\alpha)}{\delta}$.

- (b) For any initialization $\theta^0 \in \mathbb{A}(\theta^*, \sqrt{2c}(\log(1/\delta)/n)^{1/4}, 1/2)$, the sequence of Newton iterates $\theta^t := (\mathbf{M}_n^{\text{NM}})^t(\theta^0)$ satisfies the bound

$$|\theta^t - \theta^*| \leq c_2 \left(\frac{\log(1/\delta)}{n} \right)^{\frac{1}{4}} \quad \text{for all iterates } t \geq c'_2 \log n, \quad (24b)$$

as long as $n \geq c''_2 \log(1/\delta)$.

See Appendix D.1 for the proof of this corollary (and below for the proof sketch).

Corollary 7 shows that given n samples, (i) the final statistical errors achieved by the iterates generated by the gradient descent and the Newton's method are similar (of order $n^{-\frac{1}{4}}$), and (ii) the Newton's method takes a considerably smaller number (of order $\log n$) of steps in comparison to that taken by gradient ascent (of order \sqrt{n}). Finally, in Appendix D.1, we show that all the non-zero fixed points of the considered operators have a magnitude of the order $n^{-\frac{1}{4}}$ with constant probability. Therefore, the statistical radius achieved by the given optimization methods are optimal.

4.1.2 PROOF SKETCH FOR COROLLARY 7

Our proof of Corollary 7 starts with an analysis of the gradient ascent and Newton iterates on the population-level analog of the problem. In particular, taking expectations in equation (23), we obtain the following population-level optimization problem

$$\max_{\theta \in \mathbb{R}} \bar{\mathcal{L}}(\theta) \quad \text{where} \quad \bar{\mathcal{L}}(\theta) = \frac{1}{2} \log \left(1 - \frac{1}{2\sqrt{\theta^2 + 1}} \right) - \frac{\theta^2 + 1}{4}. \quad (25)$$

Let \mathbf{M}^{GA} denote the gradient update operator applied to this objective with a given step-size η , and let \mathbf{M}^{NM} denote the Newton update. In Appendix D.1 (where we also provide explicit forms of these operators), we show that with $\theta^* = 0$, the population-level operators have the following properties:

- (P1) The gradient operator \mathbf{M}^{GA} is $\text{SLOW}(\beta)$ -convergent with parameter $\beta = \frac{1}{2}$ over the Euclidean ball $\mathbb{B}(\theta^*, \frac{1}{2})$, i.e., for the sequence $\theta^t = (\mathbf{M}^{\text{GA}})^t(\theta^0)$ with $\theta^0 \in \mathbb{B}(\theta^*, \frac{1}{2})$, we have $|\theta^t - \theta^*| \leq \frac{c}{t^{1/2}}$.
- (P2) The Newton operator \mathbf{M}^{NM} is $\text{FAST}(\kappa)$ -convergent with parameter $\kappa = \frac{4}{5}$ over the Euclidean ball $\mathbb{B}(\theta^*, \frac{1}{2})$, i.e., for the sequence $\theta^t = (\mathbf{M}^{\text{NM}})^t(\theta^0)$ with $\theta^0 \in \mathbb{B}(\theta^*, \frac{1}{2})$, we have $|\theta^t - \theta^*| \leq c e^{-\kappa t}$.

Moreover in the same Appendix D.1, we show that with the noise function $\varepsilon(n, \delta) = \sqrt{\frac{\log(1/\delta)}{n}}$, the sample-level operators satisfy the following properties:

- (S1) The sample-based gradient ascent operator M_n^{GA} is $\text{STA}(\gamma)$ -stable with parameter $\gamma = 1$ over the ball $\mathbb{B}(\theta^*, \frac{1}{2})$, and
- (S2) the operator M_n^{NM} is $\text{UNS}(\gamma)$ -unstable with parameter $\gamma = -1$ over the annulus $\mathbb{A}(\theta^*, \tilde{\rho}_n, \rho)$ with $\tilde{\rho}_n = c[\varepsilon(n, \delta)]^{\frac{1}{2}}$ and $\rho = \frac{1}{2}$ where c denotes some universal positive constant.

Given these properties, we now show how our general theory yields the results stated in Corollary 7. To simplify the following discussion, we omit the universal constants and a few-logarithmic terms, and track the dependency only on the sample size n .

Results for gradient ascent: The items (P1) and (S1) establish that the gradient operators are slow-convergent and stable, and thus we can apply our general result from Theorem 5. In particular, plugging $\beta = \frac{1}{2}$, and $\gamma = 1$ in Theorem 5, we find that the statistical error for the gradient iterates $\theta^t = (M_n^{\text{GA}})^t(\theta^0)$ satisfies

$$|\theta^t - \theta^*| \lesssim [\varepsilon(n, \delta)]^{\frac{\beta}{1+\beta-\gamma\beta}} \asymp [n^{-\frac{1}{2}}]^{\frac{1/2}{1+1/2-1/2}} = n^{-\frac{1}{4}}, \quad (26a)$$

$$\text{for } t \gtrsim [\varepsilon(n, \delta)]^{-\frac{1}{1+\beta-\gamma\beta}} \asymp [n^{-\frac{1}{2}}]^{-\frac{1}{1+1/2-1/2}} = n^{\frac{1}{2}}. \quad (26b)$$

Results for Newton's method: The items (P2) and (S2) establish that the Newton operators are fast-convergent but unstable, and as a consequence our general result from Theorem 6(a) can be applied. In particular, plugging $\gamma = -1$ in Theorem 6(a), we find that the Newton iterates $\theta^t = (M_n^{\text{NM}})^t(\theta^0)$ satisfy

$$\begin{aligned} |\theta^t - \theta^*| &\lesssim \max \left\{ [\varepsilon(n, \delta)]^{\frac{1}{1+|\gamma|}}, \tilde{\rho}_n \right\} \\ &\asymp [n^{-\frac{1}{2}}]^{\frac{1}{1+1}} = n^{-\frac{1}{4}} \quad \text{for } t \gtrsim \log(1/\varepsilon(n, \delta)) \asymp \log n. \end{aligned} \quad (27)$$

Moreover, we show that (see the discussion around equation (87)) Assumption (D) holds for the Newton iterates with an initialization outside the ball $\mathbb{B}(\theta^*, \tilde{\rho}_n)$, and hence part (b) of the Corollary 7 states that the Newton iterates stay in a close vicinity of θ^* for all future iterations.

4.2 Over-specified Gaussian mixture models

We now consider the problem of parameter estimation in Gaussian mixture models; and analyze the behavior of two popular algorithms namely (a) Expectation-Maximization (EM) algorithm (Dempster et al., 1997), and (b) Newton's method. We note that EM is arguably the most widely used algorithm for parameter estimation in mixture models and other missing data problems (Dempster et al., 1997). Here we study the problem of estimating the parameters of a Gaussian mixture model given n i.i.d. samples from the model. When the number of components in the mixture is known, prior works (Balakrishnan et al., 2017; Daskalakis et al., 2017; Cai et al., To Appear) have shown that (i) the mixture parameters can be estimated at the parametric rate $n^{-\frac{1}{2}}$ with the EM algorithm and (ii) the algorithm

takes at most $\log n$ steps to converge. In the over-specified setting, i.e., when the fitted model has more components than the true model, recent works (Dwivedi et al., 2020a,b; Wu and Zhou, 2019) have established the slow convergence of EM on both the statistical and algorithmic fronts. For example, for over-specified Gaussian-location mixtures EM takes $n^{\frac{1}{2}} \gg \log n$ steps (where \gg denotes much greater than) to converge and produces an estimate for the mean parameter that has a statistical error of order $n^{-\frac{1}{4}} \gg n^{-\frac{1}{2}}$.

In the sequel, we apply our general theory to study the behavior of EM and Newton’s method for parameter estimation in over-specified Gaussian-location mixtures. First, we recover the slow convergence of EM as derived in prior works (Dwivedi et al., 2020a). Second, we prove that the Newton’s method—although an unstable algorithm in this setting—achieves a similar statistical accuracy as EM albeit in an exponentially fewer number of steps. We now formalize the details. Let $\phi(\cdot; \theta, \sigma^2)$ denote the density of $\mathcal{N}(\theta, \sigma^2)$ random variable, i.e.,

$$\phi(x; \theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \quad (28a)$$

and let X_1, \dots, X_n be n i.i.d. draws from the standard normal distribution (density $\phi(\cdot; 0, 1)$). Given this data, we fit an over-specified mixture model namely, a two-component symmetric Gaussian mixture with equal fixed weights whose density is given by

$$f_\theta(x) = \frac{1}{2}\phi(x; -\theta, 1) + \frac{1}{2}\phi(x; \theta, 1), \quad (28b)$$

where θ is the parameter to be estimated. In such a setting, the true parameter is unique and given by $\theta^* = 0$ since $f_0(\cdot) = \phi(\cdot; 0, 1)$. However, the fact that we fit a mixture that has one extra component than the true model (which has just one component) leads to interesting consequences as we now elaborate. Using \mathcal{L}_n to denote the log-likelihood function, the MLE estimate is given by

$$\hat{\theta}_{n,\text{MLE}} \in \arg \max_{\theta \in \mathbb{R}} \mathcal{L}_n(\theta) \quad \text{where} \quad \mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^n \log f_\theta(X_i). \quad (28c)$$

On one hand, it is known (Chen, 1995) that the over-specification in such a setting leads to a slower than $n^{-\frac{1}{2}}$ statistical rate for the MLE, i.e., $|\hat{\theta}_{n,\text{MLE}} - \theta^*| = \mathcal{O}(n^{-\frac{1}{4}})$. On the other hand, MLE does not admit a closed-form expression and thus it is of significant interest to understand the behavior of iterative algorithms that are used to estimate the MLE. Next, we use our general framework to provide a precise characterization of two algorithms namely, EM, and Newton’s method on maximizing the log-likelihood \mathcal{L}_n (28c).

4.2.1 THEORETICAL GUARANTEES

The next corollary provides a precise characterization of EM and Newton’s method for the over-specified setting described in the previous section. We analyze the EM updates and the pure Newton updates. Moreover, we use G_n^{EM} and G_n^{NM} respectively to denote the sample-based operators for EM and Newton’s method (see Appendix D.2 for the precise form of these operators). Finally, the scalars c, c_i, c'_i, c''_i denote some positive universal constants.

Corollary 8 *For the over-specified Gaussian mixture model (28) with $\theta^* = 0$, given some $\delta \in (0, 1)$, the following properties hold with probability at least $1 - \delta$:*

- (a) *For any fixed $\alpha \in (0, 1/4)$ and initialization $\theta^0 \in \mathbb{B}(\theta^*, 1)$, the sequence $\theta^t := (G_n^{\text{EM}})^t(\theta^0)$ of EM iterates satisfies the bound*

$$|\theta^t - \theta^*| \leq c_1 \left(\frac{\log(\frac{\log(1/\alpha)}{\delta})}{n} \right)^{\frac{1}{4}-\alpha} \quad \text{for all iterates } t \geq c'_1 \sqrt{n} \log \frac{1}{\alpha}, \quad (29a)$$

as long as $n \geq c''_1 \log \frac{\log(1/\alpha)}{\delta}$.

- (b) *For any initialization $\theta^0 \in \mathbb{A}(\theta^*, \frac{\sqrt{2c} \log^2(3n/\delta)}{n^{1/4}}, 1/3)$, the sequence of Newton iterates $\theta^t := (G_n^{\text{NM}})^t(\theta^0)$ satisfies the bound*

$$|\theta^t - \theta^*| \leq c_2 \left(\frac{\log(n/\delta)}{n} \right)^{\frac{1}{4}} \quad \text{for all iterates } t \geq c'_2 \log n, \quad (29b)$$

as long as $n \geq c''_2 \log(1/\delta)$.

See Appendix D.2 for the proof (and below for the proof sketch).

Corollary 8 establishes that the Newton EM is significantly faster than EM for the model setup (28). More precisely, it reaches ball around θ^* with a statistical radius of order $n^{-\frac{1}{4}}$ within $\log n$ steps, which is much smaller than the number of steps taken by EM. Moreover, the updates from Newton's method do not escape this ball for future iterations. This behavior is a consequence of the fact that under the assumed initialization condition, the (cubic-regularized) Newton EM sequence satisfies assumption (D).

Multivariate settings: In Figure 4, we discuss the performance of EM and Newton's method under the multivariate setting of the over-specified Gaussian mixture model (28b). Similar to the univariate setting, both algorithms converge to a statistical error of order $(d/n)^{1/4}$ around the true parameter θ^* . Furthermore, the EM algorithm takes $\sqrt{n/d}$ number of iterations to converge to the final estimate (see Appendix E.1 for a formal result) while the Newton's method takes much fewer number of iterations (which seems in agreement with the $\log n$ scaling suggested by our theory). Given that each iteration of the EM algorithm takes order $n \cdot d$ arithmetic operations, the computational complexity for the EM algorithm to reach the final estimate is of order $n^{3/2}d^{1/2}$. On the other hand, each iteration of the Newton's method takes an order of $n \cdot d + d^3$ arithmetic operations where d^3 is computational complexity of computing inverse of an $d \times d$ matrix via Gauss-Jordan elimination approach. It leads to the computational complexity at the order $(nd + d^3) \log n$ for the Newton's method to reach to the final estimate. Thus, when $d^{5/3} \ll n$, Newton's method is computationally more efficient than the EM algorithm.

4.2.2 PROOF SKETCH FOR COROLLARY 8

The proof strategy for this case is similar to that laid out in Section 4.1.2 for informative non-response model. First, to study this problem in our framework, we consider the population

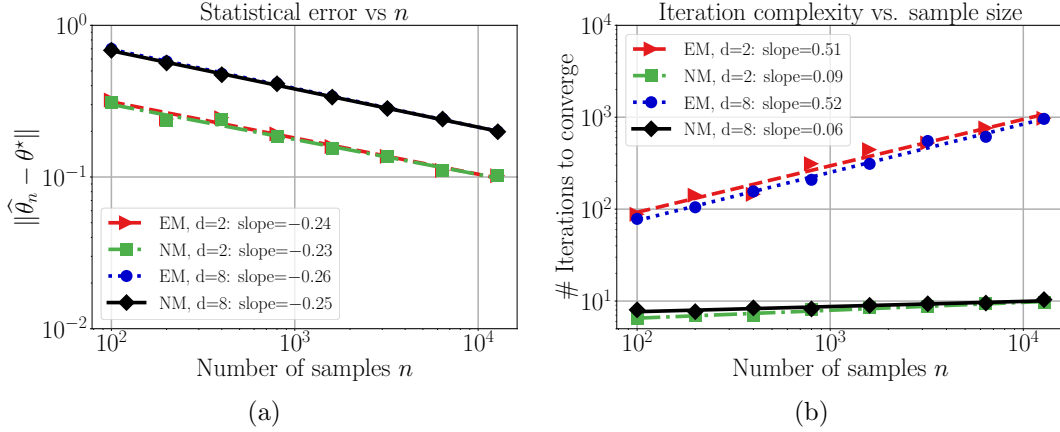


Figure 4. Plots characterizing the behavior of Expectation-Maximization (EM) and Newton’s method (NM) for two Gaussian mixture models in $d = 2$ and $d = 8$ dimensions. (a) Log-log plots of the Euclidean distance $\|\hat{\theta}_n - \theta^*\|_2$ versus the sample size. It shows that all the algorithms converge to an estimate at Euclidean distance of the order $n^{-1/4}$ from the true parameter θ^* . (b) Log-log plots for the number of iterations taken by different algorithms to converge to the final estimate. While EM takes roughly \sqrt{n} iterations, the scaling of iterations taken by Newton’s method is significantly slower.

level objective \mathcal{L} by replacing the sum over samples in equation (28c) with the corresponding expectation:

$$\mathcal{L}(\theta) := \mathbb{E}_{X \sim \mathcal{N}(0,1)} [\log f_\theta(X)] = \mathbb{E}_X \left[\frac{1}{2} \phi(X; -\theta, 1) + \frac{1}{2} \phi(X; \theta, 1) \right]. \quad (30)$$

Second, we use G^{EM} and G^{NM} respectively to denote the corresponding population-level EM and Newton’s method operators (see Appendix D.2 for the precise expressions).

Results for EM: For the case of $\theta^* = 0$, Theorem 2 and Lemma 1 of our prior works (Dwivedi et al., 2020a) show that, for any initialization θ^0 , the EM operators G^{EM} and G_n^{EM} satisfy

$$\begin{aligned} |(G^{\text{EM}})^t(\theta^0) - \theta^*| &\leq \frac{c}{t^{1/2}} \quad \text{and,} \\ \sup_{\theta \in \mathbb{B}(\theta^*, r)} |G^{\text{EM}}(\theta) - G_n^{\text{EM}}(\theta)| &\leq c_1 r \cdot \sqrt{\frac{\log(1/\delta)}{n}}, \end{aligned} \quad (31)$$

where the second bound holds with probability at least $1 - \delta$ for any fixed radius $r > 0$. In the framework of our current work, the bounds (31) imply that the operator G^{EM} exhibits $\text{SLOW}(\frac{1}{2})$ -convergence, and the operator G_n^{EM} is $\text{STA}(1)$ -stable with the noise function $\sqrt{\frac{\log(1/\delta)}{n}}$. Thus a direct application of Theorem 5 of this paper (in a fashion similar to that of equations (26a) and (26b)), recovers the main result of our prior work (Dwivedi et al., 2020a) (Theorem 3). That is, with high probability, the sequence $\theta_n^{t+1} = G_n^{\text{EM}}(\theta_n^t)$ satisfies

$$|\theta^t - \theta^*| \lesssim [n^{-1/2}]^{\frac{1/2}{1+1/2-1/2}} = n^{-1/4} \quad \text{for} \quad t \gtrsim [n^{-1/2}]^{-\frac{1}{1+1/2-1/2}} = n^{1/2}. \quad (32)$$

Results for Newton's method: In Appendix D.2, we demonstrate the following properties of Newton's method operators:

- (M1) the Newton operator G^{NM} is $\text{FAST}(\frac{7}{9})$ -convergent over the ball $\mathbb{B}(\theta^*, \frac{1}{3})$, and
- (M2) the operator G_n^{NM} is $\text{UNS}(-1)$ -unstable over the annulus $\mathbb{A}(\theta^*, \tilde{\rho}_n, 1/3)$ with noise function $\varepsilon(n, \delta) = \frac{\log(n/\delta)}{\sqrt{n}}$ where $\tilde{\rho}_n = \frac{c \log^2(3n/\delta)}{n^{1/4}}$.

Based on the results of Theorem 6(a) with $\kappa = \frac{7}{9}$ and $\gamma = -1$, the items (M1) and (M2) suggest that the Newton updates $\theta^t = (M_n^{\text{NM}})^t(\theta^0)$ satisfy

$$|\theta^t - \theta^*| \lesssim \max \left\{ [\varepsilon(n, \delta)]^{\frac{1}{1+1}}, \tilde{\rho}_n \right\} \lesssim n^{-\frac{1}{4}} \quad \text{for } t \gtrsim \log(1/\varepsilon(n, \delta)) \asymp \log n. \quad (33)$$

Furthermore, we prove that the Newton iterates satisfy Assumption (D) (see the argument with equation (97)). Therefore, the Newton iterates stay in a close vicinity of θ^* for all future iterations.

4.3 Non-linear regression model

In our third example, we consider a non-linear regression model (Carroll et al., 1997) with a known link function g . Models of this type have proven useful for applications in signal processing, econometrics, statistics, and machine learning (Ichimura, 1993; Horowitz and Härdle, 1996). For simplicity, we briefly summarize the one-dimensional version of this problem. The multivariate setting of the problem is considered in Appendix E.2. We observe the pairs of data $(X_i, Y_i) \in \mathbb{R}^2$ that are generated from the model

$$Y_i = g(X_i \theta^*) + \xi_i \quad \text{for } i = 1, \dots, n. \quad (34a)$$

Here Y_i denotes the response variable, X_i corresponds to the covariate and ξ_i denotes the additive noise assumed to have a standard Gaussian distribution, i.e., $\xi_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Note that, the Gaussianity of the additive noise is for the simplicity of the proof, and the results can be extended to sub-Gaussian errors.

