Improving Computational Complexity in Statistical Models with Local Curvature Information

Pedram Akbarian [*1] **Tongzheng Ren** [*2] **Jiacheng Zhuo** [2] **Sujay Sanghavi** [1] **Nhat Ho** [3]

Abstract

It is known that when the statistical models are singular, i.e., the Fisher information matrix at the true parameter is degenerate, the fixed step-size gradient descent algorithm takes polynomial number of steps in terms of the sample size n to converge to a final statistical radius around the true parameter, which can be unsatisfactory for the practical application. To further improve that computational complexity, we consider utilizing the local curvature information for parameter estimation. Even though there is a rich literature in using the local curvature information for optimization, the statistical rate of these methods in statistical models, to the best of our knowledge, has not been studied rigorously. The major challenge of this problem is due to the non-convex nature of sample loss function. To shed light on these problems, we specifically study the normalized gradient descent (NormGD) algorithm, a variant of gradient descent algorithm whose step size is scaled by the maximum eigenvalue of the Hessian matrix of the empirical loss function, and deal with the aforementioned issue with a population-to-sample analysis. When the population loss function is homogeneous, the NormGD iterates reach a final statistical radius around the true parameter after a logarithmic number of iterations in terms of n. Therefore, for fixed dimension d, the NormGD algorithm achieves the optimal computational complexity $\mathcal{O}(n)$ to reach the final statistical radius, which is cheaper than the complexity $\mathcal{O}(n^\tau)$ of the fixed step-size gradient descent algorithm for some $\tau > 1$.

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of Texas at Austin, Texas, USA [2]Computer Science, University of Texas at Austin, Texas, USA [3]Department of Statistics and Data Sciences, University of Texas at Austin, Texas, USA. Correspondence to: Pedram Akbarian <akbarian@utexas.edu>, Tongzheng Ren <tongzheng@utexas.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Optimization serves a critical role in statistical models by enabling the determination of fixed points in data-dependent operators. This process is fundamental to finding parameter values that either maximize likelihood functions or minimize cost functions. The asymptotic performance of these models is linked to the properties of the population-level operator, particularly under the assumption of an infinitely large sample size. Consider the problem of determining the unique minimizer, denoted as $\theta^\star$, of an objective function $f : \mathbb{R}^d \to \mathbb{R}$. Crucially, we must emphasize that our understanding of the genuine objective function f is indirect and limited; but instead, we have access to an approximate (random) objective function f_n which serves as an unbiased estimate of the true objective function. To optimize the objective function, the gradient descent (GD) algorithm has been one of the most well-known and widely used (first-order) optimization methods for approximating the true parameter for parametric statistical models (Polyak, 1987; Bubeck, 2015; Nesterov, 2018).

Problem Setup. We are interested in solving the following optimization problem

$$\widehat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^d} f_n(\theta), \tag{1}$$

where f_n denote as the *sample* loss function, applied to a set of data points $\mathcal{S} = \{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\theta^\star}$ that are independently and identically distributed according to the probability distribution $\mathcal{P}_{\theta^\star}$. In this context, $\theta^\star$ represents the true, yet unknown, parameter of the distribution $\mathcal{P}_{\theta^\star}$, and $\widehat{\theta}_n$ is a solution of (1) which can be viewed as an estimation of the true parameter $\theta^\star$. Moreover, we define the corresponding *population* version of optimization problem (1) as follows:

$$\theta^\star \in \arg\min_{\theta \in \mathbb{R}^d} f(\theta), \tag{2}$$

where $f(\theta) := \mathbb{E}_{\mathcal{S} \sim \mathcal{P}_{\theta^\star}^n}[f_n(\theta)]$ is the *population* loss function, and the expectation is taken with respect to the data. It is worth noting that we can perceive any optimization algorithm as an iterative procedure, which can be reframed as a fixed point problem concerning its corresponding operator. Now, We denote F as the operator defining the iterates at the

population level, reflecting the idealized scenario of an infinite sample size. Similarly, F_n represents the operator at the sample level, based on a dataset of size n. The population-level iterates, defined as $\theta^t := F(\theta^{t-1})$ for $t \in \mathbb{N}$, starting with an initial value of θ^0. It is assumed that these iterates converge to its fixed point $\theta^\star$ as $t \to \infty$. As discussed earlier, since we do not have access to f, we can use the sample-based iterates, defined as $\theta_n^t := F_n^t(\theta^0)$ to obtain $\widehat{\theta}_n$ as an estimate of the true parameter $\theta^\star$.