In this example, we consider the case of random design for the covariates, i.e., the covariates $\{X_i\}_{i=1}^n$ are independent and $X_i \sim \mathcal{N}(0, 1)$. Given the samples $\{(X_i, Y_i), i \in [n]\}$, we want to estimate the unknown parameter θ^* . A popular choice is the maximum-likelihood estimate (MLE):

$$\hat{\theta}_n^{\text{mle}} \in \arg \min_{\theta \in \mathbb{R}} \tilde{\mathcal{L}}_n(\theta) \quad \text{where} \quad \tilde{\mathcal{L}}_n := \frac{1}{2n} \sum_{i=1}^n (Y_i - g(X_i \theta))^2. \quad (34b)$$

Generally, the loss-function $\tilde{\mathcal{L}}_n$ is non-convex and hence the MLE does not admit a closed-form expression. Consequently, one needs to make use of certain optimization algorithms to compute an estimate $\hat{\theta}_n$, which need not be the same as $\hat{\theta}_n^{\text{mle}}$.

In the remainder of this section, we study the case when the SNR degenerates to zero. Specifically, we consider $\theta^* = 0$ and a link function of the form $g(x) = x^{2p}$ with $p \geq 1$. For such a setting, the optimization problem (34b) takes the following form:

$$\hat{\theta}_n \in \arg \min_{\theta \in \mathbb{R}} \tilde{\mathcal{L}}_n(\theta) \quad \text{where} \quad \tilde{\mathcal{L}}_n := \frac{1}{2n} \sum_{i=1}^n \left(Y_i - (X_i \theta)^{2p} \right)^2. \quad (34c)$$

4.3.1 THEORETICAL GUARANTEES

For the non-linear regression model described above with the link function $g(x) = x^{2p}$, we consider three iterative optimization methods: (a) gradient descent with a step size $\eta \in (0, \frac{1}{(4p-1)!!(2p)}]$, (b) (pure) Newton's method, and (c) cubic-regularized Newton's method with Lipschitz constant $L := (4p-1)!!(4p-1)p/3$. We denote the updates for these three methods via the operators F_n^{GD} , F_n^{NM} , and F_n^{CNM} respectively (see Appendix D.3 for the precise expressions of these operators). The next result characterizes the behavior of these three methods:

Corollary 9 *For the non-linear regression model (34) with link function $g(x) = x^{2p}$ for $p \geq 1$ and true parameter $\theta^* = 0$, given some $\delta \in (0, 1)$, the following properties hold with probability at least $1 - \delta$:*

- (a) *For any fixed $\alpha \in (0, 1/4)$ and initialization $\theta^0 \in \mathbb{B}(\theta^*, 1)$, the sequence $\theta^t := (F_n^{\text{GD}})^t(\theta^0)$ of gradient iterates satisfies the bound*

$$|\theta^t - \theta^*| \leq c_1 \left(\frac{\log^{4p}(n \frac{\log(1/\alpha)}{\delta})}{n} \right)^{\frac{1}{4p} - \alpha} \quad \text{for all iterates } t \geq c'_1 n^{\frac{2p-1}{2p}} \log \frac{1}{\alpha}, \quad (35a)$$

as long as $n \geq c''_1 \log \frac{\log(1/\alpha)}{\delta}$.

- (b) *For any initialization $\theta^0 \in \mathbb{A}(\theta^*, c^{\frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}}, 1)$, the sequence of Newton iterates $\theta^t := (F_n^{\text{NM}})^t(\theta^0)$ satisfies the bound*

$$|\theta^t - \theta^*| \leq c_2 \left(\frac{\log^{4p}(n/\delta)}{n} \right)^{\frac{1}{4p}} \quad \text{for all iterates } t \geq c'_2 \log n, \quad (35b)$$

as long as $n \geq c''_2 \log(1/\delta)$.

- (c) *The sequence of cubic-regularized Newton iterates $\theta^t := (F_n^{\text{CNM}})^t(\theta^0)$ with initialization $\theta^0 \in \mathbb{A}(\theta^*, c^{\frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}}, 1)$ satisfies the bound*

$$|\theta^t - \theta^*| \leq c_3 \left(\frac{\log^{4p}(n/\delta)}{n} \right)^{\frac{1}{4p}} \quad \text{for all iterates } t \geq c'_3 n^{\frac{4p-3}{2(4p-1)}}, \quad (35c)$$

as long as $n \geq c''_3 \log(1/\delta)$.

See Appendix D.3 for the proof (and below for the proof sketch).

This corollary shows that the final statistical errors achieved by gradient descent and the (cubic-regularized) Newton's method have the same scaling. Moreover, Newton's method, while unstable, converges to the correct statistical radius in a significantly smaller $\log n$ number of steps when compared to gradient descent, which takes $n^{\frac{2p-1}{2p}}$ steps and cubic-regularized Newton's method, which takes $n^{\frac{4p-3}{2(4p-1)}}$ steps. Moreover, we also show that

assumption (D) holds for the iterates from the (cubic-regularized) Newton method's and hence we obtain that these iterates not only converge to a ball of radius $n^{-\frac{1}{4p}}$ around θ^* , but also that they stay there for all the future iterations. Finally, in Appendix D.3 (see equation (116)) we also establish that the statistical radius $n^{-1/(4p)}$ achieved by the considered optimization methods is tight.

When $g(x) = x^2$, the model (34a) corresponds to a phase retrieval problem. In the regime of large signal-to-noise ratio (SNR), i.e., $|\theta^*| \gg 1$, and with the link function $g(x) = x^2$, there are efficient algorithms which produce an estimate $\hat{\theta}_n$ satisfying a bound $|\hat{\theta}_n - \theta^*| \lesssim n^{-\frac{1}{2}}$ (Eldar and Mendelson, 2013; Candès et al., 2015; Tan and Vershynin, 2018). However, as the SNR approaches zero these parametric rates do not apply and precise statistical behavior of these estimates are not known.

4.3.2 PROOF SKETCH FOR COROLLARY 9

In order to study these updates using our framework, we need to consider the population-level version of the optimization problem (34c), which is given by

$$\min_{\theta \in \mathbb{R}} \tilde{\mathcal{L}}(\theta) \quad \text{where} \quad \tilde{\mathcal{L}}(\theta) := \frac{1}{2} \mathbb{E}_{(X,Y)} \left[\left(Y - (X\theta)^{2p} \right)^2 \right],$$

where the expectation is taken with respect to $X \sim \mathcal{N}(0, 1)$, $Y \sim \mathcal{N}(0, 1)$ as $\theta^* = 0$. Direct computation yields that

$$\tilde{\mathcal{L}}(\theta) = \frac{1}{2} + \frac{(4p-1)!!\theta^{4p}}{2} \quad \text{and} \quad \arg \min_{\theta} \tilde{\mathcal{L}}(\theta) = 0 = \theta^*. \quad (36)$$

Like the previous proof sketches, we let F^{GD} , F^{NM} and F^{CNM} denote the population operators corresponding to the algorithms, gradient descent, Newton's method and cubic-regularized Newton's method, for the problem (36) (for a given p). See Appendix D.3 for the precise definitions of these operators. In Appendix D.3, we show that with $\theta^* = 0$, these population-level operators satisfy the following properties over the ball $\mathbb{B}(\theta^*, 1)$:

($\tilde{\text{P1}}$) the gradient operator F^{GD} is $\text{SLOW}(\frac{1}{4p-2})$ -convergent for step size $\eta \in (0, \frac{1}{(4p-1)!!(2p)}]$,

($\tilde{\text{P2}}$) the Newton operator F^{NM} is $\text{FAST}(\frac{4p-2}{4p-1})$ -convergent, and

($\tilde{\text{P3}}$) the cubic-regularized Newton operator F^{CNM} is $\text{SLOW}(\frac{2}{4p-3})$ -convergent.

Moreover in the Appendix D.3, we show that with the noise function $\varepsilon(n, \delta) = \sqrt{\frac{\log^{4p}(n/\delta)}{n}}$, the sample-level operators satisfy the following properties:

($\tilde{\text{S1}}$) the operator F_n^{GD} is $\text{STA}(2p-1)$ -stable over the ball $\mathbb{B}(\theta^*, 1)$,

($\tilde{\text{S2}}$) the operator F_n^{NM} is $\text{UNS}(-(2p-1))$ -unstable over the annulus $\mathbb{A}(\theta^*, \tilde{\rho}_n, 1)$ with inner radius $\tilde{\rho}_n = c \log^{p/(2p-1)}(n/\delta)/n^{1/4(2p-1)}$, and

($\tilde{\text{S3}}$) the operator F_n^{CNM} is $\text{UNS}(-\frac{1}{2})$ -unstable over the annulus $\mathbb{A}(\theta^*, \tilde{\rho}_n, 1)$.

. See the proofs of equations (111) and (117) in Appendix D.3 for more details.

These properties show that the gradient descent is a slow-converging stable method and we can apply Theorem 5. On the other hand, Newton’s method is a fast-converging unstable method, and Theorem 6(a) can be applied. Finally, cubic-regularized Newton’s method is a slow-converging unstable method and Theorem 6(b) can be applied. In the subsequent proof-sketch, we track the dependency only on the sample size n and ignore logarithmic factors and universal constants. Moreover, since the computations here mimic the discussion from Section 4.1.2, we keep the discussion briefer.

Results for gradient descent: Applying Theorem 5 with $\beta = \frac{1}{4p-2}$, and $\gamma = 2p - 1$ (items $(\tilde{P}1)$ and $(\tilde{S}1)$ respectively), we find that the statistical error for the gradient iterates $\theta^t = (F_n^{\text{GD}})^t(\theta^0)$ satisfy

$$|\theta^t - \theta^*| \lesssim [\varepsilon(n, \delta)]^{\frac{\beta}{1+\beta-\gamma\beta}} \lesssim n^{-\frac{1}{2p}} \quad \text{for } t \gtrsim [\varepsilon(n, \delta)]^{-\frac{1}{1+\beta-\gamma\beta}} \asymp n^{\frac{2p-1}{2p}}. \quad (37)$$

Results for Newton’s method: Next applying Theorem 6(a) for the Newton’s method with $\kappa = \frac{4p-2}{4p-1}$, and $\gamma = -(2p - 1)$ (see items $(\tilde{P}2)$ and $(\tilde{S}2)$), we conclude that the updates $\theta^t = (F_n^{\text{NM}})^t(\theta^0)$ from the Newton’s method have the following property:

$$|\theta^t - \theta^*| \lesssim \max \left\{ [\varepsilon(n, \delta)]^{\frac{1}{1+|\gamma|}}, \tilde{\rho}_n \right\} \lesssim n^{-\frac{1}{2p}} \quad \text{for } t \gtrsim \log(1/\varepsilon(n, \delta)) \asymp \log n. \quad (38)$$

Results for cubic-regularized Newton’s method: Finally by using Theorem 6(b) for the cubic-regularized Newton’s method with $\beta = \frac{2}{4p-3}$, and $\gamma = -\frac{1}{2}$ (see items $(\tilde{P}3)$ and $(\tilde{S}3)$), the following results hold for the cubic-regularized Newton iterates $\theta^t = (F_n^{\text{CNM}})^t(\theta^0)$:

$$\begin{aligned} |\theta^t - \theta^*| &\lesssim \max \left\{ [\varepsilon(n, \delta)]^{\frac{\beta}{1+\beta-\gamma\beta}}, \tilde{\rho}_n \right\} \\ &\lesssim n^{-\frac{1}{2p}} \quad \text{for } t \gtrsim [\varepsilon(n, \delta)]^{-\frac{1}{1+\beta}} \asymp n^{\frac{4p-3}{2(4p-1)}}. \end{aligned} \quad (39)$$

5. Discussion

In this paper, we established several results characterizing the statistical radius achieved by a sequence of updates $\{F_n^t(\theta_n^0)\}_{t \geq 0}$, induced by an operator F_n and a given initial point θ_n^0 . We established these results by analyzing the interplay between (in)-stability of the operator F_n for its population operator F and the local convergence of F around its fixed point θ^* . We then applied our general theory to derive sharp algorithmic and statistical guarantees for several iterative algorithms by analyzing the corresponding sample and population operators, in three different statistical settings. In particular, we studied the behavior of gradient methods and higher-order (cubic-regularized) Newton’s method for parameter estimation—in the weak signal-to-noise ratio regime—in Gaussian mixture models, non-linear regression models, and informative non-response models. We showed that for such models, despite instability, fast algorithms like Newton’s method may still be preferred over a stable one like gradient descent since they achieve the same statistical accuracy as that of the stable counterpart in exponentially fewer steps.

We now discuss a few questions that arise naturally from our work. First, our results, as stated, are not directly applicable to the settings of accelerated optimization methods or quasi-Newton methods, e.g., accelerated gradient descent (Nesterov, 2013) and L-BFGS (Fletcher, 1987). On the one hand, the updates from an accelerated gradient descent

method require that the operators F_n and F to change with each iteration. On the other hand, the updates from the L-BFGS method would require additional machinery to deal with the preconditioning matrices in each step. Developing a general theory to characterize the statistical performance of algorithms associated with a time-varying operator F_n is an interesting direction for future research.

Secondly, it is desirable to understand the behavior of optimization methods to a wider range of statistical problems. In the context of mixture models, recent work by Dwivedi et al. (Dwivedi et al., 2020b) established that for over-specified mixtures with both location and scale parameter unknown, EM takes an $\mathcal{O}(n^{\frac{3}{4}})$ steps to return estimates with minimax statistical error of order $n^{-\frac{1}{8}}$ and $n^{-\frac{1}{4}}$ for the location and scale parameter, respectively. Whether an unstable method like (cubic-regularized) Newton’s EM proves computationally advantageous (without losing statistical accuracy) in such more challenging non-convex landscapes remains an open problem.

Finally, our theory does not easily extend to the settings with dependent data, such as time series. When the samples are (time) dependent, taking the limit of infinite sample size does not yield a natural population-level operator. One possible fix is to borrow the technique of truncating the sample operator from the analysis of the Baum-Welch algorithm for hidden Markov models (Yang et al., 2017). However, even with the help of such a technique, ample technical challenges remain towards developing a general theory for such non-i.i.d. settings.

In this supplementary material, we provide the details of proofs and results that were deferred from the main paper. Appendices A and C contain the proofs of Theorems 5 and 6, respectively, including all the details of the localization argument and the proofs of all auxiliary technical lemmas. In Appendix B, we construct a simple class of problems to demonstrate that the guarantees Theorems 5 and 6 are unimprovable in general. Finally, in Appendix D, we collect the proofs of several corollaries stated in the paper. Finally, we discuss an extension of the theoretical results in the main text to multivariate settings in Appendix E.

Appendix A. Proofs of main results

In this section, we provide the proofs of our main results, namely Theorems 5 and 6.

A.1 Proof of Theorem 5

The reader should recall the proof outline provided following the statement of the theorem. Our proof here follows this outline, making each step precise. For the remainder of the proof, we assume without loss of generality that $\theta^* = 0$ and $r_0 = 1$. Proofs for the cases $\theta^* \neq 0$ or $r_0 > 1$ can be reduced to this case in a straightforward fashion and are thereby omitted.

A.1.1 NOTATION FOR STABLE CASE

For each positive integer $\ell = 1, 2, \dots$, let T_ℓ denote the number of iterations during the ℓ -th epoch, and let S_ℓ denote the total number of iterations taken up to the completion of epoch

ℓ . In order to describe some recursions satisfied by these quantities, we define

$$\begin{aligned} T_\ell^{(1)} &:= C\varepsilon(n, \delta^*)^{-\frac{\lambda_\ell - 1(\gamma) + 1}{1+\beta}} \quad \text{and} \quad T_\ell^{(2)} := C'\varepsilon(n, \delta^*)^{-\frac{\lambda_\ell(\gamma) + 1}{1+\beta}}, \\ \text{for } C &:= (c_2 2^\gamma)^{-\frac{1}{(1+\beta)}} \quad \text{and} \quad C' := C(c')^{\frac{\gamma}{1+\beta}}, \end{aligned} \quad (40a)$$

where $c' := (c_2 2^\gamma)^{\frac{\beta}{1+\beta}} = C^{-\beta}$ and hence we have $C' = C^{\frac{1+\beta+\beta\gamma}{1+\beta}}$. Here the constant c_2 is the constant from the the stability definition (11). The sequences $\{T_\ell\}$ and $\{S_\ell\}$ have the following properties: with the initialization $T_0 := 0$, we have

$$T_\ell := \left\lceil T_\ell^{(1)} + T_\ell^{(2)} \right\rceil \quad \text{and} \quad S_\ell := \sum_{j=0}^{\ell} T_j \quad \text{for } \ell = 1, 2, \dots \quad (40b)$$

Our proof is based on studying the sequence of real-numbers $\{\lambda_\ell\}_{\ell \geq 0}$ given by

$$\lambda_0 = 0 \quad \text{and} \quad \lambda_{\ell+1} = \lambda_\ell \nu + \nu', \quad \text{where } \nu = \frac{\beta\gamma}{1+\beta} \text{ and } \nu' = \frac{\beta}{1+\beta}. \quad (40c)$$

Note that Assumption (B) implies that $\nu \in (0, 1)$ and hence

$$\lambda_\ell = \nu_\star(1 - \nu^\ell) \uparrow \nu_\star \quad \text{where} \quad \nu_\star := \frac{\beta}{1 + \beta - \gamma\beta}, \quad (40d)$$

where $\lambda_\ell \uparrow \nu_\star$ means that the sequence λ_ℓ is monotonically increasing and converges to ν_\star . In the epoch-based argument, we need to control the deviation $\sup_{\|\theta\| \leq r} \|F(\theta) - F_n(\theta)\|$ uniformly for each radii $r \in \mathcal{R}'$. To this end, for any tolerance $\delta \in (0, 1)$, we define the event \mathcal{E} by

$$\mathcal{E} := \left\{ \sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|F(\theta) - F_n(\theta)\| \leq c_2 r^\gamma \varepsilon(n, \delta^*) \quad \text{uniformly for all } r \in \mathcal{R}' \right\}, \quad (41)$$

where $\delta^* = \delta \cdot \frac{\log(\frac{1+\beta}{\beta\gamma})}{8 \log(\frac{\beta}{\alpha(1+\beta-\gamma\beta)})}$ was defined in equation (14) and the radii-set \mathcal{R}' is defined as

$$\begin{aligned} \mathcal{R}' &:= \mathcal{R} \cup 2\mathcal{R}, \quad \text{with} \\ \mathcal{R} &:= \left\{ \varepsilon(n, \delta^*)^{\lambda_0}, \dots, \varepsilon(n, \delta^*)^{\lambda_{\ell_\alpha}}, c' \varepsilon(n, \delta^*)^{\lambda_0}, \dots, c' \varepsilon(n, \delta^*)^{\lambda_{\ell_\alpha}} \right\}, \\ \ell_\alpha &= \lceil \log(1/\alpha) \rceil \quad \text{and} \quad c' = (c_2 2^\gamma)^{\frac{\beta}{1+\beta}}. \end{aligned} \quad (42)$$

Combining the STA(γ)-stability assumption (11) with a standard application of union bound we conclude that

$$\mathbb{P}(\mathcal{E}) \geq 1 - \delta. \quad (43)$$

Before we start the main argument, we state a lemma useful in the proof of our theorem:

Lemma 10 *Assume that the assumptions of Theorem 5 are in force. Then conditioned on the event \mathcal{E} (41. 43), for all radius r in the set \mathcal{R} (42), we have*

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|F^t(\theta) - F_n^t(\theta)\| \leq c_2 (2r)^\gamma \varepsilon(n, \delta^*) \cdot t \quad \text{for all } t \leq \tilde{\mathcal{T}}(r), \quad (44)$$

where $\tilde{\mathcal{T}}(r) := \frac{r^{1-\gamma}}{2^\gamma c_2 \varepsilon(n, \delta^*)}$. Furthermore, for all $\ell \leq \ell_\alpha$ we have

$$T_{\ell+1}^{(1)} \leq \tilde{\mathcal{T}}(\varepsilon(n, \delta^*)^{\lambda_\ell}) \quad \text{and} \quad T_{\ell+1}^{(2)} \leq \tilde{\mathcal{T}}(c' \varepsilon(n, \delta^*)^{\lambda_{\ell+1}}). \quad (45)$$

See Appendix C.1 for the proof of this lemma.

A.1.2 MAIN ARGUMENT

We claim that the sequence $\{\theta_n^t\}_{t \geq 1}$ satisfies

$$\|\theta_n^{S_\ell}\|_2 \leq \varepsilon(n, \delta^*)^{\lambda_\ell} \quad \text{uniformly for all } \ell \in \{0, 1, \dots, \ell_\alpha\}, \quad \text{and} \quad (46a)$$

$$\|\theta_n^{S_{\ell_\alpha} + t}\| \leq 2\varepsilon(n, \delta^*)^{\nu_\star - \alpha} \quad \text{uniformly for all } t \in \{0, 1, 2, \dots\}, \quad (46b)$$

with probability at least $1 - \delta$. The quantities λ_ℓ, S_ℓ and ℓ_α are defined in equations (40a) through equation (40c). With these claims at our disposal, it remains to prove an upper bound on the scalar S_{ℓ_α} . Towards this end, doing some straightforward algebra we find that

$$T_\ell \leq T_{\ell_\alpha} \leq c' \varepsilon(n, \delta^*)^{-\frac{\nu_\star}{\beta}} \quad \text{for any } 0 \leq \ell \leq \ell_\alpha. \quad (47)$$

Combining the above bounds on T_ℓ with the definition of S_ℓ from equation (40b) yields an upper bound on S_{ℓ_α} . Substituting the upper bound on S_{ℓ_α} in inequality (46b) yields the claimed bound (15) of Theorem 5. We now prove the claims (46a) and (46b) using induction.

A.1.3 PROOF OF CLAIM (46a)

We condition on the event \mathcal{E} defined in the equation (41), which occurs with probability at least $1 - \delta$, and establish the claim using induction on the epoch index ℓ . The base case $\ell = 0$ is immediate. We now establish the inductive step, i.e., given $\|\theta_n^{S_\ell}\| \leq \varepsilon(n, \delta^*)^{\lambda_\ell}$ for some $\ell \leq \ell_\alpha - 1$, we show that $\|\theta_n^{S_{\ell+1}}\| \leq \varepsilon(n, \delta^*)^{\lambda_{\ell+1}}$. We split the proof in two parts (primarily to handle the constants):

$$\|\theta_n^{S_\ell + T_{\ell+1}^{(1)}}\| \leq c' \varepsilon(n, \delta^*)^{\lambda_{\ell+1}} \quad \text{and} \quad (48a)$$

$$\|\theta_n^{S_\ell + T_{\ell+1}^{(1)} + T_{\ell+1}^{(2)}}\| \leq \varepsilon(n, \delta^*)^{\lambda_{\ell+1}}, \quad (48b)$$

where $c' > 1$ is a universal constant. These claims together imply the induction hypothesis and thereby the claim (46a).

Proof of claim (48a) Inequality (45) implies that $T_{\ell+1}^{(1)} \leq \tilde{\mathcal{T}}(\varepsilon(n, \delta^*)^{\lambda_\ell})$, and hence we can apply the bound (44) from Lemma 10 with $r = \varepsilon(n, \delta^*)^{\lambda_\ell} \in \mathcal{R}$ for any $t \leq T_{\ell+1}^{(1)}$. Applying the triangle inequality yields

$$\begin{aligned} \|\theta_n^{t+S_\ell}\| &= \|F_n^t(\theta_n^{S_\ell})\| \leq \|F^t(\theta_n^{S_\ell})\| + \|F^t(\theta_n^{S_\ell}) - F_n^t(\theta_n^{S_\ell})\| \\ &\stackrel{(i)}{\leq} \frac{1}{t^\beta} + \|F^t(\theta_n^{S_\ell}) - F_n^t(\theta_n^{S_\ell})\| \end{aligned} \quad (49)$$

$$\stackrel{(ii)}{\leq} \frac{1}{t^\beta} + c_2(2\varepsilon(n, \delta^*)^{\lambda_\ell})^\gamma \varepsilon(n, \delta^*)^t, \quad (50)$$

for any $t \leq T_{\ell+1}^{(1)}$; where step (i) follows from the $\text{SLOW}(\beta)$ -convergence (8) of the operator F along with the assumption that $\theta^* = 0$, and step (ii) follows by using the inductive hypothesis $\|\theta_n^{S_\ell}\| \leq \varepsilon(n, \delta^*)^{\lambda_\ell}$ and applying Lemma 10 with $r = \varepsilon(n, \delta^*)^{\lambda_\ell}$. Note that in the final bound (50) the first term decreases with iteration t while the second term increases with t . In order to trade off these two terms, we set $t = T_{\ell+1}^{(1)}$ (40a) in the bound (50) and find that

$$\begin{aligned} \|\theta_n^{S_\ell + T_{\ell+1}^{(1)}}\| &\leq \frac{1}{(T_{\ell+1}^{(1)})^\beta} + c_2(2\varepsilon(n, \delta^*)^{\lambda_\ell})^\gamma \varepsilon(n, \delta^*) T_{\ell+1}^{(1)} \\ &= \underbrace{2(c_2 2^\gamma)^{\frac{\beta}{1+\beta}}}_{=: c'} \cdot \varepsilon(n, \delta^*)^{1 - \frac{\lambda_\ell \gamma + 1}{1+\beta} + \lambda_\ell \gamma} \\ &= c' \varepsilon(n, \delta^*)^{\frac{\lambda_\ell(\beta\gamma) + \beta}{1+\beta}} \\ &= c' \varepsilon(n, \delta^*)^{\lambda_{\ell+1}}, \end{aligned}$$

where the last equality follows from the relation (40c) between λ_ℓ and $\lambda_{\ell+1}$. The claim (48a) now follows.

Proof of claim (48b) For any $t \leq \tilde{\mathcal{T}}(c' \varepsilon(n, \delta^*)^{\lambda_{\ell+1}})$, we have

$$\begin{aligned} \|\theta_n^{t + S_\ell + T_{\ell+1}^{(1)}}\| &\leq \|F^t(\theta_n^{S_\ell + T_{\ell+1}^{(1)}})\| + \|F^t(\theta_n^{S_\ell + T_{\ell+1}^{(1)}}) - F_n^t(\theta_n^{S_\ell + T_{\ell+1}^{(1)}})\| \\ &\leq \frac{1}{t^\beta} + c_2(2c' \varepsilon(n, \delta^*)^{\lambda_{\ell+1}})^\gamma \varepsilon(n, \delta^*) t, \end{aligned}$$

where the last inequality follows from arguments similar to those used to establish the inequalities (49) and (50) above. Next, recalling the inequality $T_{\ell+1}^{(2)} \leq \tilde{\mathcal{T}}(c' \varepsilon(n, \delta^*)^{\lambda_{\ell+1}})$ from equation (45) and plugging $t = T_{\ell+1}^{(2)}$ (40a) in the above inequality, we find that

$$\|\theta_n^{S_{\ell+1}}\| \leq \underbrace{2(c_2 2^\gamma)^{\frac{\beta}{1+\beta}} c'^{\frac{\beta\gamma}{1+\beta}}}_{=: \tilde{C}} \cdot \varepsilon(n, \delta^*)^{\frac{\lambda_{\ell+1}\beta\gamma + \beta}{1+\beta}} = \tilde{C} \varepsilon(n, \delta^*)^{\lambda_{\ell+2}}.$$

In order to complete the proof, it remains to show that last quantity is upper bounded by $\varepsilon(n, \delta^*)^{\lambda_{\ell+1}}$; equivalently, we need to verify the following upper bound

$$\varepsilon(n, \delta^*) \leq \frac{1}{\tilde{C}^{\lambda_{\ell+2} - \lambda_{\ell+1}}}, \quad (51)$$

which is equivalent to the large sample-size assumption (C) (see condition (79) for a more precise statement) if we establish that

$$\lambda_{\ell+2} - \lambda_{\ell+1} \geq \alpha_\star := \frac{\alpha(1 + \beta - \beta\gamma)}{1 + \beta}. \quad (52)$$

In order to do so, we use the fact (40d) that $\lambda_\ell = \nu_\star(1 - \nu^\ell)$ and obtain that

$$\lambda_\ell \leq \nu_\star - \alpha \quad \text{and consequently that} \quad \nu_\star \nu^\ell \geq \alpha$$

. We ignore the effect of the ceiling function $\lceil \cdot \rceil$ to simplify the computations

for all $\ell \in \{0, 1, \dots, \ell_\alpha\}$. Putting together the pieces we have

$$\lambda_{\ell+2} - \lambda_{\ell+1} = \nu_\star \nu^{\ell+1} (1 - \nu) \geq \alpha(1 - \nu) = \alpha_\star,$$

which yields the claimed bound (52) and we are done.

A.1.4 PROOF OF CLAIM (46b)

The proof of this claim follows a similar road-map as that in the previous Section, and hence we simply sketch it. Conditional on the event \mathcal{E} , we claim that

$$\|\theta_n^{S_{\ell_\alpha} + kT_{\ell_\alpha}}\| \leq \varepsilon(n, \delta^\star)^{\nu_\star - \alpha} \quad \text{uniformly for all } k \in \{0, 1, 2, \dots\}. \quad (53)$$

Assuming this bound is given for now, we complete the proof. Invoking inequality (75) from the proof of Lemma 10, we obtain that

$$\|\theta_n^{S_{\ell_\alpha} + kT_{\ell_\alpha} + t}\| \leq 2\varepsilon(n, \delta^\star)^{\nu_\star - \alpha} \quad (54)$$

for all $k \in \{1, 2, \dots\}$ and $t \leq \tilde{\mathcal{T}}(\varepsilon(n, \delta^\star)^{\nu_\star - \alpha})$. Mimicking the arguments from claims (48a) and (48b), and using the large sample-size assumption (C) (condition (79)) yields the claim (54) for any $t \leq \varepsilon(n, \delta^\star)^{-\frac{\nu_\star}{\beta}}$. Putting this together with the fact (47) that $T_{\ell_\alpha} \leq \varepsilon(n, \delta^\star)^{-\frac{\nu_\star}{\beta}}$ implies the claim (46b).

Turning to the proof of claim (53), we note that the base case $k = 0$ follows from the claim (46a) by plugging in $\ell = \ell_\alpha$. For the inductive step, assuming $\|\theta_n^{S_{\ell_\alpha} + kT_{\ell_\alpha}}\| \leq \varepsilon(n, \delta^\star)^{\nu_\star - \alpha}$, arguments similar to that in the proof of claims (48a) and (48b) yield

$$\begin{aligned} \|\theta_n^{S_{\ell_\alpha} + kT_{\ell_\alpha} + T_{\ell_\alpha}^{(1)}}\| &\leq c' \varepsilon(n, \delta^\star)^{\nu_\star - \alpha} \quad \text{and,} \\ \|\underbrace{\theta_n^{S_{\ell_\alpha} + kT_{\ell_\alpha} + T_{\ell_\alpha}^{(1)} + T_{\ell_\alpha}^{(2)}}_{\theta_n^{S_{\ell_\alpha} + (k+1)T_{\ell_\alpha}}}\| &\leq \varepsilon(n, \delta^\star)^{\nu_\star - \alpha}, \end{aligned}$$

thereby establishing the induction hypothesis.

A.2 Proof of Theorem 6

We divide the proof into two subsections, corresponding to parts (a) and (b) of Theorem 6.

A.2.1 PROOF OF PART (A)

We introduce the shorthands $\tilde{\varepsilon}(n, \delta) = (\varepsilon(n, \delta))^{\frac{1}{1+\gamma}}$ and $T_f = \frac{1}{(1+\gamma)} \cdot \frac{\log(\rho/\varepsilon(n, \delta))}{\log(1/\kappa)}$. Without loss of generality, we can assume that

$$\|\theta_n^t - \theta^\star\| > \frac{(2 - \kappa)}{(1 - \kappa)} \tilde{\varepsilon}(n, \delta) \quad \text{for all } t \in \{0, \dots, T_f - 1\}. \quad (55)$$

Otherwise, the claim is immediate. Given the condition (55), we prove the following two claims:

$$\theta_n^t \in \mathbb{A}(\theta^\star, \tilde{\varepsilon}(n, \delta), \rho) \quad \text{for all } t \in \{0, \dots, T_f - 1\}, \quad (56a)$$

$$\text{and } \|\theta_n^{T_f} - \theta^\star\| \leq \frac{(2 - \kappa)}{(1 - \kappa)} \tilde{\varepsilon}(n, \delta). \quad (56b)$$

The latter claim (56b) completes the proof of part (a) of the theorem.

Proof of claim (56a) With the condition (55) in hand, it remains to prove that $\|\theta_n^t - \theta^*\| \leq \rho$. The base case of $t = 0$ is immediate from the initialization conditions. For the induction step, assuming $\theta_n^t \in \mathbb{A}(\theta^*, \tilde{\varepsilon}(n, \delta), \rho)$, we have

$$\begin{aligned}
\|\theta_n^{t+1} - \theta^*\| &= \|F_n(\theta_n^t) - \theta^*\| \leq \|F_n(\theta_n^t) - F(\theta_n^t)\| + \|F(\theta_n^t) - \theta^*\| \\
&\stackrel{(i)}{\leq} \sup_{\theta \in \mathbb{A}(\theta^*, \tilde{\varepsilon}(n, \delta), \rho)} \|F_n(\theta) - F(\theta)\| + \kappa \|\theta_n^t - \theta^*\| \\
&\stackrel{(ii)}{\leq} \varepsilon(n, \delta) \max \left\{ \frac{1}{\tilde{\varepsilon}(n, \delta)^\gamma}, \rho \right\} + \kappa \rho \\
&= \frac{\varepsilon(n, \delta)}{\tilde{\varepsilon}(n, \delta)^\gamma} + \kappa \rho \\
&= \varepsilon(n, \delta)^{\frac{1}{1+\gamma}} + \kappa \rho \stackrel{(iii)}{\leq} \rho,
\end{aligned} \tag{57}$$

where the inequality (i) follows from the induction hypothesis that $\theta_n^t \in \mathbb{A}(\theta^*, \tilde{\varepsilon}(n, \delta), \rho)$ and the fact that operator F is κ -contractive in the ball $\mathbb{B}(\theta^*, \rho)$; inequality (ii) follows from the first inequality from condition (19a) that implies that $\tilde{\varepsilon}(n, \delta) = \varepsilon(n, \delta)^{\frac{1}{1+\gamma}} \geq \tilde{\rho}$ and then invoking the instability condition (12) with $r = \tilde{\varepsilon}(n, \delta)$ and $\rho_2 = \rho$. Finally, the last inequality (iii) follows from the second bound of the condition (19a). The inductive step is thus established.

Proof of claim (56b) We observe that

$$\|\theta_n^{T_f} - \theta^*\| = \|F_n(\theta_n^{T_f-1}) - \theta^*\| \tag{58}$$

$$\begin{aligned}
&\leq \|F_n(\theta_n^{T_f-1}) - F(\theta_n^{T_f-1})\| + \|F(\theta_n^{T_f-1}) - \theta^*\| \\
&\stackrel{(i)}{\leq} \sup_{\theta \in \mathbb{A}(\theta^*, \tilde{\varepsilon}(n, \delta), \rho)} \|F_n(\theta) - F(\theta)\| + \kappa \|\theta_n^{T_f-1} - \theta^*\| \\
&\stackrel{(ii)}{\leq} \varepsilon(n, \delta) \max \left\{ \frac{1}{\tilde{\varepsilon}(n, \delta)^\gamma}, \rho \right\} + \kappa \|\theta_n^{T_f-1} - \theta^*\|,
\end{aligned} \tag{59}$$

where inequality (i) follows from our earlier claim (56a) and the κ -contractivity of the operator F on the ball $\mathbb{B}(\theta^*, \rho)$; inequality (ii) follows from an argument similar to the one used to establish the inequality (57). Finally, recursing equation (59) T_f times, we obtain that

$$\begin{aligned}
\|\theta_n^{T_f} - \theta^*\| &\leq \varepsilon(n, \delta) \max \left\{ \frac{1}{\tilde{\varepsilon}(n, \delta)^\gamma}, \rho \right\} \cdot (1 + \kappa + \dots + \kappa^{T_f-1}) + \kappa^{T_f} \|\theta_n^0 - \theta^*\| \\
&\leq \frac{\varepsilon(n, \delta)}{(1 - \kappa)} \max \left\{ \frac{1}{\tilde{\varepsilon}(n, \delta)^\gamma}, \rho \right\} + \kappa^{T_f} \rho \\
&\leq \frac{\tilde{\varepsilon}(n, \delta)}{(1 - \kappa)} + \tilde{\varepsilon}(n, \delta) = \frac{(2 - \kappa)}{(1 - \kappa)} \tilde{\varepsilon}(n, \delta),
\end{aligned}$$

where the last step follows from the upper bound on iteration T_f , which in turn implies that $\kappa^{T_f} \rho \leq \tilde{\varepsilon}(n, \delta)$. The proof is now complete.

A.2.2 PROOF OF PART (B)

The proof for Theorem 6(b) borrows ideas from the proof of Theorem 5 as well as the proof of part (a) of Theorem 6. We introduce the following definitions:

$$T_s := [\varepsilon(n, \delta)]^{-\frac{1-|\gamma|\nu_\star}{1+\beta}}, \quad \text{where} \quad \nu_\star := \frac{\beta}{1+\beta-\gamma\beta}.$$

In order to prove the result (20b), we can, without loss of generality, assume that

$$\|\theta_n^t - \theta^\star\| > 2[\varepsilon(n, \delta)]^{\nu_\star} \quad \text{for all} \quad t \in \{0, \dots, T_s - 1\}, \quad (60)$$

and show that $\|\theta_n^{T_s} - \theta^\star\| \leq 2[\varepsilon(n, \delta)]^{\nu_\star}$. We only prove the result for $\theta^\star = 0$ as the more general case can be derived in a similar fashion.

In order to proceed further, we make use of a result similar to Lemma 10 adapted to the unstable case. Given two positive scalars $r_1 < r_2$, we define

$$\tilde{\mathcal{T}}(r_1, r_2) := \frac{r_2 r_1^{|\gamma|}}{\varepsilon(n, \delta)}. \quad (61)$$

Lemma 11 *Suppose that the assumptions for part (b) of Theorem 6 hold. Further, suppose that the operator F_n satisfies $\|F_n^t(\theta)\| \geq r_1$ for any point θ such that $\|\theta\| \in [r_1, r_2]$ and for all $t \leq \tilde{\mathcal{T}}(r_1, r_2)$, where $\tilde{\rho} \leq r_1 \leq r_2 \leq \rho/2$. Then with probability at least $1 - \delta$, we have*

$$\sup_{\theta \in \mathbb{A}(\theta^\star, r_1, r_2)} \|F^t(\theta) - F_n^t(\theta)\| \leq t \cdot \frac{\varepsilon(n, \delta)}{r_1^{|\gamma|}} \quad \text{for all} \quad t \leq \tilde{\mathcal{T}}(r_1, r_2). \quad (62)$$

See Appendix C.2 for its proof.