Sub-optimality of fixed step size GD: An important insight here is that the statistical and computational complexities of fixed step size sample GD iterates $\theta_{n,\mathtt{GD}}^t$ are determined by the singularity of Hessian matrix of the population loss function f at $\theta^\star$. In particular, when the Hessian matrix of f at $\theta^\star$ is non-singular, i.e., $\nabla^2 f(\theta^\star) \succ 0$, the previous works (Balakrishnan et al., 2017; Ho et al., 2020) demonstrate that $\theta_{n,\mathtt{GD}}^t$ converge to a neighborhood of the true parameter $\theta^\star$ with the optimal statistical radius $\mathcal{O}((d/n)^{1/2})$ after $\mathcal{O}(\log(n/d))$ number of iterations. The logarithmic number of iterations is a direct consequence of the linear convergence of fixed step size GD algorithm for solving the strongly convex population loss function (2). When the Hessian matrix of f at $\theta^\star$ is singular, i.e., $\det(\nabla^2 f(\theta^\star)) = 0$, which we refer to as *singular statistical models*, $\theta_{n,\mathtt{GD}}^t$ can only converge to a neighborhood of $\theta^\star$ with the statistical radius larger than $\mathcal{O}((d/n)^{1/2})$ and the iteration complexity becomes polynomial in n. In particular, the work of (Ho et al., 2020) demonstrates that when the optimization rate of fixed step size population GD iterates follows the order of $1/t^{\frac{1}{\alpha'}}$ for some $\alpha' > 0$, and the noise magnitude of $\nabla f_n(\theta)$ is at the order of $\mathcal{O}(r^{\gamma'}(d/n)^{1/2})$ for some $\alpha' \geq \gamma'$ as long as $\|\theta - \theta^\star\| \leq r$, then the statistical rate of fixed step size sample GD iterates is $\mathcal{O}((d/n)^{\frac{1}{2(\alpha'+1-\gamma')}})$ after $\mathcal{O}((n/d)^{\frac{\alpha'}{2(\alpha'+1-\gamma')}})$ number of iterations. Given that the per iteration cost of fixed step size GD is $\mathcal{O}(nd)$, the total computational complexity of fixed step size GD for solving singular statistical models is $\mathcal{O}(n^{1+\frac{\alpha'}{2(\alpha'+1-\gamma')}})$ for a fixed dimension d, which is much more expensive than the optimal computational complexity $\mathcal{O}(n)$.

Contribution. To improve the computational complexity of the GD algorithm, we consider the utilization of the *local curvature information for parameter estimation in statistical models*. These methods go beyond simple gradient-based methods by incorporating second-order information. Newton's method is known for its fast convergence properties, making it suitable for finding precise solutions to optimization problems. However, it can be expensive to compute and store the Hessian matrix, particularly in high-dimensional settings. Indeed, each iteration of Newton's method demands a computational cost at the order of $\mathcal{O}(nd + d^3)$ [1] per iteration. On the other hand, BFGS is a quasi-Newton method that approximates the Hessian matrix, resulting in more computational efficiency. It requires a total of $\mathcal{O}(nd+d^2)$ arithmetic operations per iteration. Nonetheless, the question of whether both methods exhibit statistical optimality in high-dimensional settings has remained unclear. These considerations raise the following question:

Is there a method that achieves a balance between computational efficiency and provable statistical optimality at a reasonable per-iteration computational cost?

We explore this inquiry and demonstrate that *normalized gradient descent (NormGD) algorithm* can attain both statistical optimality and computational efficiency. NormGD is a variant of the gradient descent algorithm whose step size is scaled by the maximum eigenvalue of the Hessian matrix of the sample loss function. Calculating the complete spectrum of a $d \times d$ matrix typically demands $\mathcal{O}(d^3)$ computations. A more efficient alternative is to use the power iteration method, requiring just $\mathcal{O}(d^2)$ computations to find the maximum eigenvalue. Consequently, each iteration of NormGD necessitates $\mathcal{O}(nd + d^2)$ computations. We show that NormGD provably reaches the optimal statistical radius after a logarithmic number of iterations. Our results can be summarized as follows:

1. General theory: We study the computational and statistical complexities of NormGD iterates when the population loss function is homogeneous in all directions and the stability of first-order and second-order information holds. In particular, when the population loss function f is homogeneous with all fast directions, i.e., it is locally strongly convex and smooth, and the concentration bounds between the gradients and Hessian matrices of the sample and population loss functions are at the order of $\mathcal{O}((d/n)^{1/2})$, then the NormGD iterates reach the final statistical radius $\mathcal{O}((d/n)^{1/2})$ after $\log(n/d)$ number of iterations. When the function f is homogeneous, which corresponds to singular statistical models, with the fastest and slowest directions are at the order of $\|\theta - \theta^\star\|^\alpha$ for some $\alpha > 0$, and the concentration bound between Hessian matrices of the sample and population loss functions is $\mathcal{O}(r^\gamma (d/n)^{1/2})$ for some $\gamma \geq 0$ and $\alpha \geq \gamma + 1$, the NormGD iterates converge to a radius $\mathcal{O}((d/n)^{\frac{1}{2(\alpha-\gamma)}})$ within the true parameter after $\log(n/d)$ number of iterations. Therefore, for fixed dimension d the total computational complexity of NormGD to reach the final statistical radius is at the order of $\mathcal{O}((nd + d^2)\log(n/d))$, which is cheaper than that of the fixed step size GD, which is of the order of $\mathcal{O}(n^{1+\frac{\alpha}{2(\alpha-\gamma)}})$. Details of these results are in Theorem 2.3 and Proposition 2.4.

2. Examples: We illustrate the general theory for the sta-

[1] The cubic term, d^3, represents the computational complexity associated with calculating the inverse of a $d \times d$ matrix using the Gauss-Jordan elimination approach.

Table 1: Overview of Results for GLM with Link Function $g(r) = r^p$ in low SNR regime with $\theta^\star = 0$. The second and third columns represent the optimization properties of the respective algorithms. The fourth column presents the statistical performance of each algorithm, where NormGD attains the optimal rate while BFGS does not. The last column illustrates the computational complexity of each algorithm, showcasing that NormGD outperforms Gradient Descent, Newton's method, and BFGS for a fixed final statistical radius.

Algorithm	Optimization Rate	Iterations for convergence	Statistical error on convergence	Computational complexity
Gradient Descent	$t^{-1/(2p-2)}$	$(n/d)^{\frac{p-1}{p}}$	$(d/n)^{\frac{1}{2p}}$	$n^{\frac{2p-1}{p}} d^{\frac{1}{p}}$
Newton's Method (Ho et al., 2020)	$e^{-\kappa_{\text{NWT}} t}$	$\log(n/d)$	$(d/n)^{\frac{1}{2p}}$ *	$(nd + d^3)\log(n/d)$
BFGS (Jin et al., 2024)	$e^{-\kappa_{\text{BFGS}} t}$	$\log(n/d)$	$(d/n)^{\frac{1}{2p+2}}$ *	$(nd + d^2)\log(n/d)$
NormGD (Ours)	$e^{-\kappa_{\text{NGD}} t}$	$\log(n/d)$	$(d/n)^{\frac{1}{2p}}$	$(nd + d^2)\log(n/d)$

* Note that the statistical behavior of Newton's method has not been established for multivariate case. In the univariate case $d = 1$, both Newton's method and BFGS reach the optimal final statistical radius of $(d/n)^{\frac{1}{2p}}$.

tistical guarantee of NormGD under two popular statistical models: generalized linear models (GLM) and Gaussian mixture models (GMM). For GLM, we consider the settings when the link function $g(r) = r^p$ for $p \in \mathbb{N}$ and $p \geq 2$. Under the strong signal-to-noise (SNR) regime, namely, when the norm of the true parameter is sufficiently large, the NormGD iterates reach the statistical radius $\mathcal{O}((d/n)^{1/2})$ around the true parameter after $\log(n/d)$ number of iterations. On the other hand, for the low SNR regime, specifically, we assume $\theta^\star = 0$, the final statistical radius of NormGD updates is $\mathcal{O}((d/n)^{\frac{1}{2p}})$ and it is achieved after $\log(n/d)$ number of iterations. Furthermore, we compare NormGD to gradient descent, Newton's method, and BFGS in terms of computational efficiency and statistical optimality. Under low SNR regimes, the computational complexity of GD iterations scales as $\mathcal{O}(n^{\frac{2p-1}{2p}} d^{\frac{1}{p}})$. Additionally, the computational complexity for Newton's method is $\mathcal{O}((nd+d^3)\log(n/d))$, while both BFGS and NormGD exhibit a computational complexity of $\mathcal{O}((nd+d^2)\log(n/d))$ which are more efficient than GD and Newton's method. However, the statistical optimality of BFGS remains suboptimal (Jin et al., 2024). An overview of the results for the low SNR GLM is presented in Table 1. Moving to the GMM, we specifically consider the symmetric two-component location setting, which has been considered widely to study the statistical behaviors of Expectation-Maximization (EM) algorithm (Balakrishnan et al., 2017; Dwivedi et al., 2020b). We demonstrate that the statistical radius of NormGD iterates under strong and low SNR regimes are respectively $\mathcal{O}((d/n)^{1/2})$ and $\mathcal{O}((d/n)^{1/4})$. Both of these results are obtained after $\log(n/d)$ number of iterations.