We are now ready for the main argument. We have

$$\begin{aligned} \|\theta_n^t\| &= \|F_n^t(\theta_n^0)\| \leq \|F^t(\theta_n^0)\| + \|F^t(\theta_n^0) - F_n^t(\theta_n^0)\| \\ &\stackrel{(i)}{\leq} \frac{1}{t^\beta} + \|F^t(\theta_n^0) - F_n^t(\theta_n^0)\| \end{aligned} \quad (63)$$

$$\stackrel{(ii)}{\leq} \frac{1}{t^\beta} + t \cdot \frac{\varepsilon(n, \delta)}{[\varepsilon(n, \delta)]^{\nu_\star |\gamma|}}, \quad \text{for all} \quad t \leq \tilde{\mathcal{T}}([\varepsilon(n, \delta)]^{\nu_\star}, \rho), \quad (64)$$

with probability at least $1 - \delta$. Here, inequality (i) follows from the $\text{SLOW}(\beta)$ -convergence condition (8) of the operator F along with the assumptions that $\theta^\star = 0$ and $\|\theta_n^0\| \leq \rho$; inequality (ii) follows by applying Lemma 11 with $r_1 = [\varepsilon(n, \delta)]^{\nu_\star}$ and $r_2 = \rho$ in light of the condition (60). In the final bound (64), the first term decreases with iteration t while the second term increases with t . In order to trade off the two terms, we plug in $t = T_s \stackrel{(\dagger)}{\leq} \tilde{\mathcal{T}}([\varepsilon(n, \delta)]^{\nu_\star}, \rho)$ (where the inequality (\dagger) holds due to the second bound in assumption (20a)), and perform some algebra to obtain that

$$\|\theta_n^{T_s}\| \leq \frac{1}{T_s^\beta} + T_s \frac{\varepsilon(n, \delta)}{[\varepsilon(n, \delta)]^{\nu_\star |\gamma|}} \leq 2[\varepsilon(n, \delta)]^{\nu_\star},$$

which yields the claim.

Appendix B. Tightness of general results

In this appendix, we construct a simple class of problems to demonstrate that the guarantees Theorems 5 and 6 in this paper are unimprovable in general.

B.1 Constructing the family of operators

We establish our lower bounds by considering the following pairs of optimization problems:

$$\min_{\theta \in \mathbb{R}^d} f(\theta), \quad \text{where } f(\theta) := \frac{\|\theta\|^p}{p}, \quad \text{and} \quad (65)$$

$$\min_{\theta \in \mathbb{R}^d} f_n(\theta), \quad \text{where } f_n(\theta) := f(\theta) - \varepsilon_n \frac{\|\theta\|^q}{q}, \quad (66)$$

where p, q are positive reals satisfying $q \geq 2$, and $p > q+1$, and the scalar ε_n is a perturbation term. The perturbation ε_n is any non-increasing function in n , that decays to zero as the sample size n increases, so that the problem (66) can be seen as a noisy “finite-sample instantiation” of the “population-level” problem (65).

To study the tightness of our general results, we study the guarantees for three different algorithms: (a) gradient descent method, (b) Newton’s method, and (c) cubic-regularized Newton’s method (for $d = 1$). We note that for the population-level updates, there is a unique global optimal $\theta^* = 0$, and for the sample-level objective, the global minima θ_n^* satisfies $\epsilon_n^{\text{stat}} := \|\theta_n^*\| = \varepsilon_n^{\frac{1}{p-q}}$.

For our lower bounds, we analyze the behavior of three different algorithms: (a) gradient descent method, (b) Newton’s method, and (c) cubic-regularized Newton’s method (for $d = 1$), with the population-level operators Q_n^{GD} , Q_n^{NM} , and Q_n^{CNM} defined as follows:

$$Q^{\text{GD}}(\theta) = \theta - \eta \nabla f(\theta) = \theta (1 - \eta \|\theta\|^{q-2}), \quad (67a)$$

$$Q^{\text{NM}}(\theta) = \theta - [\nabla^2 f(\theta)]^{-1} \nabla f(\theta) = \left(1 - \frac{1}{p-1}\right) \theta, \quad \text{and} \quad (67b)$$

$$Q^{\text{CNM}}(\theta) = \arg \min_{y \in \mathbb{R}} \left\{ \nabla f(\theta)(y - \theta) + \frac{1}{2} \nabla^2 f(\theta)(y - \theta)^2 + c_p |y - \theta|^3 \right\}, \quad (67c)$$

where $c_p := \frac{1}{6}(p-1)(p-2)$, and $\eta > 0$ denotes the step-size of gradient descent algorithm. The corresponding sample-level updates are generated by the operators Q_n^{GD} , Q_n^{NM} , and Q_n^{CNM} , as follows:

$$Q_n^{\text{GD}}(\theta) = \theta - \eta \nabla f_n(\theta) = \theta - \eta (\|\theta\|^{p-2} - \varepsilon \|\theta\|^{q-2}) \theta, \quad (68a)$$

$$Q_n^{\text{NM}}(\theta) = \theta - [\nabla^2 f_n(\theta)]^{-1} \nabla f_n(\theta) = \frac{(p-2)\|\theta\|^{p-2} - (q-2)\varepsilon\|\theta\|^{q-2}}{(p-1)\|\theta\|^{p-2} - \varepsilon(q-1)\|\theta\|^{q-2}} \theta, \quad (68b)$$

$$Q_n^{\text{CNM}}(\theta) = \arg \min_{y \in \mathbb{R}} \left\{ \nabla f_n(\theta)(y - \theta) + \frac{1}{2} \nabla^2 f_n(\theta)(y - \theta)^2 + c_p |y - \theta|^3 \right\}. \quad (68c)$$

Standard algebra with the update equations (67a)-(67c) yields the following properties with the population-level operators:

($\hat{\text{P1}}$) the operator Q^{GD} is $\text{SLOW}(\frac{1}{p-2})$ -convergent on the ball $\mathbb{B}(\theta^*, 1)$ for small enough $\eta > 0$,

($\widehat{P}2$) the operator Q_n^{NM} is $\text{FAST}(\frac{p-2}{p-1})$ -convergent towards $\theta^* = 0$, and

($\widehat{P}3$) the operator Q_n^{CNM} is $\text{SLOW}(\frac{2}{p-3})$ -convergent on the ball $\mathbb{B}(\theta^*, 1)$.

Moving to the (in)-stability of sample-level operators, we can verify that:

($\widehat{S}1$) the operator Q_n^{GD} is $\text{STA}(q-1)$ -stable over the Euclidean ball $\mathbb{B}(\theta^*, 1)$;

($\widehat{S}2$) the operator Q_n^{NM} is $\text{UNS}(-p+q+1)$ -unstable over the annulus $\mathbb{A}(\theta^*, c_1 r_*, 1)$, and

($\widehat{S}3$) the operator Q_n^{CNM} is $\text{UNS}(-\frac{p+1}{2}+q)$ -unstable over the annulus $\mathbb{A}(\theta^*, c_2 r_*, 1)$ ($d=1$)

with respect to the corresponding population-level operators, and the noise function ε_n .

B.2 Lower bounds showing sharpness

In this section, we demonstrate the our general upper bounds on statistical accuracy and the iteration count, when specialized to the set-up above, are unimprovable. More precisely, the following result applies to the gradient descent updates (68a) with step size $\eta \in (0, \frac{1}{2}]$, along with the cubic regularized and standard Newton updates.

Proposition 12 *Let $p > q+1$ and $q \geq 2$, and define $\epsilon_n^{\text{stat}} := \varepsilon_n^{\frac{1}{p-q}}$, then for the set-up (66), given an initialization θ^0 with $\|\theta^0\| = 1$, we have*

$$\|(Q_n^{\text{GD}})^t(\theta^0) - \theta^*\| \begin{cases} \leq 2\epsilon_n^{\text{stat}} & \text{for all } t \geq c_1 \varepsilon_n^{-\frac{p-2}{p-q}}, \\ \geq \epsilon_n^{\text{stat}} & \text{for all } t \geq 1, \\ \geq 2\epsilon_n^{\text{stat}} & \text{for all } t \leq c'_1 \varepsilon_n^{-\frac{p-2}{p-q}}, \end{cases} \quad (69)$$

$$\|(Q_n^{\text{NM}})^t(\theta^0) - \theta^*\| \begin{cases} \leq 2\epsilon_n^{\text{stat}} & \text{for all } t \geq c_2 \log(\varepsilon_n^{-1}), \\ \geq \epsilon_n^{\text{stat}} & \text{for all } t \geq 1, \\ \geq 2\epsilon_n^{\text{stat}} & \text{for all } t \leq c'_2 \log(\varepsilon_n^{-1}), \end{cases} \quad \text{and} \quad (70)$$

$$\|(Q_n^{\text{CNM}})^t(\theta^0) - \theta^*\| \begin{cases} \leq 2\epsilon_n^{\text{stat}} & \text{for all } t \geq c_3 \varepsilon_n^{-\frac{p-3}{p-1}}, \\ \geq \epsilon_n^{\text{stat}} & \text{for all } t \geq 1, \end{cases} \quad (71)$$

where $\theta^* = 0$ denotes the fixed point of the operators Q^{GD} , Q^{NM} , and Q^{CNM} , and $c_1 > c'_1, c_2 > c'_2$, and c_3 denote universal constants depending on p, q and independent of n .

It is worth understanding how Proposition 12 establishes the tightness of the general upper bounds given in Theorems 5 and 6. Note that the properties ($\widehat{P}1$) – ($\widehat{P}3$) and ($\widehat{S}1$) – ($\widehat{S}3$), in conjunction with our general results in Theorems 5 and 6, provide an upper bound on the statistical error given sufficiently many iterations as summarized in bounds (69), (70) and (71), e.g., for the GD iterates, substituting $\beta = \frac{1}{p-2}$ and $\gamma = q-1$ in Theorem 5, we conclude that

$$\|(Q_n^{\text{GD}})^t(\theta^0) - \theta^*\| \leq 2\varepsilon_n^{\frac{1}{p-q}} = 2\epsilon_n^{\text{stat}} \quad \text{for all } t \geq c_1 \varepsilon_n^{-\frac{p-2}{p-q}}.$$

Furthermore, Proposition 12 guarantees that, up to a the constant pre-factor 2, this statistical error is the best possible since

$$\|(\mathbf{Q}_n^{\text{GD}})^t(\theta^0) - \theta^*\| \geq \epsilon_n^{\text{stat}} \quad \text{for all } t \geq 1.$$

Finally, the proposition also asserts that the GD updates take at least order $\epsilon_n^{-\frac{p-2}{q-2}}$ iterations to converge by the following additional bounds, as we also have the following bound from the display (69):

$$\|(\mathbf{Q}_n^{\text{GD}})^t(\theta^0) - \theta^*\| \geq 2\epsilon_n^{\text{stat}} \quad \text{for all } t = c'_2 \epsilon_n^{-\frac{p-2}{p-q}}.$$

A similar tightness of the statistical and computational guarantee can be argued for fast unstable methods stated in Theorem 6(a) via the guarantee (70) for the Newton's method. Finally, for slow unstable operators, we establish the tightness for the statistical error guarantee of Theorem 6(b) via the bound (71) for the CNM algorithm. (Showing the tightness of iteration complexity for this case requires fairly involved technical analysis, and is left for future work.) In a nutshell, Proposition 12 shows that the upper bound on the final statistical errors, and the lower bound on the number of iterations needed to obtain that final estimate, as stated in Theorems 5 and 6 are tight for the class of problems (66).

B.3 Proof of Proposition 12

As noted earlier, the upper bounds on the statistical error, and the corresponding lower bound on the number of iterations follow directly by substituting appropriate β and γ parameters from the properties listed above in Theorems 5 and 6. Since the arguments are very similar to those in Appendix D, we omit a detailed derivation.

In order to see that the statistical error cannot decrease any further, we note that in our example the iterates from gradient descent and (cubic-regularized) Newton's methods always converge to the global minima θ_n^* of f_n . Thus, we also have

$$\|\mathbf{Q}_n^{\text{GD}}(\theta)\| \geq \epsilon_n^{\text{stat}}, \quad \|\mathbf{Q}_n^{\text{NM}}(\theta)\| \geq \epsilon_n^{\text{stat}}, \quad \text{and} \quad |\mathbf{Q}_n^{\text{CNM}}(\theta)| \geq \epsilon_n^{\text{stat}}$$

for all $\|\theta\| \geq \epsilon_n^{\text{stat}}$. Consequently, we conclude that the error for all iterations are lower bounded by ϵ_n^{stat} .

Next, we establish the lower bounds on the number of iterations to converge to within $2\epsilon_n^{\text{stat}}$ for Gradient descent and Newton's method. Introducing the shorthand $\varepsilon := \varepsilon_n$, and rearranging terms in equations (68a) and (68b), we find that

$$\mathbf{Q}_n^{\text{GD}}(\theta) = (1 - \eta\|\theta\|^{p-2} + \eta\varepsilon\|\theta\|^{q-2})\theta, \tag{72}$$

$$\mathbf{Q}_n^{\text{NM}}(\theta) = \left(1 - \frac{\|\theta\|^{p-2} - \|\theta\|^{q-2}\varepsilon}{(p-1)\|\theta\|^{p-2} - (q-1)\|\theta\|^{q-2}\varepsilon}\right)\theta. \tag{73}$$

Proof for gradient descent iterates: Recursing the update (72), we find that

$$\theta^{t+T} = \theta^t \cdot \prod_{j=1}^T (1 - \eta\|\theta^{t+j}\|^{p-2} + \eta\varepsilon\|\theta^{t+j}\|^{q-2}). \tag{74}$$

Note that it suffices to show that with $\|\theta^t\| = 2\Delta := 4\epsilon_n^{\text{stat}} = 4\epsilon^{\frac{1}{p-q}}$, the smallest T_Δ such that $\|\theta^{T_\Delta+t}\| \leq \Delta$ satisfies $T_\Delta = \Omega(\Delta^{2-p})$.

Since the sequence $\{\|\theta^{t+j}\|\}_{j=1}^{T_\Delta}$ is a decreasing sequence, we find that

$$\Delta \leq \|\theta^{t+j}\| \leq 2\Delta \quad \text{for all } j = 1, \dots, T_\Delta.$$

Using $\Delta := 2 \cdot \epsilon^{\frac{1}{p-q}}$ and the update (74), we have

$$\|\theta^{t+T_\Delta+1}\| \geq (1 - c\eta\Delta^{p-2})^{T_\Delta} \|\theta^t\| = 2\Delta \cdot (1 - c\eta\Delta^2)^{T_\Delta}.$$

where $c = 2^{2p-4} - 2^{q-p} > 0$ under the assumptions $p > q + 1$ and $q \geq 2$. In order to ensure that $\|\theta^{t+T_\Delta}\| \leq \Delta$, we need to have

$$(1 - c\eta\Delta^{p-2})^{T_\Delta} \leq \frac{1}{2}.$$

Rearranging the last equation yields $T_\Delta \geq \frac{c'}{\Delta^{p-2}} \geq c'\epsilon^{-\frac{p-2}{p-q}}$, where c' is a universal constant which depends only on the pair (p, q) . This completes the proof.

Proof for Newton's method iterates: Following an argument similar to the last paragraph and using $\|\theta_t\| = 2\Delta \geq 2 \cdot \epsilon^{\frac{1}{p-q}}$, we find that

$$\|\theta^{t+T_\Delta}\| \geq \left(1 - \frac{1}{p-q}\right)^{T_\Delta} \|\theta^t\| = 2\Delta \cdot \left(1 - \frac{1}{p-q}\right)^{T_\Delta}.$$

Recalling that $p - q \geq 2$, we have that $T_\Delta \geq \frac{\log 2}{\log\left(\frac{p-q}{p-q-1}\right)}$. Consequently, in order to achieve an accuracy of $\frac{1}{\epsilon^{p-q}}$, we need at least $\frac{\log 2}{\log\left(\frac{p-q}{p-q-1}\right)} \cdot \log(\epsilon^{-(p-q)}) = c' \cdot \log(1/\epsilon)$ steps. Here, the universal constant c' only depends on (p, q) . This completes the proof of the sharpness of the Newton's method.

B.4 Undesirable behavior of unstable operators

In this appendix, we prove that the minimum over all iterates $k \in \{1, 2, \dots, t\}$ in Theorem 6 is necessary. In particular, we consider the following example

$$\mathcal{L}(\theta) = -\theta^4(\theta - 2)^2 \quad \text{and} \quad \mathcal{L}_n(\theta) = -\left(\theta^4 - \frac{\theta^2}{\sqrt{n}}\right)(\theta - 2)^2.$$

We let F and F_n denote the operators corresponding to the Newton's method as applied to the functions \mathcal{L} and \mathcal{L}_n , respectively (Consequently, the operator F has three fixed points). Following some simple algebra, it can be verified there are universal constants (c_1, c_2) such that the operators F and F_n defined above satisfy the conditions of Theorem 6 (a) with $\theta^* = 0$ for some $\kappa < 1$, $\gamma = -1$, $\epsilon(n, \delta) = n^{-\frac{1}{2}}$, $\tilde{\rho} = c_1 n^{-\frac{1}{4}}$ and $\rho = c_2$. In panel (a) of Figure 5, we plot the two functions \mathcal{L} and \mathcal{L}_n and illustrate the radii $\tilde{\rho}, \rho$ (for a fixed n). Some additional algebra shows that there exists $\theta_n^0 \in \mathbb{B}(\theta^*, \tilde{\rho})$ such that the iterates

corresponding to the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ satisfy $\|\theta_n^t - \theta^*\| \geq 1 \gg n^{-\frac{1}{4}}$ for all iterations $t = 1, 2, \dots$. See, in particular, the red (diamond) iterates in panel (b) of Figure 5 which are generated with a starting point $\theta_n^0 = c_3 n^{-\frac{1}{4}}$ (which is below the controlled instability threshold $\tilde{\rho}$). Clearly, we see that the first iterate produced by Newton's method escapes the local basin of attraction and the subsequent iterates converge to a very different fixed point of the function \mathcal{L}_n . On the other hand, when the Newton's method is initialized in the annulus $\mathbb{A}(\theta^*, \tilde{\rho}, \rho)$, the sequence θ_n^t (blue circles) converges quickly to the vicinity of θ^* as guaranteed by Theorem 6. Furthermore, the iterates do not escape this local neighborhood. Via this simple example, we have demonstrated that if no further regularity assumptions

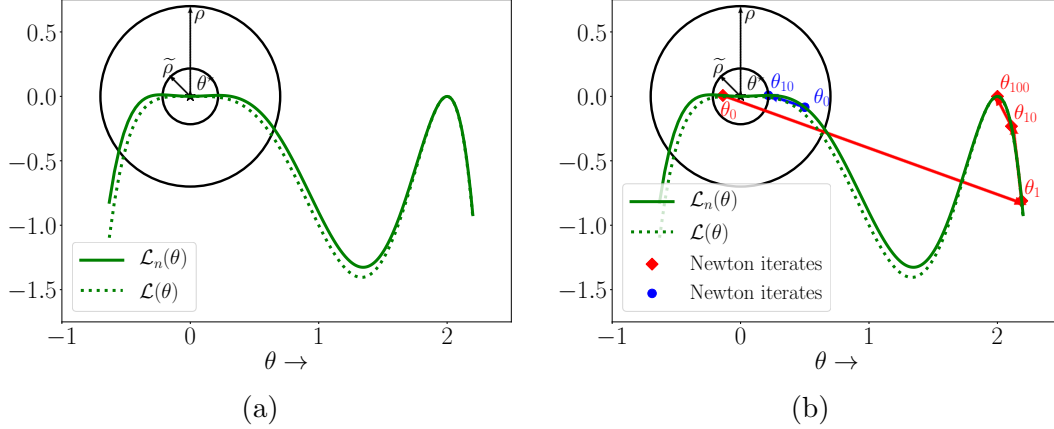


Figure 5. Instability of Newton's method for the example discussed above (figure best viewed in color). When the algorithm is initialized too close to θ^* (red diamonds), the instability of Newton's method forces the iterates to jump too far away from θ^* and converge to another fixed point. On the other hand, if the initial point is initialized in the annulus $\mathbb{A}(\theta^*, \tilde{\rho}, \rho)$, the Newton iterates (blue circles), do not leave this annulus and converge monotonically to a small neighborhood of θ^* .

are made, then starting an unstable algorithm from a point that is too close to θ^* , the subsequent iterates can be quite far from the true parameter.

Appendix C. Proofs of auxiliary results

In this appendix, we collect the proofs of Lemmas 10 and 11 that are central to the proofs of our main theorems.

C.1 Proof of Lemma 10

We fix a radius $r \in \mathcal{R}$. Our proof is based on the following auxiliary claim: conditioned on the event \mathcal{E} from equation (41), we have

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|F_n^t(\theta)\| \leq 2r \quad \text{for all } t \leq \tilde{\mathcal{T}}(r) = \frac{r^{1-\gamma}}{2^\gamma c_2 \varepsilon(n, \delta^*)}. \quad (75)$$

Taking this claim as given for the moment, we now establish the bound (44) claimed in the lemma. We do so via induction on the iteration $t \in \{0, 1, \dots, \tilde{\mathcal{T}}(r)\}$. Note that the

base-case $t = 0$ holds trivially, since $\|F^0(\theta) - F_n^0(\theta)\| = \|\theta - \theta\| = 0$. Given the induction hypothesis for t , we establish the claim for $t' = t + 1$. For any $\theta \in \mathbb{B}(\theta^*, r)$, we have

$$\begin{aligned}
 \|F^{t'}(\theta) - F_n^{t'}(\theta)\| &= \|F^{t+1}(\theta) - F_n^{t+1}(\theta)\| \\
 &\leq \|F(F^t(\theta)) - F(F_n^t(\theta))\| + \|F(F_n^t(\theta)) - F_n(F_n^t(\theta))\| \\
 &\stackrel{(i)}{\leq} \|F^t(\theta) - F_n^t(\theta)\| + \sup_{\tilde{\theta} \in \mathbb{B}(\theta^*, 2r)} \|F(\tilde{\theta}) - F_n(\tilde{\theta})\| \\
 &\stackrel{(ii)}{\leq} \sup_{\theta \in \mathbb{B}(\theta^*, r)} \|F^t(\theta) - F_n^t(\theta)\| + c_2(2r)^\gamma \varepsilon(n, \delta^*) \\
 &\stackrel{(iii)}{\leq} c_2(2r)^\gamma \varepsilon(n, \delta^*) t + c_2(2r)^\gamma \varepsilon(n, \delta^*) \\
 &= (t + 1) c_2(2r)^\gamma \varepsilon(n, \delta^*).
 \end{aligned} \tag{76}$$

In the above sequence of inequalities, we have made use of the following facts. In step (i), we have used the 1-Lipschitzness (6) of the operator F for the first term and the bound (75) on $F_n^t(\theta)$ for the second term. In order to establish step (ii), we have used the fact that $\theta \in \mathbb{B}(\theta^*, r)$ for the first term, while for the second term we have invoked the definition of the event \mathcal{E} in equation (41) with radius $2r$ (note that $2\mathcal{R} \subset \mathcal{R}'$ and the event \mathcal{E} is defined for all $r' \in \mathcal{R}'$). Finally step (iii) follows directly from the induction hypothesis. Noting that the bound (75) holds for any $t \leq \tilde{\mathcal{T}}(r)$ and taking supremum over $\theta \in \mathbb{B}(\theta^*, r)$ on the LHS of equation (81), we obtain the desired proof of the inductive step.

C.1.1 PROOF OF CLAIM (75)

We establish the claim (75) by proving the following stronger result: For any fixed $r \in \mathcal{R}$, and any $\theta \in \mathbb{B}(\theta^*, r)$, we have

$$\|F_n^t(\theta)\| \leq r + c_2(2r)^\gamma \varepsilon(n, \delta^*) \cdot t \quad \text{for all iterations } t = 0, 1, \dots, \tilde{\mathcal{T}}(r). \tag{77}$$

We note that the claim (75) is a direct application of this result along with the definition $\tilde{\mathcal{T}}(r) = \frac{r^{1-\gamma}}{2^\gamma c_2^\gamma \varepsilon(n, \delta^*)}$. We now use an induction argument on the iteration t (similar to the ones used in the paragraph above) to establish the claim (77). The base-case $t = 0$ holds trivially. Let us assume that $\|F_n^t(\theta)\| \leq r + c_2(2r)^\gamma \varepsilon(n, \delta^*) \cdot t$ and establish the claim (77) for $t' = t + 1$. Note that since $t \leq \tilde{\mathcal{T}}(r)$, this assumption trivially yields that $\|F_n^t(\theta)\| \leq 2r$. We have

$$\begin{aligned}
 \|F_n^{t+1}(\theta)\| &\leq \|F(F_n^t(\theta))\| + \|F(F_n^t(\theta)) - F_n(F_n^t(\theta))\| \\
 &\stackrel{(i)}{\leq} \|F_n^t(\theta)\| + \sup_{\tilde{\theta} \in \mathbb{B}(\theta^*, 2r)} \|F(\tilde{\theta}) - F_n(\tilde{\theta})\| \\
 &\stackrel{(ii)}{\leq} (r + c_2(2r)^\gamma \varepsilon(n, \delta^*) \cdot t) + c_2(2r)^\gamma \varepsilon(n, \delta^*) \\
 &= r + c_2(2r)^\gamma \varepsilon(n, \delta^*) (t + 1),
 \end{aligned}$$

where in step (i), we have used the 1-Lipschitzness (6) of the operator F for the first term and the observation that $\|F_n^t(\theta)\| \leq 2r$ for the second term. On the other hand, in step (ii),

we have used the induction hypothesis to bound the first term, and invoked the definition of the event \mathcal{E} in equation (41) with radius $2r$ to bound the second term. Taking supremum over $\theta \in \mathbb{B}(\theta^*, r)$ completes the proof.

C.1.2 PROOF OF CLAIM (45)

Combining the relation $\lambda_\ell = \nu_\star(1 - \nu^\ell)$ with the two inequalities in equation (45), we find that it suffices to prove the following two bounds:

$$\varepsilon(n, \delta^*)^{-\frac{\beta\nu^\ell}{1+\beta}} \geq (2^\gamma c_2)^{\frac{\beta}{1+\beta}} \quad \text{and} \quad \varepsilon(n, \delta^*)^{-\frac{\beta\nu^{\ell+1}}{1+\beta}} \geq (2^\gamma c_2)^{\frac{\beta}{1+\beta}} (c')^{-\frac{\beta}{\nu_\star(1+\beta)}}. \quad (78)$$

Observe that $\lambda_\ell \leq \nu_\star - \alpha/4$; consequently, we find that $1/\nu^\ell \leq 4\nu_\star/\alpha$ for all $\ell \leq \ell_\alpha$. Finally, invoking assumption (14) we find that

$$\varepsilon(n, \delta^*) \leq \frac{1}{(2^\gamma c_2)^{\frac{4\nu_\star}{\alpha}} \cdot \max\left\{1, (c')^{\frac{4}{\alpha}}\right\}}. \quad (79)$$

The rest of the proof follows by noting that the upper bound (79) implies the bounds in equation (78).

C.2 Proof of Lemma 11

Fix an arbitrary pair of radii $r_1, r_2 \in \mathcal{R}$. Our proof is based on the following intermediate claim

$$\|F_n^t(\theta)\| \leq 2r_2 \quad \text{for all } t \leq \tilde{\mathcal{T}}(r_1, r_2). \quad (80)$$

We prove this claim at the end of this appendix. Assuming that this claim is given at the moment, we now establish the bound (62) claimed in the lemma. We do so by using induction on the iteration $t \in \{0, 1, \dots, \tilde{\mathcal{T}}(r_1, r_2)\}$ where we note that the base-case $t = 0$ holds trivially, since $\|F^0(\theta) - F_n^0(\theta)\| = \|\theta - \theta\| = 0$. Turning to the induction step (with $t' = t + 1$), for any θ with $\|\theta\| \in [r_1, r_2]$, we have

$$\begin{aligned} \|F^{t'}(\theta) - F_n^{t'}(\theta)\| &= \|F^{t+1}(\theta) - F_n^{t+1}(\theta)\| \\ &\leq \|F(F^t(\theta)) - F(F_n^t(\theta))\| + \|F(F_n^t(\theta)) - F_n(F_n^t(\theta))\| \\ &\stackrel{(i)}{\leq} \|F^t(\theta) - F_n^t(\theta)\| + \sup_{r_1 \leq \|\tilde{\theta}\| \leq 2r_2} \|F(\tilde{\theta}) - F_n(\tilde{\theta})\| \\ &\stackrel{(ii)}{\leq} \sup_{r_1 \leq \|\theta\| \leq 2r_2} \|F^t(\theta) - F_n^t(\theta)\| + \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} \\ &\stackrel{(iii)}{\leq} t \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} = (t+1) \cdot \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}}. \end{aligned} \quad (81)$$

In step (i), we have used the 1-Lipschitzness (6) of the operator F for the first term and the upper bound (75) on $F_n^t(\theta)$ for the second term. In step (ii), the upper bound for the first term follows from the sequence of inequalities

$$\tilde{\rho} \leq r_1 \leq \|\theta\| \leq r_2 \leq 2r_2 \leq \rho,$$

whereas for the second term we have invoked the bound $\|\tilde{\theta}\| := F_n^{t'}(\theta) \leq 2r_2$ (80) and applied the instability condition (12). Finally, step (iii) follows from a direct application of the induction hypothesis. Note that the bound (75) holds for any $t \leq \tilde{\mathcal{T}}(r)$. By taking supremum over $\theta \in \mathbb{B}(\theta^*, r)$ on the LHS of equation (81), we obtain the desired proof of the inductive step.

C.2.1 PROOF OF BOUND (80)

We use an inductive argument to show that

$$\|F_n^t(\theta)\| \leq t \cdot \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + r_2 \quad \text{for all } 1 \leq t \leq \tilde{\mathcal{T}}(r_1, r_2), \quad (82)$$

which immediately implies the claim (80) once we plug in the definition of $\tilde{\mathcal{T}}$ (61).

For the base-case $t = 0$, invoking the properties of the operators F and F_n we have

$$\begin{aligned} \|F_n(\theta)\| &\leq \|F_n(\theta) - F(\theta)\| + \|F(\theta)\| \stackrel{(i)}{\leq} \sup_{r_1 \leq \|\theta\| \leq r_2} \|F_n(\theta) - F(\theta)\| + \|\theta\| \\ &\stackrel{(ii)}{\leq} \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + r_2, \end{aligned}$$

where step (i) follows since $\|\theta\| \in [r_1, r_2]$ and the operator F is 1-Lipschitz, and step (ii) follows from the instability condition (12). This proves the base case of the induction hypothesis (82).

Now we prove the inductive step. In particular, we assume that the induction hypothesis (82) holds for $t \leq \tilde{\mathcal{T}}(r_1, r_2) - 1$ and show that the upper bound (82) holds for $t' = t + 1$. Towards this end, unwrapping the expression for $\|F_n^{t+1}(\theta)\|$ we have

$$\begin{aligned} \|F_n^{t'}(\theta)\| &\leq \|F_n^{t+1}(\theta) - F(F_n^t(\theta))\| + \|F(F_n^t(\theta))\| \\ &\stackrel{(iii)}{\leq} \sup_{r_1 \leq \|\theta\| \leq 2r_2} \|F_n(\theta) - F(\theta)\| + \|F_n^t(\theta)\| \\ &\stackrel{(iv)}{\leq} \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + t \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + r_2 \\ &= (t + 1) \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + r_2. \end{aligned}$$

Here, step (iii) follows from the fact that $\|F_n^t(\theta)\| \geq r_1$ and the $\text{LL}(\rho)$ condition (6); step (iv) stems from the instability condition (12) and the induction hypothesis. This completes the proof of the intermediate claim (82).

Appendix D. Proofs of corollaries

We now collect the proofs of several corollaries stated in the paper. As a high-level summary, our analysis in all three examples in Section 4 involves applying Theorem 5 to analyze gradient descent/ascent and EM, both of which are stable algorithms and exhibit slow

convergence for the considered examples. We invoke Theorem 6(b) to characterize the cubic-regularized Newton algorithm, a slowly convergent and unstable algorithm. Finally, the analysis of Newton's method in all the examples relies on Theorem 6(a). Appendices D.1 and D.2 are devoted to the proofs of Corollaries 7 and 8, respectively. We then prove Corollary 9 in Appendix D.3. In this section, the values of universal constants (e.g., c , c' etc.) can change from line-to-line.

D.1 Proof of Corollary 7

In this appendix, we demonstrate the convergence and stability of the gradient and Newton methods. The operators for the gradient method and Newton's method take the following forms

$$M^{\text{GA}}(\theta) = \theta + \eta \bar{\mathcal{L}}'(\theta), \quad \text{and} \quad M_n^{\text{GA}}(\theta) = \theta + \eta \bar{\mathcal{L}}'_n(\theta), \quad (83a)$$

$$M^{\text{NM}}(\theta) = \theta - \left[\frac{\bar{\mathcal{L}}'(\theta)}{\bar{\mathcal{L}}''(\theta)} \right], \quad \text{and} \quad M_n^{\text{NM}}(\theta) = \theta - \left[\frac{\bar{\mathcal{L}}'_n(\theta)}{\bar{\mathcal{L}}''_n(\theta)} \right]. \quad (83b)$$

D.1.1 PROOFS FOR THE GRADIENT OPERATORS

In lieu of the discussion around Corollary 7 it remains to establish that (a) the operator M^{GA} exhibits a slow convergence condition $\text{SLOW}(\frac{1}{2})$ over the Euclidean ball $\mathbb{B}(\theta^*, 1/2)$ and (b) the operator M_n^{GA} satisfies a stability condition $\text{STA}(1)$ over the Euclidean ball $\mathbb{B}(\theta^*, 1/2)$ with noise function $\varepsilon(n, \delta) = \sqrt{\log(1/\delta)/n}$ when $n \geq c \log(1/\delta)$ for some universal constant $c > 0$.

Slow convergence of M^{GA} Direct computation with the gradient of population log-likelihood function $\bar{\mathcal{L}}$ leads to

$$\begin{aligned} \bar{\mathcal{L}}'(\theta) &:= \frac{\theta}{2(\theta^2 + 1)(2\sqrt{1 + \theta^2} - 1)} - \frac{\theta}{2} \\ \implies M^{\text{GA}}(\theta) &= \theta \left[1 - \eta \left(\frac{1}{2} - \frac{1}{2(\theta^2 + 1)(2\sqrt{1 + \theta^2} - 1)} \right) \right]. \end{aligned} \quad (84)$$

Noting that the fixed point of the population operator is $\theta^* = 0$ and that $\eta \leq 8/3$, we find that

$$\begin{aligned} |M^{\text{GA}}(\theta) - \theta^*| &= |\theta| \left[1 - \eta \left(\frac{1}{2} - \frac{1}{2(\theta^2 + 1)(2\sqrt{1 + \theta^2} - 1)} \right) \right] \\ &\leq |\theta| \left[1 - \eta \left(\frac{1}{2} - \frac{1}{2(\theta^2 + 1)} \right) \right] \\ &\leq |\theta| \left(1 - \frac{\eta\theta^2}{4} \right) \quad \text{for all } |\theta| \in [0, 1/2]. \end{aligned}$$

Thus the population operator M^{GA} satisfies a slow convergence condition $\text{SLOW}(\frac{1}{2})$ over the ball $\mathbb{B}(\theta^*, 1/2)$.

Stability of the sample operator M_n^{GA} We have

$$\begin{aligned} |M_n^{\text{GA}}(\theta) - M^{\text{GA}}(\theta)| &= \eta |\nabla \bar{\mathcal{L}}(\theta) - \nabla \bar{\mathcal{L}}_n(\theta)| \\ &\leq \eta \left(\left| \frac{\theta}{2(\theta^2 + 1)(2\sqrt{1 + \theta^2} - 1)} \left(\frac{2}{n} \sum_{i=1}^n (1 - R_i) - 1 \right) \right| \right. \\ &\quad \left. + \left| \theta \left(\frac{1}{2} - \frac{1}{n} \sum_{i=1}^n R_i Y_i^2 \right) \right| \right). \end{aligned}$$

Recall that, R_1, \dots, R_n are i.i.d. samples from Bernoulli distribution with probability $1/2$. Invoking Hoeffding's inequality yields that

$$\left| \frac{2}{n} \sum_{i=1}^n (1 - R_i) - 1 \right| \leq c \sqrt{\frac{\log(1/\delta)}{n}}, \quad (85)$$

with probability at least $1 - \delta$. Additionally, as Y_1, \dots, Y_n are i.i.d. samples from standard Gaussian distribution $\mathcal{N}(0, 1)$ and R_1, \dots, R_n are independent of Y_1, \dots, Y_n , by following the same argument as that in the proof of Lemma 1 from the paper (Dwivedi et al., 2020a), we can demonstrate that

$$\left| \frac{1}{n} \sum_{i=1}^n R_i Y_i^2 - \frac{1}{2} \right| \leq c_1 \sqrt{\frac{\log(1/\delta)}{n}}, \quad (86)$$

as long as the sample size $n \geq c_2 \log(1/\delta)$ with probability at least $1 - \delta$ where c_1 and c_2 are some universal constants.

Combining the inequalities (85) and (86) yields the following bound

$$\begin{aligned} &\sup_{\theta \in \mathbb{B}(\theta^*, r)} |M_n^{\text{GA}}(\theta) - M^{\text{GA}}(\theta)| \\ &\leq c_3 \sqrt{\frac{\log(1/\delta)}{n}} \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left(\frac{|\theta|}{2(\theta^2 + 1)(2\sqrt{1 + \theta^2} - 1)} + |\theta| \right) \\ &\leq \frac{3c_3 r}{2}, \end{aligned}$$

with probability at least $1 - 2\delta$ for any $r > 0$. Here, the second inequality in the above display follows from the fact that $(\theta^2 + 1)(2\sqrt{1 + \theta^2} - 1) \geq 1$ for all $\theta \in \mathbb{R}$. Thus, the sample-level operator M_n^{GA} is STA(1)-stable over the Euclidean ball $\mathbb{B}(\theta^*, 1/2)$ with noise function $\varepsilon(n, \delta) = \sqrt{\log(1/\delta)/n}$ when $n \geq c \log(1/\delta)$ for some universal constant $c > 0$.

D.1.2 PROOF FOR THE NEWTON OPERATORS

Similar to the proof for Newton operators in over-specified Gaussian mixtures (see Appendix D.2.1), we first verify the geometric convergence of population operator M^{NM} and the instability condition of sample operator M_n^{NM} . Then, we validate Assumption (D) by

showing that the Newton updates are monotone decreasing and satisfy the following lower bound

$$|M^{\text{NM}}(\theta)| \geq |\theta_n^*|, \quad (87)$$

for all $|\theta| \in [|\theta_n^*|, 1/2]$ for any global maxima θ_n^* of the sample log-likelihood function $\bar{\mathcal{L}}_n$ in equation (23).

Geometric convergence of M^{NM} We can verify that $\bar{\mathcal{L}}''(\theta) < 0$ for all $\theta \in \mathbb{R}$. Additionally, we have the following equation

$$|M^{\text{NM}}(\theta) - \theta^*| = |\theta - \theta^*| \frac{\theta^2 T_2(\theta)}{T_1(\theta) + \theta^2 T_2(\theta)},$$

where the functions T_1 and T_2 are defined as

$$T_1(\theta) := \frac{1}{2} - \frac{1}{2(\theta^2 + 1)(2\sqrt{\theta^2 + 1} - 1)}, \quad \text{and}$$

$$T_2(\theta) := \frac{1}{2(\theta^2 + 1)^2(2\sqrt{\theta^2 + 1} - 1)} \left(3 + \frac{1}{2\sqrt{\theta^2 + 1} - 1} \right).$$

From the earlier proof argument for slow convergence of M^{GA} , we have $T_1(\theta) \geq \frac{\theta^2}{8}$ for all $|\theta| \in [0, 1/2]$. Given the above lower bound of T_1 , we directly obtain that

$$|M^{\text{NM}}(\theta) - \theta^*| \leq |\theta - \theta^*| \frac{T_2(\theta)}{1/8 + T_2(\theta)} \leq |\theta - \theta^*| \frac{T_2(1/2)}{1/8 + T_2(1/2)} \leq \frac{4}{5} |\theta - \theta^*|,$$

for all $|\theta| \in [0, 1/2]$ where the last inequality is due to the fact that $T_2(\theta)/(c + T_2(\theta))$ achieves its maximum value at $|\theta| = 1/2$. Therefore, the population operator M^{NM} is FAST(4/5)-convergent on the ball $\mathbb{B}(\theta^*, 1/2)$.