Organization. The paper is organized as follows. In Section 2 and Appendix A, we provide a general theory for the statistical guarantee of the NormGD algorithm for solving parameter estimation in parametric statistical models when the population loss function is homogeneous. We illustrate the general theory with generalized linear models and mixture models in Section 3. We conclude the paper with a few discussions in Section 5. Finally, proofs of the general theory are in Appendix B while proofs of the examples are in the remaining appendices in the supplementary material.

Notation. For any $n \in \mathbb{N}$, we denote $[n] = \{1, 2, \ldots, n\}$. For any matrix $A \in \mathbb{R}^{d\times d}$, we denote $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ respectively the maximum and minimum eigenvalues of the matrix A. Throughout the paper, $\|\cdot\|$ denotes the ℓ_2 norm of some vector while $\|\cdot\|_{\text{op}}$ denotes the operator norm of some matrix. For any two sequences $\{a_n\}_{n\geq 1}, \{b_n\}_{n\geq 1}$, the notation $a_n = \mathcal{O}(b_n)$ is equivalent to $a_n \leq C b_n$ for all $n \geq 1$ where C is some universal constant.

2. General Theory of Normalized Gradient Descent

In this section, we provide statistical and computational complexities of NormGD updates for homogeneous settings when all the directions of the population loss function f have similar behaviors. For the inhomogeneous population loss function, to the best of our knowledge, the theories for these settings are only for specific statistical models (Dwivedi et al., 2020a; Zhuo et al., 2021). The general theory for these settings is challenging and hence we leave this direction for future work. To simplify the ensuing presentation, we denote the NormGD iterates for solving the samples and population losses functions (1) and (2) as follows $\theta_n^{t+1} := F_n^{\text{NGD}}(\theta_n^t)$ and $\theta^{t+1} := F^{\text{NGD}}(\theta^t)$, respectively, where sample and population NormGD operators are defined as follows

$$F_n^{\text{NGD}}(\theta_n^t) := \theta_n^t - \frac{\eta}{\lambda_{\max}(\nabla^2 f_n(\theta_n^t))}\nabla f_n(\theta_n^t),$$
$$F^{\text{NGD}}(\theta^t) := \theta^t - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta^t))}\nabla f(\theta^t).$$

Furthermore, we call $\{\theta_n^t\}_{t\geq 0}$ and $\{\theta^t\}_{t\geq 0}$ as the sample and population NormGD iterates respectively.

Locally strongly convex and smooth settings: For the homogeneous setting when all directions are fast, namely, when the population loss function is locally strongly convex, we defer the general theory of these settings to Appendix A.

Locally convex and smooth settings: Here, we focus on homogeneous settings where all directions are slow. To characterize the homogeneous settings, we assume that the population loss function f is locally convex in the ball $\mathbb{B}(\theta^\star, r) := \{\theta \in \mathbb{R}^d : \|\theta - \theta^\star\| \leq r\}$ for some given radius r. Apart from the local convexity assumption, we also utilize the following assumption on the population loss function f.

(W.1) **(Homogeneous Property)** Given the constant $\alpha > 0$ and the radius $r > 0$, for all $\theta \in \mathbb{B}(\theta^\star, r)$ we have

$$\lambda_{\min}(\nabla^2 f(\theta)) \geq c_1 \|\theta - \theta^\star\|^\alpha,$$
$$\lambda_{\max}(\nabla^2 f(\theta)) \leq c_2 \|\theta - \theta^\star\|^\alpha,$$

where $c_1 > 0$ and $c_2 > 0$ are some universal constants depending on r.