Instability of the sample Newton operator M_n^{NM} Given the formulations of population operator M^{NM} and sample operator M_n^{NM} from Newton's method, we have the following inequality

$$|M_n^{\text{NM}}(\theta) - M^{\text{NM}}(\theta)| \leq \underbrace{\left| \frac{\bar{\mathcal{L}}'(\theta) - \bar{\mathcal{L}}'_n(\theta)}{\bar{\mathcal{L}}''(\theta)} \right|}_{:=J_1} + \underbrace{\left| \bar{\mathcal{L}}'_n(\theta) \left(\frac{1}{\bar{\mathcal{L}}''(\theta)} - \frac{1}{\bar{\mathcal{L}}''_n(\theta)} \right) \right|}_{:=J_2}.$$

We claim the following upper bounds of J_1 and J_2 :

$$J_1 \leq c_1 \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}}, \quad (89)$$

with probability at least $1 - 2\delta$ as long as $|\theta| \in [0, 1/2]$ and $n \geq c' \log(1/\delta)$, and

$$J_2 \leq c_2 \cdot \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}}, \quad (90)$$

with probability at least $1 - 6\delta$ when $|\theta| \geq \sqrt{2c}(\log(1/\delta)/n)^{1/4}$.

With the upper bounds (89) and (90) of J_1 and J_2 respectively, we arrive at the following inequality

$$|M_n^{\text{NM}}(\theta) - M^{\text{NM}}(\theta)| \leq c'' |\theta|^{-1} \sqrt{\log(1/\delta)/n},$$

with probability at least $1 - 8\delta$ as long as $\sqrt{2c}(\log(1/\delta)/n)^{1/4} \leq |\theta| \leq 1/2$. As a consequence, the sample operator M_n^{NM} satisfies instability condition $\text{UNS}(1)$ over the annulus $\mathbb{A}(\theta^*, \sqrt{2c}(\log(1/\delta)/n)^{1/4}, 1/2)$ with noise function $\varepsilon(n, \delta) = \sqrt{\frac{\log(1/\delta)}{n}}$ as long as $n \geq c' \log(1/\delta)$.

Proof for the upper bound of J_1 When $n \geq c' \log(1/\delta)$, we can validate that

$$|\bar{\mathcal{L}}'(\theta) - \bar{\mathcal{L}}'_n(\theta)| \leq c |\theta| \sqrt{\frac{\log(1/\delta)}{n}},$$

for any $|\theta| \in [0, 1/2]$ with probability at least $1 - 2\delta$ where c and c' are some universal constants. Furthermore, based on the computations in Appendix D.1.2, we find that

$$|\bar{\mathcal{L}}''(\theta)| = T_1(\theta) + \theta^2 T_2(\theta) \geq \frac{\theta^2}{8} + \theta^2 T_2(1/2) \geq \frac{11\theta^2}{32}, \quad (91)$$

for any $|\theta| \in [0, 1/2]$. Combining the previous inequalities, we have the following upper bound with J_1 :

$$J_1 \leq c_1 \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1 - 2\delta$ as long as $|\theta| \in [0, 1/2]$ and $n \geq c' \log(1/\delta)$.

Proof for the upper bound of J_2 In order to derive an upper bound for J_2 , we make use of the following bounds:

$$|\bar{\mathcal{L}}'_n(\theta)| \leq c_1 \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \right), \quad (92a)$$

$$|\bar{\mathcal{L}}''_n(\theta) - \bar{\mathcal{L}}''(\theta)| \leq c_2 \sqrt{\frac{\log(1/\delta)}{n}}, \quad (92b)$$

$$|\bar{\mathcal{L}}''_n(\theta)| \geq c_3 \left(\theta^2 - c \cdot \sqrt{\frac{\log(1/\delta)}{n}} \right), \quad (92c)$$

for all $|\theta| \in [0, 1/2]$ with probability at least $1 - 2\delta$ when $n \geq c' \log(1/\delta)$. Here, c, c_1, c_2, c_3 in the above bounds are universal constants independent of δ .

Deferring the proofs of these claims to later, we now proceed to give an upper bound for J_2 based on the given bounds in the above display. In particular, from the formulation

of J_2 , we achieve that

$$\begin{aligned} J_2 &\leq \frac{32c_1c_2}{11c_3} \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \right) \frac{\sqrt{\frac{\log(1/\delta)}{n}}}{\theta^2 \left(\theta^2 - c\sqrt{\frac{\log(1/\delta)}{n}} \right)} \\ &\leq C \cdot \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}} \end{aligned}$$

with probability at least $1 - 6\delta$ when $|\theta| \geq \sqrt{2c} (\log(1/\delta)/n)^{1/4}$ where C is some universal constant. Here, the last inequality is due to $|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \leq |\theta|^3 \left(1 + \frac{1}{2c}\right)$ and $\theta^2 - c\sqrt{\frac{\log(1/\delta)}{n}} \geq |\theta|^2/2$ as long as we have $|\theta| \geq \sqrt{2c} (\log(1/\delta)/n)^{1/4}$.

Proof of claim (92a) Invoking triangle inequality, when $n \geq c' \log(1/\delta)$ we have

$$|\bar{\mathcal{L}}'_n(\theta)| \leq c|\theta| \left(\sqrt{\frac{\log(1/\delta)}{n}} + \frac{1}{2} - \frac{1}{2(\theta^2 + 1) \left(2\sqrt{\theta^2 + 1} - 1 \right)} \right),$$

with probability at least $1 - 2\delta$ for any $|\theta| \in [0, 1/2]$ where the inequality in the above display is due to the inequalities (85) and (86). Furthermore, we can validate that

$$\frac{1}{2} - \frac{1}{2(\theta^2 + 1) \left(2\sqrt{\theta^2 + 1} - 1 \right)} \leq \frac{3\theta^2}{2}$$

for any $|\theta| \in [0, 1/2]$. In light of the previous inequalities, we arrive at the following inequality

$$|\bar{\mathcal{L}}'_n(\theta)| \leq \frac{3c|\theta|}{2} \left(\sqrt{\frac{\log(1/\delta)}{n}} + \theta^2 \right),$$

with probability at least $1 - 2\delta$ for all $|\theta| \in [0, 1/2]$. As a consequence, we reach the conclusion of claim (92a).

Proof of claims (92b) and (92c) The proof of claim (92b) is a direct application of triangle inequality and the fact that $|\theta| \in [0, 1/2]$. In addition, we have

$$|\bar{\mathcal{L}}''_n(\theta)| \geq |\bar{\mathcal{L}}''(\theta)| - |\bar{\mathcal{L}}''_n(\theta) - \bar{\mathcal{L}}''(\theta)| \geq c' \left(\theta^2 - c\sqrt{\frac{\log(1/\delta)}{n}} \right),$$

with probability at least $1 - 2\delta$ for any $|\theta| \in [0, 1/2]$ where c, c' are universal constants independent of δ and the last inequality in the above display is due the results from equation (91) and claim (92b). As a consequence, we achieve the conclusion of claim (92c).

Lower bound and monotonicity of Newton updates Now, we proceed to verify the lower bound of Newton updates in claim (87). In order to ease the ensuing presentation, we denote $f(\theta) := \frac{1}{(\theta^2+1)(2\sqrt{\theta^2+1}-1)}$ for all θ . The global maxima θ_n^* of the sample log-likelihood function $\bar{\mathcal{L}}_n$ are the solutions of the following equation

$$\theta_n^* f(\theta_n^*) \left(\frac{1}{n} \sum_{i=1}^n (1 - R_i) \right) = \theta_n^* \left(\frac{1}{n} \sum_{i=1}^n R_i Y_i^2 \right).$$

The specific forms of θ_n^* depend on the values of R_i, Y_i for $i \in [n]$. In particular, when $\sum_{i=1}^n R_i Y_i^2 < \sum_{i=1}^n (1 - R_i)$, namely, the Hessian of sample likelihood function $\bar{\mathcal{L}}_n$ at 0 is positive, the function $\bar{\mathcal{L}}_n$ is bimodal and symmetric around 0. Additionally, θ_n^* are different from 0 and become the solution of the following equation

$$f(\theta_n^*) \left(\frac{1}{n} \sum_{i=1}^n (1 - R_i) \right) = \left(\frac{1}{n} \sum_{i=1}^n R_i Y_i^2 \right). \quad (93)$$

On the other hand, when $\sum_{i=1}^n R_i Y_i^2 > \sum_{i=1}^n (1 - R_i)$, the function $\bar{\mathcal{L}}_n$ is unimodal and symmetric around 0. Under this case, $\theta_n^* = 0$ is the unique global maximum.

Without loss of generality, we assume that $\theta > 0$ and the global maxima are solutions of equation (93). From the formulation of M_n^{NM} , the inequality $M_n^{\text{NM}}(\theta) > 0$ is equivalent to

$$\theta f'(\theta) + f(\theta) < f(\theta_n^*),$$

which holds for all $\theta \geq |\theta_n^*|$ since $f(\theta) < f(\theta_n^*)$ and $f'(\theta) < 0$ as $\theta \geq |\theta_n^*|$. Therefore, we have $M_n^{\text{NM}}(\theta) > 0$ for all $\theta \geq |\theta_n^*|$. Now, in order to demonstrate that $M_n^{\text{NM}}(\theta) \geq |\theta_n^*|$ for $\theta \geq |\theta_n^*|$, it is equivalent to

$$(|\theta_n^*| - \theta) \theta f'(\theta) + |\theta_n^*| (f(\theta) - f(\theta_n^*)) \geq 0. \quad (94)$$

Invoking mean value theorem, we can find some constant $\bar{\theta} \in (|\theta_n^*|, \theta)$ such that

$$f(\theta) - f(\theta_n^*) = f(\theta) - f(|\theta_n^*|) = f'(\bar{\theta})(\theta - |\theta_n^*|).$$

Given the above equation, the inequality (94) can be rewritten as

$$|\theta_n^*| f'(\bar{\theta}) \geq \theta f'(\theta) \quad (95)$$

for all $\theta \geq |\theta_n^*|$. Since the function $\theta f'(\theta)$ is a decreasing function in $(0, 1/2]$, we have $\theta f'(\theta) \leq \bar{\theta} f'(\bar{\theta})$ for any $\bar{\theta} < \theta$. Since $f'(\bar{\theta}) < 0$ and $\bar{\theta} > |\theta_n^*|$, we find that $\bar{\theta} f'(\bar{\theta}) \leq |\theta_n^*| f'(\bar{\theta})$. In light of these two inequalities, we achieve the inequality (95). As a consequence, we reach the conclusion of claim (87).

D.2 Proof of Corollary 8

Under the model (28b), the sample EM operator takes the form

$$G_n^{\text{EM}}(\theta) = \frac{1}{n} \sum_{i=1}^n X_i \tanh(\theta X_i),$$

where $\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$ is the hyperbolic tangent. In our prior work (cf. Theorem 3 in the paper (Dwivedi et al., 2020a)), we studied the sample EM operator for this model.

Accordingly, in this paper, we limit our analysis to the Newton updates; see Appendix D.2.1 for the details. The sample and population Newton updates take the form

$$G^{\text{NM}}(\theta) = \theta - [\mathcal{L}''(\theta)]^{-1} \mathcal{L}'(\theta) = \theta + \frac{\mathbb{E}[X \tanh(X\theta)] - \theta}{\mathbb{E}[X^2 \tanh^2(X\theta)]}, \quad \text{and} \quad (96a)$$

$$\begin{aligned} G_n^{\text{NM}}(\theta) &= \theta - [\mathcal{L}_n''(\theta)]^{-1} \mathcal{L}_n'(\theta) \\ &= \theta + \frac{(\frac{1}{n} \sum_{i=1}^n X_i \tanh(X_i \theta)) - \theta}{\frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i \theta) + 1 - \frac{1}{n} \sum_{i=1}^n X_i^2}. \end{aligned} \quad (96b)$$

D.2.1 PROOFS FOR NEWTON OPERATORS

We begin by verifying the fast convergence of the operator G^{NM} and then the instability of the operator G_n^{NM} with respect to G^{NM} in Theorem 6. Then, we demonstrate that the Newton updates satisfy Assumption (D). Noting that it can be done by establishing that the Newton updates are monotone decreasing and admit the following lower bound

$$|G_n^{\text{NM}}(\theta)| \geq |\theta_n^*| \quad (97)$$

for all $|\theta| \in [|\theta_n^*|, 1/3]$ for any global maximum θ_n^* of \mathcal{L}_n .

Fast convergence of the population-level operator G^{NM} We provide the full proof for the case $\theta \in (0, \frac{1}{3}]$; the proof for the case $\theta \in [-\frac{1}{3}, 0)$ is analogous. We make use of the following known bounds (Dwivedi et al., 2020b) on the hyperbolic function $x \mapsto x \tanh(x)$:

$$x^2 - \frac{x^4}{3} \leq x \tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15} \quad \text{for all } x \in \mathbb{R}. \quad (98)$$

Applying this bound, we find that

$$\begin{aligned} \mathbb{E}[X \tanh(X\theta)] &\leq \frac{1}{\theta} \mathbb{E}[(X\theta)^2 - (X\theta)^4/3 + 2(X\theta)^6/15] = \theta - \theta^3 + 2\theta^5, \quad \text{as well as} \\ \mathbb{E}[X^2 \tanh^2(X\theta)] &\leq \frac{1}{\theta^2} \mathbb{E}[(X\theta)^4] = 3\theta^2, \end{aligned}$$

and consequently that

$$\frac{\theta - \mathbb{E}[X \tanh(X\theta)]}{\mathbb{E}[X^2 \tanh^2(X\theta)]} \geq \frac{\theta - (\theta - \theta^3 + 2\theta^5)}{3\theta^2} = \frac{\theta - 2\theta^3}{3} \stackrel{(\theta \in (0, \frac{1}{3}])}{\geq} \frac{2\theta}{9}.$$

Noting that $G^{\text{NM}}(\theta) = \theta - \frac{\theta - \mathbb{E}[X \tanh(X\theta)]}{\mathbb{E}[X^2 \tanh^2(X\theta)]}$ and $\theta^* = 0$, we conclude that the population Newton operator G^{NM} is $\text{FAST}(\frac{7}{9})$ -convergent over the ball $\mathbb{B}(\theta^*, \frac{1}{3})$.

Instability of the sample-level operator G_n^{NM} Let us introduce the shorthand

$$A_n := \frac{1}{n} \sum_{i=1}^n X_i \tanh(X_i \theta), \quad \text{and} \quad B_n := \frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i \theta) + 1 - \frac{1}{n} \sum_{i=1}^n X_i^2.$$

Using the definitions (96b) of the operators G_n^{NM} and G^{NM} , we find that

$$\begin{aligned} & |G_n^{\text{NM}}(\theta) - G^{\text{NM}}(\theta)| \\ &= \left| \frac{\mathbb{E}[X \tanh(X\theta)] - \theta}{\mathbb{E}[X^2 \tanh^2(X\theta)]} - \frac{A_n - \theta}{B_n} \right| \\ &\leq \underbrace{\frac{|\mathbb{E}[X \tanh(X\theta)] - A_n|}{\mathbb{E}[X^2 \tanh^2(X\theta)]}}_{:=J_1} + \underbrace{|A_n - \theta| \left| \frac{1}{\mathbb{E}[X^2 \tanh^2(X\theta)]} - \frac{1}{B_n} \right|}_{:=J_2}. \end{aligned} \tag{99}$$

Thus, in order to bound the difference $|G_n^{\text{NM}}(\theta) - G^{\text{NM}}(\theta)|$, it suffices to derive bounds for the terms J_1 and J_2 .

Upper bound for J_1 For a given $\delta \in (0, 1)$, as long as the sample size $n \geq C \log(1/\delta)$ for some universal constant C , we can apply Lemma 1 from the paper (Dwivedi et al., 2020a) to assert that

$$|\mathbb{E}[X \tanh(X\theta)] - A_n| \leq c |\theta| \sqrt{\frac{\log(1/\delta)}{n}} \quad \text{for all } |\theta| \in (0, \tfrac{1}{3}) \tag{100}$$

with probability $1 - \delta$. Moreover, the bound (98) implies that

$$\mathbb{E}[X^2 \tanh^2(X\theta)] \geq \frac{1}{\theta^2} \mathbb{E} \left[\left((X\theta)^2 - \frac{(X\theta)^4}{3} \right)^2 \right] = 3\theta^2 - 10\theta^4 + \frac{35\theta^6}{33} \geq 2\theta^2,$$

for $\theta \in [-\frac{1}{3}, \frac{1}{3}]$. Combining the above inequalities yields

$$J_1 = \frac{|\mathbb{E}[X \tanh(X\theta)] - A_n|}{\mathbb{E}[X^2 \tanh^2(X\theta)]} \leq c \frac{|\theta| \sqrt{\frac{\log(1/\delta)}{n}}}{2\theta^2} \leq c' \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}}, \tag{101}$$

for all $|\theta| \in (0, 1/3)$ with probability at least $1 - \delta$.

Upper bound for J_2 In order to obtain an upper bound for J_2 , we claim the following key bounds appearing in its formulation:

$$|A_n - \theta| \leq c_1 \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \right), \tag{102a}$$

$$|B_n| \geq c_2 \left(\theta^2 - c \frac{\log^4(3n/\delta)}{\sqrt{n}} \right), \tag{102b}$$

$$|\mathbb{E}[X^2 \tanh^2(X\theta)] - B_n| \leq c_3 \frac{\log(n/\delta)}{\sqrt{n}}, \tag{102c}$$

for all $|\theta| \in (0, 1/3]$ with probability at least $1 - 2\delta$ as long as the sample size $n \geq c \log(1/\delta)$. Here, c, c_1, c_2, c_3 in the above probability bounds are universal constants independent of δ . Assume that the above claims are given at the moment. The results in these claims lead to

$$\begin{aligned}
 J_2 &= |A_n| \left| \frac{\mathbb{E}[X^2 \tanh^2(X\theta)] - B_n}{B_n \mathbb{E}[X^2 \tanh^2(X\theta)]} \right| \\
 &\leq c' \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \right) \frac{\frac{\log(n/\delta)}{\sqrt{n}}}{\theta^2 \left(\theta^2 - c \frac{\log^4(3n/\delta)}{\sqrt{n}} \right)} \\
 &\leq c'' \frac{1}{|\theta|} \frac{\log(n/\delta)}{\sqrt{n}}
 \end{aligned} \tag{103}$$

with probability at least $1 - 5\delta$. Here, the last inequality is due to the facts that

$$|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \leq |\theta|^3 \left(1 + \frac{1}{2c} \right) \quad \text{and} \quad \theta^2 - c \frac{\log^4(3n/\delta)}{\sqrt{n}} \geq |\theta|^2 / 2,$$

as long as $|\theta| \geq \sqrt{2c} \log^2(3n/\delta) / n^{1/4}$. Plugging the bounds (101) and (103) into equation (99), the operator G_n^{NM} is $\text{UNS}(-1)$ -unstable over the annulus $\mathbb{A}(\theta^*, \frac{\sqrt{2c} \log^2(3n/\delta)}{n^{1/4}}, 1/3)$ with noise function $\varepsilon(n, \delta) = \frac{\log(n/\delta)}{\sqrt{n}}$ as long as the sample size $n \geq C \frac{\log^8(3n/\delta)}{n^{1/4}}$.

Proof of claim (102a) Invoking the concentration bound (100) and applying the triangle inequality, we find that

$$\begin{aligned}
 |A_n - \theta| &\leq \left| \frac{1}{n} \sum_{i=1}^n X_i \tanh(X_i \theta) - \mathbb{E}[X \tanh(X\theta)] \right| + |\mathbb{E}[X \tanh(X\theta)] - \theta| \\
 &\leq c \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + \frac{1}{|\theta|} |\mathbb{E}[X\theta \tanh(X\theta)] - \theta^2| \right)
 \end{aligned}$$

for all $|\theta| \in (0, 1/3]$ with probability $1 - \delta$. Next, taking expectation on both sides in the bounds (98), we find that

$$\begin{aligned}
 \mathbb{E}[X\theta \tanh(X\theta)] - \theta^2 &\leq \mathbb{E} \left[(X\theta)^2 - \frac{(X\theta)^4}{3} + \frac{2(X\theta)^6}{15} \right] - \theta^2 \\
 &= -\theta^4 + 2\theta^6 \leq -\frac{7\theta^4}{9}, \quad \text{and} \\
 \mathbb{E}[X\theta \tanh(X\theta)] - \theta^2 &\geq \mathbb{E} \left[(X\theta)^2 - \frac{(X\theta)^4}{3} \right] - \theta^2 = -\theta^4.
 \end{aligned}$$

Putting these pieces together yields the claim (102a).

Proof of claim (102b) Invoking standard chi-squared concentration bounds and applying triangle inequality, we obtain that

$$\begin{aligned} |B_n| &\geq \frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i \theta) - \left| \frac{1}{n} \sum_{i=1}^n X_i^2 - 1 \right| \\ &\geq c \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i \theta) - \sqrt{\frac{\log(1/\delta)}{n}} \right) \end{aligned}$$

with probability at least $1 - \delta$. Using the lower bound from inequality (98), we find that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i \theta) &\geq \frac{1}{n} \sum_{i=1}^n \left(\theta X_i^2 - \frac{\theta^3 X_i^4}{3} \right)^2 \\ &= \theta^2 \left(\frac{1}{n} \sum_{i=1}^n X_i^4 \right) - \frac{2\theta^4}{3} \left(\frac{1}{n} \sum_{i=1}^n X_i^6 \right) + \frac{\theta^6}{9} \left(\frac{1}{n} \sum_{i=1}^n X_i^8 \right) \\ &\stackrel{(i)}{\geq} \theta^2 \left(3 - c' \frac{\log^2(3n/\delta)}{\sqrt{n}} \right) - \frac{2\theta^4}{3} \left(15 + c' \frac{\log^3(3n/\delta)}{\sqrt{n}} \right) \\ &\quad + \frac{\theta^6}{9} \left(105 - c' \frac{\log^4(3n/\delta)}{\sqrt{n}} \right) \\ &\geq \theta^2 - c' \frac{\log^4(3n/\delta)}{\sqrt{n}}, \end{aligned}$$

with probability at least $1 - \delta$ for some universal constant c . Here step (i) makes use of the following concentration bound for higher moments of Gaussian random variables (Lemma 5 (Dwivedi et al., 2020b)):

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n X_i^{2k} - \mathbb{E} [X^{2k}] \right| \leq c' \frac{\log^k(3n/\delta)}{n^{\frac{1}{2}}} \right] \geq 1 - \frac{\delta}{3} \quad \text{for } k \in \{2, 4, 6\}$$

with probability at least $1 - \delta/3$ for $k \in \{2, 4, 6\}$. Putting together the pieces yields the claim (102b).

Proof of claim (102c) Applying the triangle inequality yields

$$\begin{aligned} &|\mathbb{E} [X^2 \tanh^2(X \theta)] - B_n| \tag{104} \\ &\leq \left| \frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i \theta) - \mathbb{E} [X^2 \tanh^2(X \theta)] \right| + \left| \frac{1}{n} \sum_{i=1}^n X_i^2 - 1 \right| \\ &\leq \left| \frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i \theta) - \mathbb{E} [X^2 \tanh^2(X \theta)] \right| + c \sqrt{\frac{\log(1/\delta)}{n}} \end{aligned}$$

with probability at least $1 - \delta$. By adapting the truncation argument from the proof of Lemma 5 in the paper (Dwivedi et al., 2020b) for the random variable $X \tanh(X)$ with $X \sim \mathcal{N}(0, 1)$, it follows that

$$\left| \frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i \theta) - \mathbb{E} [X^2 \tanh^2(X \theta)] \right| \leq c' \frac{\log(n/\delta)}{\sqrt{n}},$$

for all $|\theta| \in (0, 1/3]$ with probability at least $1 - \delta$. Putting the results together yields the claim (102c).

Lower bound and monotonicity of Newton updates We first make some observations about the structure of the log-likelihood function \mathcal{L}_n . Define

$$f(\theta) := \theta - \frac{1}{n} \sum_{i=1}^n X_i \tanh(X_i \theta).$$

When $\sum_{i=1}^n X_i^2 > n$, it can be shown (by computing the gradient and Hessian) that the log-likelihood \mathcal{L}_n is bimodal and symmetric around 0. It has multiple global maxima θ_n^* that are non-zero, and are solutions of the equation $f(\theta) = 0$. On the other hand, when $\sum_{i=1}^n X_i^2 \leq n$, the function \mathcal{L}_n is unimodal and symmetric around 0, and the point $\theta_n^* = 0$ is the unique global maximum of the log likelihood.

Next we verify the lower bound of Newton updates $G_n^{\text{NM}}(\theta)$ in claim (97); the proof of monotonicity can be argued similarly. Without loss of generality, we only consider the setting when the global maxima θ_n^* are different from 0 and $\theta > 0$. Under that case, the Hessian of the function \mathcal{L}_n at $|\theta_n^*|$ is negative. A direct computation with the gradient of the function f leads to

$$\begin{aligned} f'(\theta) &= 1 - \frac{1}{n} \sum_{i=1}^n X_i^2 \text{sech}^2(X_i \theta) = 1 - \frac{1}{n} \sum_{i=1}^n X_i^2 \text{sech}^2(|X_i| |\theta|) \\ &\geq 1 - \frac{1}{n} \sum_{i=1}^n X_i^2 \text{sech}^2(|X_i| |\theta_n^*|) \\ &= -\nabla^2 \mathcal{L}_n(\theta_n^*) > 0 \end{aligned}$$

for any $\theta > |\theta_n^*|$. Therefore, the function f is a strictly increasing function when $\theta > |\theta_n^*|$. It leads to the inequality $f(\theta) \geq f(\theta_n^*) = 0$ for all $\theta \geq |\theta_n^*|$. Further computation with second derivative of f yields that

$$f''(\theta) = \frac{2}{n} \sum_{i=1}^n X_i^3 \tanh(X_i \theta) \text{sech}^2(X_i \theta) > 0$$

for all $\theta > 0$. The above inequality is due to $X_i \tanh(X_i \theta) > 0$ for all $\theta > 0$ and $i \in [n]$. Thus, the function f' is strictly increasing when $\theta > 0$.

Now the inequality $G_n^{\text{NM}}(\theta) \geq |\theta_n^*|$ for all $\theta \geq |\theta_n^*|$ is equivalent to

$$f'(\theta)(\theta - |\theta_n^*|) \geq f(\theta) - f(\theta_n^*). \quad (105)$$

Invoking the mean value theorem, we find that

$$f(\theta) - f(\theta_n^*) = f(\theta) - f(|\theta_n^*|) = f'(\bar{\theta})(\theta - |\theta_n^*|)$$

for some $\bar{\theta} \in (|\theta_n^*|, \theta)$. Given that equality, the equality (105) can be rewritten as $f'(\theta) \geq f'(\bar{\theta})$ for all $\theta \geq |\theta_n^*|$. This inequality is true since f' is an increasing function when $\theta > 0$. As a consequence, we achieve the conclusion of claim (97).

D.3 Proof of Corollary 9

In this appendix, we demonstrate the convergence and stability properties of operators from gradient descent and (cubic-regularized) Newton's methods in the non-linear regression model. The sample operators of these methods take the following forms

$$F_n^{\text{GD}}(\theta) = \theta - \eta \tilde{\mathcal{L}}'_n(\theta) = \theta - \eta \left(\frac{2p}{n} \sum_{i=1}^n X_i^{4p} \theta^{4p-1} - \frac{2p}{n} \sum_{i=1}^n Y_i X_i^{2p} \theta^{2p-1} \right), \quad (106a)$$

$$\begin{aligned} F_n^{\text{NM}}(\theta) &= \theta - \left[\tilde{\mathcal{L}}''_n(\theta) \right]^{-1} \tilde{\mathcal{L}}'_n(\theta) \\ &= \theta - \frac{\left(\frac{1}{n} \sum_{i=1}^n X_i^{4p} \right) \theta^{2p+1} - \left(\frac{1}{n} \sum_{i=1}^n Y_i X_i^{2p} \right) \theta}{\left(\frac{4p-1}{n} \sum_{i=1}^n X_i^{4p} \right) \theta^{2p} - \frac{2p-1}{n} \sum_{i=1}^n Y_i X_i^{2p}}, \quad \text{and} \end{aligned} \quad (106b)$$

$$F_n^{\text{CNM}}(\theta) = \arg \min_{y \in \mathbb{R}} \left\{ \tilde{\mathcal{L}}'_n(\theta)(y - \theta) + \frac{1}{2} \tilde{\mathcal{L}}''_n(\theta)(y - \theta)^2 + L |y - \theta|^3 \right\}, \quad (106c)$$

where $L := (4p-1)!!(4p-1)p/3$. Noting that the specific choice of L in the formulation of the cubic-regularized Newton operator F_n^{CNM} arises because the second-order derivative of $\tilde{\mathcal{L}}_n$ is Lipschitz continuous with constant L . Similarly, the population-level operators are given by

$$F^{\text{GD}}(\theta) = \theta - \eta \tilde{\mathcal{L}}'(\theta) = \theta \left[1 - (4p-1)!!(2p)\eta\theta^{4p-2} \right], \quad (107a)$$

$$F^{\text{NM}}(\theta) = \theta - \left[\tilde{\mathcal{L}}''(\theta) \right]^{-1} \tilde{\mathcal{L}}'(\theta) = \frac{(4p-2)}{4p-1} \theta, \quad \text{and} \quad (107b)$$

$$F^{\text{CNM}}(\theta) = \arg \min_{y \in \mathbb{R}} \left\{ \tilde{\mathcal{L}}'(\theta)(y - \theta) + \frac{1}{2} \tilde{\mathcal{L}}''(\theta)(y - \theta)^2 + L |y - \theta|^3 \right\}. \quad (107c)$$

D.3.1 PROOFS FOR THE GRADIENT DESCENT OPERATORS

In order to achieve the conclusion of the corollary with convergence rate of updates from gradient descent method, it is sufficient to demonstrate that the sample gradient operator F_n^{GD} is $\text{STA}(2p-1)$ -stable over the Euclidean ball $\mathbb{B}(\theta^*, 1)$ with noise function $\varepsilon(n, \delta) = \frac{\log^{2p}(n/\delta)}{\sqrt{n}}$. By using the similar truncation argument as that in equation (104), we can verify the following concentration bound

$$\left| \frac{1}{n} \sum_{i=1}^n Y_i X_i^{2p} \right| \leq c \log^{2p}(n/\delta) / \sqrt{n}, \quad (108)$$

with probability $1 - \delta$ where c is some universal constant. An application of triangle inequality yields

$$|F^{\text{GA}}(\theta) - F_n^{\text{GA}}(\theta)| \leq \left| \frac{1}{n} \sum_{i=1}^n X_i^{4p} - (4p-1)!! \right| |\theta|^{4p-1} + c \frac{\log^{2p}(n/\delta)}{\sqrt{n}} |\theta|^{2p-1}. \quad (109)$$

Based on known concentration bounds for moments of Gaussian random variables (cf. Lemma 5 in (Dwivedi et al., 2020b)), we have

$$\left| \frac{1}{n} \sum_{i=1}^n X_i^{4p} - (4p-1)!! \right| \leq c' \log^{2p}(n/\delta) / \sqrt{n} \quad (110)$$

with probability $1 - \delta$ where c' is some universal constant. Substituting the inequality (110) into equation (109) yields the above claim with the stability of F_n^{GD} .

D.3.2 PROOFS FOR THE NEWTON OPERATORS

Moving to the convergence rates of updates from Newton's method, it is sufficient to establish the instability of F_n^{NM} with respect to F^{NM} , and moreover that, for any global minimum θ_n^* of the sample least-squares function $\tilde{\mathcal{L}}_n$ in equation (34b), we have

$$|F_n^{\text{NM}}(\theta)| \geq |\theta_n^*|, \quad (111)$$

for all $|\theta| \in [|\theta_n^*|, 1]$.

Instability of the sample Newton operator F_n^{NM} Let us introduce the following shorthand notation:

$$\begin{aligned} A_n &:= \left(\frac{2p}{n} \sum_{i=1}^n X_i^{4p} \right) \theta^{4p-1} - \left(\frac{2p}{n} \sum_{i=1}^n Y_i X_i^{2p} \right) \theta^{2p-1}, \\ B_n &:= \left(\frac{2p(4p-1)}{n} \sum_{i=1}^n X_i^{4p} \right) \theta^{4p-2} - \left(\frac{2p(2p-1)}{n} \sum_{i=1}^n Y_i X_i^{2p} \right) \theta^{2p-2}. \end{aligned}$$

Applying the triangle inequality yields

$$\begin{aligned} |F_n^{\text{NM}}(\theta) - F^{\text{NM}}(\theta)| &\leq \underbrace{\frac{|(4p-1)!!(2p)\theta^{4p-1} - A_n|}{(4p-1)!!(2p)(4p-1)\theta^{4p-2}}}_{:=J_1} \\ &\quad + \underbrace{|A_n| \left| \frac{1}{(4p-1)!!(2p)(4p-1)\theta^{4p-2}} - \frac{1}{B_n} \right|}_{:=J_2}. \end{aligned}$$

Upper bound for J_1 Invoking triangle inequality, we obtain that

$$\begin{aligned} &|A_n - (4p-1)!!(2p)\theta^{4p-1}| \\ &\leq 2p \left| \frac{1}{n} \sum_{i=1}^n X_i^{4p} - (4p-1)!! \right| |\theta|^{4p-1} + \left| \frac{2p}{n} \sum_{i=1}^n Y_i X_i^{2p} \right| |\theta|^{2p-1} \\ &\leq c \frac{\log^{2p}(n/\delta)}{\sqrt{n}} \left(|\theta|^{4p-1} + |\theta|^{2p-1} \right), \end{aligned}$$

where the last inequality is due to concentration bounds for moments of Gaussian random variables (108). With the above inequality, we have

$$J_1 \leq \frac{c \log^{2p}(n/\delta) \left(|\theta|^{4p-1} + |\theta|^{2p-1} \right)}{(4p-1)!!(2p)(4p-1)\sqrt{n}|\theta|^{4p-2}} \leq \frac{2c}{|\theta|^{2p-1}} \frac{\log^{2p}(n/\delta)}{\sqrt{n}}, \quad (112)$$

for all $|\theta| \leq 1$ with probability at least $1 - 2\delta$.

Upper bound for J_2 In order to obtain an upper bound for J_2 , we exploit the following concentration bounds

$$|A_n| \leq c_1 \left(|\theta|^{4p-1} + \frac{\log^{2p}(n/\delta)}{\sqrt{n}} |\theta|^{2p-1} \right), \quad (113a)$$

$$|B_n - (4p-1)!!(2p)(4p-1)\theta^{4p-2}| \leq c_2 \frac{\log^{2p}(n/\delta)}{\sqrt{n}}, \quad (113b)$$

$$|B_n| \geq c_3 \left((4p-1)!!(2p)(4p-1)\theta^{4p-2} - c \frac{\log^{2p}(n/\delta)}{\sqrt{n}} \right), \quad (113c)$$

for all $|\theta| \leq 1$ with probability at least $1 - 2\delta$. Here, c, c_1, c_2, c_3 are universal constants independent of δ . The proofs of the above claims are direct applications of triangle inequalities and concentration bounds we utilized earlier with gradient descent operators in Appendix D.3.1; therefore, they are omitted. In light of the above bounds, we can bound J_2 as follows:

$$\begin{aligned} J_2 &\leq \frac{c_1 c_2}{c_3} \left(|\theta|^{4p-1} + \frac{\log^{2p}(n/\delta)}{\sqrt{n}} |\theta|^{2p-1} \right) \\ &\quad \times \frac{\frac{\log^{2p}(n/\delta)}{\sqrt{n}}}{\theta^{4p-2} \left((4p-1)!!(2p)(4p-1)\theta^{4p-2} - c \frac{\log^{2p}(n/\delta)}{\sqrt{n}} \right)} \\ &\leq \frac{2c_1 c_2}{c_3 c} \frac{1}{|\theta|^{2p-1}} \frac{\log^{2p}(n/\delta)}{\sqrt{n}}, \end{aligned} \quad (114)$$

for all $|\theta| \in [C \cdot \log^{p/(2p-1)}(n/\delta)/n^{1/4(2p-1)}, 1]$ with probability $1 - 6\delta$ where C is solution of the equation $(4p-1)!!(2p)(4p-1)\theta^{4p-2} = 2c \frac{\log^{2p}(n/\delta)}{\sqrt{n}}$. Combining the results from equations (112) and (114), we achieve that

$$|F_n^{\text{NM}}(\theta) - F^{\text{NM}}(\theta)| \leq c' \frac{1}{|\theta|^{2p-1}} \frac{\log^{2p}(n/\delta)}{\sqrt{n}} \quad (115)$$

for all $|\theta| \in [C \log^{p/(2p-1)}(n/\delta)/n^{1/4(2p-1)}, 1]$ with probability $1 - 8\delta$ where c' is some universal constant.

As a consequence, the sample operator F_n^{NM} is $\text{UNS}(-2p+1)$ -unstable over the annulus $\mathbb{A}(\theta^*, c_1 \log^{p/(2p-1)}(n/\delta)/n^{1/4(2p-1)}, 1)$ with noise function $\varepsilon(n, \delta) = \frac{\log^{2p}(n/\delta)}{\sqrt{n}}$.

Lower bound and monotonicity of Newton updates Moving to the claim (111), we first study the global minima θ_n^* of the sample least-squares function $\tilde{\mathcal{L}}_n$ in equation (34b). In particular, they satisfy the equation $\nabla \tilde{\mathcal{L}}_n(\theta_n^*) = 0$, which is equivalent to

$$\left(\frac{1}{n} \sum_{i=1}^n X_i^{4p} \right) (\theta_n^*)^{4p-1} - \left(\frac{1}{n} \sum_{i=1}^n Y_i X_i^{2p} \right) (\theta_n^*)^{2p-1} = 0.$$

Given the above equation, the specific form of θ_n^* depends on the sign of second derivative of $\tilde{\mathcal{L}}_n$ at 0. In particular, when $\sum_{i=1}^n Y_i X_i^{2p} > 0$, the function $\tilde{\mathcal{L}}_n$ is bimodal and symmetric around 0. Additionally, global minima θ_n^* have the form

$$(\theta_n^*)^{2p} = \left(\frac{1}{n} \sum_{i=1}^n Y_i X_i^{2p} \right) / \left(\frac{1}{n} \sum_{i=1}^n X_i^{4p} \right). \quad (116)$$

On the other hand, when $\frac{1}{n} \sum_{i=1}^n Y_i X_i^{2p} \leq 0$, the function $\tilde{\mathcal{L}}_n$ is unimodal and symmetric around 0. Furthermore, it has only global minimum $\theta_n^* = 0$.

Now, we focus on the case $\theta > 0$ and $\sum_{i=1}^n Y_i X_i^{2p} > 0$, i.e., the global minima θ_n^* are different from 0 and the solutions of equation (116). A simple calculation demonstrates that $B_n > 0$ and $F_n^{\text{NM}}(\theta) > 0$ as long as $\theta > |\theta_n^*|$. Now, the inequality $F_n^{\text{NM}}(\theta) \geq |\theta_n^*|$ is equivalent to

$$\begin{aligned} & \left(\frac{4p-2}{n} \sum_{i=1}^n X_i^{4p} \right) \theta^{2p+1} + \left(\frac{2p-1}{n} \sum_{i=1}^n Y_i X_i^{2p} \right) |\theta_n^*| \\ & \geq \left(\frac{4p-1}{n} \sum_{i=1}^n X_i^{4p} \right) \theta^{2p} |\theta_n^*| + \left(\frac{2p-2}{n} \sum_{i=1}^n Y_i X_i^{2p} \right) \theta \end{aligned}$$

for $\theta \geq |\theta_n^*|$. In light of the closed form expression of $|\theta_n^*|$ in equation (116), a simple algebra with the above inequality leads to the inequality

$$(4p-2)\theta^{2p+1} + (2p-1)|\theta_n^*|^{2p+1} \geq (2p-2)(\theta_n^*)^{2p}\theta + (4p-1)|\theta_n^*|\theta^{2p},$$

which holds true due to AM-GM inequality. Thus, we have established the claim (111).

D.3.3 PROOFS FOR THE CUBIC-REGULARIZED NEWTON OPERATORS

Our proof is divided into three separate steps. First, we establish the slow convergence of operator F^{CNM} . Then, we proceed to establishing the instability of operator F_n^{CNM} . Finally, we demonstrate the monotonicity of cubic-regularized Newton updates and their lower bound

$$|F_n^{\text{CNM}}(\theta)| \geq |\theta_n^*|, \quad (117)$$

for all $|\theta| \in [|\theta_n^*|, 1]$ for any global minima θ_n^* of the sample least-squares function $\tilde{\mathcal{L}}_n$ in equation (34b).

Slow convergence of F^{CNM} Without loss of generality, we assume that $\theta \in (0, 1]$. Direct computation leads to

$$\begin{aligned} F^{\text{CNM}}(\theta) &= \theta + \theta^{4p-2} - \sqrt{\theta^{8p-4} + \frac{2}{4p-1}\theta^{4p-1}} \\ &= \theta - \frac{\frac{2}{4p-1}\theta^{4p-1}}{\theta^2 + \sqrt{\theta^{8p-4} + \frac{2}{4p-1}\theta^{4p-1}}} \leq \theta \left(1 - c_1\theta^{(4p-3)/2}\right), \end{aligned}$$

for any $\theta \in (0, 1]$ where $c_1 < 1$ is some universal constant. As a consequence, the operator F^{CNM} satisfies slow convergence condition $\text{SLOW}(2/(4p-3))$ over the Euclidean ball $\mathbb{B}(\theta^*, 1)$.

Instability of the sample operator F_n^{CNM} Suppose that $\theta > |\theta_n^*|$, where θ_n^* are global minima of the sample least-squares function $\tilde{\mathcal{L}}_n$. With this condition, direct computation of $F_n^{\text{CNM}}(\theta)$ leads to

$$F_n^{\text{CNM}}(\theta) = \theta - \frac{2\tilde{\mathcal{L}}'_n(\theta)}{\tilde{\mathcal{L}}''_n(\theta) + \sqrt{\left(\tilde{\mathcal{L}}''_n(\theta)\right)^2 + 12L \cdot \tilde{\mathcal{L}}'_n(\theta)}} := \theta - \frac{2\tilde{\mathcal{L}}'_n(\theta)}{T_n}.$$

Similar to the previous proofs with cubic-regularized Newton operators, we find that

$$|F^{\text{CNM}}(\theta) - F_n^{\text{CNM}}(\theta)| \leq 2 \frac{\tilde{\mathcal{L}}'(\theta) |T_n - T| + T |\tilde{\mathcal{L}}'_n(\theta) - \tilde{\mathcal{L}}'(\theta)|}{TT_n},$$

where $T := \tilde{\mathcal{L}}''(\theta) + \sqrt{\left(\tilde{\mathcal{L}}''(\theta)\right)^2 + 12L \cdot \tilde{\mathcal{L}}'(\theta)} \geq \sqrt{12L\tilde{\mathcal{L}}'(\theta)} \geq C \cdot \theta^{(4p-1)/2}$ for some universal constant $C > 0$. Additionally, we have

$$|T_n - T| \leq c' \cdot \theta^{-1/2} \frac{\log^{2p}(n/\delta)}{\sqrt{n}}$$

when $\theta \geq c \cdot \max \left\{ |\theta_n^*|, \frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}} \right\}$ with probability $1 - 10\delta$ for some universal constants c and c' . Furthermore, we can check that $T_n \geq \sqrt{12L \cdot \tilde{\mathcal{L}}'_n(\theta)} \geq c''\theta^{(4p-1)/2}$ as long as $\theta \geq c \cdot \max \left\{ |\theta_n^*|, \frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}} \right\}$ with probability $1 - 2\delta$ for some universal constant c'' . These inequalities guarantee that

$$|F^{\text{CNM}}(\theta) - F_n^{\text{CNM}}(\theta)| \leq c_1\theta^{-1/2} \frac{\log^{2p}(n/\delta)}{\sqrt{n}}$$

for all $\theta \geq c \cdot \max \left\{ |\theta_n^*|, \frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}} \right\}$ with probability $1 - 14\delta$. As a consequence, we conclude that the operator F_n^{CNM} is $\text{UNS}(-1/2)$ -unstable over the annulus $\mathbb{A}(\theta^*, c \frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}, 1)$ with noise function $\varepsilon = \frac{\log^{2p}(n/\delta)}{\sqrt{n}}$ where c is some universal constant.

Lower bound and monotonicity of cubic-regularized Newton updates To simplify the presentation, we only consider $\theta > 0$ and the setting when global minima θ_n^* are different from 0. As $\theta \geq |\theta_n^*|$, the inequality $F_n^{\text{CNM}}(\theta) \geq |\theta_n^*|$ is equivalent to

$$\tilde{\mathcal{L}}_n''(\theta) + \sqrt{\left(\tilde{\mathcal{L}}_n''(\theta)\right)^2 + 12L\tilde{\mathcal{L}}_n'(\theta)} > 2\tilde{\mathcal{L}}_n''(\tilde{\theta})$$

for some $\tilde{\theta} \in (|\theta_n^*|, \theta)$. This inequality holds since $\tilde{\mathcal{L}}_n'$ and $\tilde{\mathcal{L}}_n''$ are positive and strictly increasing when $\theta > |\theta_n^*|$, thereby completing the proof of claim (117).

Appendix E. Extension to multivariate settings

In this appendix, we discuss some extensions of the theoretical results in Section 4 to multivariate settings. Here we state detailed theoretical results for the EM algorithm and gradient descent for multivariate versions of the over-specified mixture model and the non-linear regression model. We explore the behavior of Newton's method via experimental studies in both Figures 4 and 6.

E.1 Over-specified mixture model

We denote by $\phi(\cdot; \theta, \sigma^2 I_d)$ the density of $\mathcal{N}(\theta, \sigma^2 I_d)$ random variable, i.e.,

$$\phi(x; \theta, \sigma^2 I_d) = (2\pi\sigma^2)^{-d/2} e^{-\frac{\|x-\theta\|^2}{2\sigma^2}}.$$

Assume that X_1, \dots, X_n be n i.i.d. samples from $\mathcal{N}(0, I_d)$. We then fit a two-component symmetric Gaussian mixture with equal fixed weights whose density is given by:

$$f_\theta(x) = \frac{1}{2}\phi(x; -\theta, I_d) + \frac{1}{2}\phi(x; \theta, I_d), \quad (118)$$

where $\theta \in \mathbb{R}^d$ is the parameter to be estimated. Given the model, the true parameter is unique and given by $\theta^* = 0$. Similar to the univariate setting in Section 4.2, we also use the EM algorithm to estimate $\theta^* = 0$. Direct calculation of the sample EM operator yields that

$$G_n^{\text{EM}}(\theta) = \frac{1}{n} \sum_{i=1}^n X_i \tanh(\theta^\top X_i),$$

where $\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$ for all $x \in \mathbb{R}$. The result characterizing the behavior of sample EM operator in the multivariate setting (118) is already proven in our prior work (Dwivedi et al., 2020a) (see Theorem 3 in that paper). So as to keep our discussion self-contained, we restate it here:

Corollary 13 *For the over-specified Gaussian mixture model (118) with $\theta^* = 0$, given some $\delta \in (0, 1)$ and for any fixed $\alpha \in (0, 1/4)$ and initialization $\theta^0 \in \mathbb{B}(\theta^*, 1)$, with probability at least $1 - \delta$ the sequence $\theta^t := (G_n^{\text{EM}})^t(\theta^0)$ of EM iterates satisfies the bound*

$$\|\theta^t - \theta^*\| \leq c_1 \left(\frac{d + \log(\frac{\log(1/\alpha)}{\delta})}{n} \right)^{\frac{1}{4} - \alpha} \quad \text{for all iterates } t \geq c'_1 \sqrt{\frac{n}{d}} \log \frac{1}{\alpha},$$

as long as $n \geq c''_1(d + \log \frac{\log(1/\alpha)}{\delta})$.

The result of Corollary 13 shows that the EM iterates converge to a radius of convergence $(d/n)^{1/4}$ around the true parameter $\theta^* = 0$ after $\sqrt{n/d}$ number of iterations. Note that our simulation results for EM, as shown in Figure 4, are consistent with this theoretical prediction.

E.2 Non-linear regression model

We now turn to the multivariate instantiation of the non-linear regression model considered in the main text. Suppose that we observe pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ generated from the model

$$Y_i = g(X_i^\top \theta^*) + \xi_i \quad \text{for} \quad i = 1, \dots, n, \quad (119)$$

where $\xi_i \sim \mathcal{N}(0, 1)$.

We assume that the covariate vectors X_i are drawn i.i.d. from the multivariate Gaussian $\mathcal{N}(0, I_d)$. As in our study of the univariate case, we consider the family of link functions $g(x) = x^{2p}$ for $p \geq 1$ and the unknown parameter $\theta^* = \mathbf{0}$. With this set-up, the maximum likelihood estimate for θ^* is based on the minimization problem

$$\min_{\theta \in \mathbb{R}^d} \tilde{\mathcal{L}}_n(\theta) \quad \text{where} \quad \tilde{\mathcal{L}}_n(\theta) := \frac{1}{2n} \sum_{i=1}^n \left(Y_i - (X_i^\top \theta)^{2p} \right)^2. \quad (120)$$

By taking the expectation of $\tilde{\mathcal{L}}_n$ with respect to $X_1, \dots, X_n \sim \mathcal{N}(0, I_d)$, we find that the corresponding population version of $\tilde{\mathcal{L}}$ takes the form

$$\tilde{\mathcal{L}}(\theta) := \frac{1}{2} \mathbb{E} \left[\left(Y - (X^\top \theta)^{2p} \right)^2 \right] = \frac{1 + (4p-1)!! \|\theta - \theta^*\|^{4p}}{2}. \quad (121)$$

The sample operator for the gradient method is given by

$$\begin{aligned} F_n^{\text{GD}}(\theta) &= \theta - \eta \nabla \tilde{\mathcal{L}}_n(\theta) \\ &= \theta - \eta \left(\frac{2p}{n} \sum_{i=1}^n X_i (X_i^\top \theta)^{4p-1} - \frac{2p}{n} \sum_{i=1}^n Y_i X_i (X_i^\top \theta)^{2p-1} \right), \end{aligned} \quad (122)$$

whereas the population level operator corresponding to the operator F_n^{GD} takes the form

$$F^{\text{GD}}(\theta) = \theta - \eta \nabla \tilde{\mathcal{L}}(\theta) = \theta \left(1 - (4p-1)!! 2p\eta \|\theta\|^{4p-2} \right), \quad (123)$$

where I_d denotes the identity matrix in d dimension.

We first state a result concerning the contraction and stability properties of the population and sample operators F^{GD} and F_n^{GD} .

Lemma 14 (a) For any step size $\eta \in (0, \frac{1}{(4p-1)!!(2p)}]$, the gradient operator F^{GD} is $\text{SLOW}(\frac{1}{4p-2})$ -convergent over the ball $\mathbb{B}(\theta^*, 1)$.

(b) The operator F_n^{GD} is $\text{STA}(2p-1)$ -stable over the ball $\mathbb{B}(\theta^*, 1)$ with noise function $\varepsilon(n, \delta) = \sqrt{\frac{d + \log(1/\delta)}{n}}$.

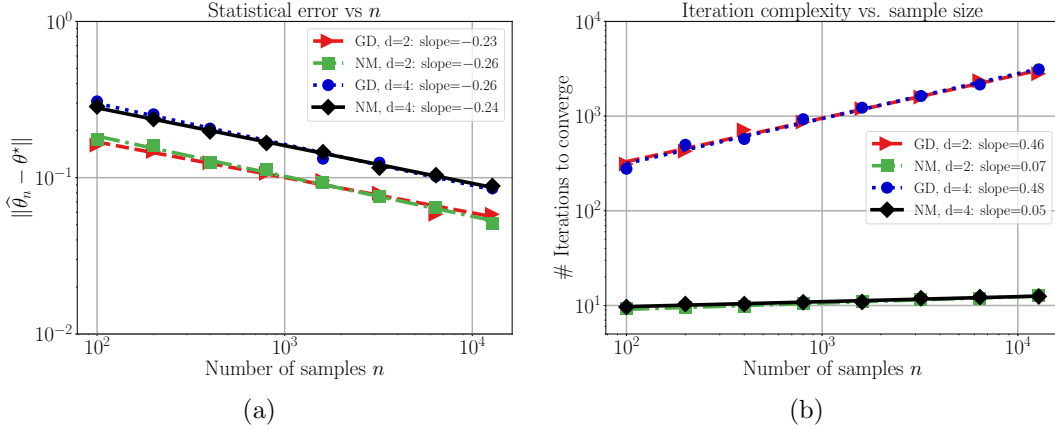


Figure 6. Plots characterizing the behavior of Gradient Descent (GD) and Newton's method (NM) for the non-linear regression with $p = 1$ for $d = 2$ and $d = 4$ dimensions. (a) Log-log plots of the Euclidean distance $\|\hat{\theta}_n - \theta^*\|_2$ versus the sample size. It shows that all the algorithms converge to an estimate at Euclidean distance of the order $n^{-1/4}$ from the true parameter θ^* . (b) Log-log plots for the number of iterations taken by different algorithms to converge to the final estimate.

The proof of Lemma 14 is deferred to the end of this appendix. Based on the result of that lemma, we have the following result characterizing the behavior of the updates from the gradient descent algorithm for solving $\tilde{\mathcal{L}}_n$.

Corollary 15 *For the non-linear regression model (119) with $\theta^* = 0$, given some $\delta \in (0, 1)$ and for any fixed $\alpha \in (0, 1/4)$ and initialization $\theta^0 \in \mathbb{B}(\theta^*, 1)$, with probability at least $1 - \delta$ the sequence $\theta^t := (F_n^{\text{GD}})^t(\theta^0)$ generated by gradient descent satisfies the bound*

$$\|\theta^t - \theta^*\| \leq c_1 \left(\frac{d + \log(\frac{\log(1/\alpha)}{\delta})}{n} \right)^{\frac{1}{4p} - \alpha} \quad \text{for all iterates } t \geq c'_1 \left(\frac{n}{d} \right)^{\frac{2p-1}{2p}} \log \frac{1}{\alpha},$$

as long as $n \geq c''_1 (d + \log \frac{\log(1/\alpha)}{\delta})^{4p}$.

Based on the result of Corollary 15, the updates from the gradient method converge to a ball of radius of the order of $(d/n)^{1/4p}$ around the true parameter $\theta^* = 0$ after an order of $(n/d)^{(2p-1)/2p}$ number of iterations. We further illustrate these behaviors of the gradient method when $p = 1$ in Figure 6. Based on these results, the computational complexity of the gradient method is at the order of $n^{\frac{4p-1}{2p}} d^{\frac{1}{2p}}$. For the Newton's method, the experimental results in Figure 6 show that the Newton iterates also converge to the similar radius of convergence $(d/n)^{1/4p}$ after $\log(n)$ number of iterations. Since each iteration of the Newton's method takes an order of $n \cdot d + d^3$ arithmetic operations where d^3 is computational complexity of computing inverse of an $d \times d$ matrix via Gauss-Jordan elimination, the overall complexity required to reach to the final estimate scales as $(nd + d^3) \log n$. Thus, when $d^{\frac{6p-1}{4p-1}} \ll n$, Newton's method is computationally more efficient than the gradient descent method.

E.2.1 PROOF OF LEMMA 14

The slow contraction of the population gradient operator F^{GD} follows immediately from its definition. Furthermore, the proof of the stability of the sample operator F_n^{GD} follows from the concentration bound in Corollary 3 in (Mou et al., 2019). In fact, from the proof of Corollary 3 in (Mou et al., 2019), as long as $r \leq 1$ we have

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} |F_n^{\text{GD}}(\theta) - F^{\text{GD}}(\theta)| \leq Cr^{2p-1} \sqrt{\frac{d + \log(1/\delta)}{n}},$$

as long as $n \geq C'(d + \log(d/\delta))^{4p}$ where C and C' are some universal constants. As a consequence, we obtain the conclusion of the lemma with the contraction and stability of the operators F^{GD} and F_n^{GD} .

References

- A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5):2452–2482, 2012.
- Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *2008 IEEE International Symposium on Information Theory*, pages 2454–2458. IEEE, 2008.
- S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017.
- Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.
- T. T. Cai, J. Ma, and L. Zhang. CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *Annals of Statistics*, To Appear.
- E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66:1241–1274, 2012.
- E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61:1985–2007, 2015.
- R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92:477–489, 1997.
- Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754, 2018.

- J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1): 221–233, 1995.
- Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. Technical report, UC Berkeley, September 2015. [arxiv:1509.03025.pdf](https://arxiv.org/abs/1509.03025).
- Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018a.
- Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, pages 1–33, 2018b.
- H. Chernoff. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16:31–41, 1964.
- C. Daskalakis, C. Tzamos, and M. Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–38, 1997.
- P. Diggle and M. G. Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43:49–93, 1994.
- R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *To appear, Annals of Statistics*, 2020a.
- R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Sharp analysis of expectation-maximization for weakly identifiable models. *AISTATS*, 2020b.
- Y. C. Eldar and S. Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36:473–494, 2013.
- R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 1987.
- Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, volume 9, pages 337–344, 2009.
- Elaine T Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for ℓ_1 -minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.
- Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <http://proceedings.mlr.press/v48/hardt16.html>.

- T. Hastie, R. Tibshirani, and M. J. Wainwright. *Statistical Learning with Sparsity: The Lasso and generalizations*. CRC Press, 2015.
- J. J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492, 1976.
- N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016.
- Joel Horowitz and Wolfgang Härdle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996.
- H. Ichimura. Semiparametric least squares (SLS) and weighted (SLS) estimation of single index models. *Journal of Econometrics*, 58:71–120, 1993.
- Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824, 2018.
- L. F. Lee. Asymptotic distribution of the maximum likelihood estimator for a stochastic frontier function model with a singular information matrix. *Econometric Theory*, 9:413–430, 1993.
- L. F. Lee and A. Chesher. Specification testing when score test statistic are identically zero. *Journal of Econometrics*, 31:121–149, 1986.
- Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.
- Z. Ma. Sparse principal component analysis and iterative thresholding. *Annals of Statistics*, 41(2):772–801, 2013.
- C. F. Manski. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3:205–228, 1975.
- Wenlong Mou, Nhat Ho, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. A diffusion process perspective on the posterior contraction rates for parameters. *arXiv preprint arXiv:1909.00966*, 2019.
- Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013.
- A. Rotnitzky, D. R. Cox, M. Bottai, and J. Robins. Likelihood-based inference with singular information matrix. *Bernoulli*, 6:243–284, 2000.

- J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:689–710, 2011.
- P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.
- Ronald G Shaiko, Diana Dwyre, Mark O’Gorman, Jeffrey M Stonecash, and James Vike. Pre-election political polling and the non-response bias issue. *International Journal of Public Opinion Research*, 3(1):86–99, 1991.
- Y. S. Tan and R. Vershynin. Phase retrieval via randomized Kaczmarz: Theoretical guarantees. *Information and Inference: A journal of the IMA*, 2018.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
- Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.
- Yihong Wu and Harrison H Zhou. Randomly initialized EM algorithm for two-component gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *arXiv preprint arXiv:1908.10935*, 2019.
- F. Yang, S. Balakrishnan, and M. Wainwright. Statistical and computational guarantees for the Baum-Welch algorithm. *Journal of Machine Learning Research*, 18:1–53, 2017.
- Xinyang Yi and Constantine Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.
- Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(Apr):899–925, 2013.
- C.-H Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.
- Huishuai Zhang, Yi Zhou, Yingbin Liang, and Yuejie Chi. A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *The Journal of Machine Learning Research*, 18(1):5164–5198, 2017.