The condition $\alpha > 0$ is to ensure that the Hessian matrix is singular at the true parameter $\theta^\star$. For the setting $\alpha = 0$, corresponding to the locally strongly convex setting, the detailed analysis of NormGD is provided in Appendix A. A simple example of Assumption (W.1) is $f(\theta) = \|\theta - \theta^\star\|^{\alpha+2}$ for all $\theta \in \mathbb{B}(\theta^\star, r)$. The Assumption (W.1) is satisfied by several statistical models, such as low signal-to-noise regime of generalized linear models with polynomial link functions (see Section 3.1) and symmetric two-component mixture model when the true parameter is close to 0 (see Section 3.2). The homogeneous assumption (W.1) was also considered before to study the statistical and computational complexities of optimization algorithms (Ren et al., 2022). It is important to highlight that Assumption (W.1) is applicable to missing data problems, including the informative non-response model (refer to Section 4.1 in (Ho et al., 2020)), and stochastic frontier models (Lee & Chesher, 1986).

Statistical rate of sample NormGD operator $F_n^{\texttt{NGD}}$: To establish the statistical and computational complexities of sample NormGD updates θ_n^t, we utilize the population to sample analysis (Yi & Caramanis, 2015; Balakrishnan et al., 2017; Ho et al., 2020; Kwon et al., 2021). In particular, using triangle inequality, we have

$$\|\theta_n^t - \theta^\star\| \leq \underbrace{\|\theta_n^t - \theta^t\|}_{=:\varepsilon_{\text{stab}}^t} + \underbrace{\|\theta^t - \theta^\star\|}_{=:\varepsilon_{\text{opt}}^t}, \tag{3}$$

Given this decomposition, the statistical error of θ_n^t is controlled by two terms:

(1) $\varepsilon_{\text{stab}}^t$: the uniform concentration of the sample operator F_n^{NGD} around the population operator F^{NGD};

(2) $\varepsilon_{\text{opt}}^t$: the contraction rate of population operator.

For $\varepsilon_{\text{opt}}^t$ in equation (3), the homogeneous assumption (W.1) entails the following contraction rate of population NormGD operator.

Lemma 2.1. *Suppose Assumption (W.1) holds for some $\alpha > 0$ and some universal constants c_1, c_2. Then, if the step-size $\eta \leq \frac{c_1^2}{2c_2^2}$, then we have that*

$$\|F(\theta) - \theta^\star\| \leq \kappa \|\theta - \theta^\star\|,$$

where $\kappa < 1$ is a universal constant that only depends on η, c_1, c_2, α.

The proof of Lemma 2.1 can be found in Appendix B.1. For $\varepsilon_{\text{stab}}^t$ in equation (3), the uniform concentration bound between F_n^{NGD} and F^{NGD}, can be obtained via the following assumption on the concentration bound of $\|\nabla^2 f_n(\theta) - \nabla^2 f(\theta)\|_{\text{op}}$ for $\theta \in \mathbb{B}(\theta^\star, r)$.

(W.2) **(Stability of Second-order Information)** For a given parameter $\gamma \geq 0$, there exist a noise function $\varepsilon : \mathbb{N} \times (0, 1] \to \mathbb{R}^+$, universal constant $c_3 > 0$, and some positive parameter $\rho > 0$ such that

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla^2 f_n(\theta) - \nabla^2 f(\theta)\|_{\text{op}} \leq c_3 r^\gamma \varepsilon(n, \delta),$$

for all $r \in (0, \rho)$ with probability $1 - \delta$.

The idea of Assumption (W.2) is to control the growth of the noise function, which is the difference between the population and sample loss functions, via the second-order information of these loss functions. The second-order stability assumption has been previously examined by (Mei et al., 2018). It is essential to emphasize that while both our study and (Mei et al., 2018) address the uniform convergence of the Hessian, our specific emphasis lies in leveraging the additional r^γ dependency within the uniform concentration bound to attain a more precise statistical dependency. A simple example for Assumption (W.2) is when $f_n(\theta) = \frac{\|\theta\|^{2p}}{2p} - \omega \frac{\|\theta\|^{2q}}{2q} \sqrt{\frac{d}{n}}$ where $\omega \sim \mathcal{N}(0, 1)$ and p, q are some positive integer numbers such that $p > q$. Then, $f(\theta) = \|\theta\|^{2p}/2p$. The Assumption (W.2) is satisfied with $\gamma = 2q - 2$ and with the noise function $\varepsilon(n, \delta) = \sqrt{\frac{d \log(1/\delta)}{n}}$. For concrete statistical examples, we demonstrate later in Section 3 that Assumption (W.2) is satisfied by the generalized linear model and mixture model. Given Assumption (W.2), we have the following uniform concentration bound between the sample and population NormGD operators

Lemma 2.2. *Assume that Assumptions (W.1) and (W.2) hold with $\alpha \geq \gamma + 1$. Furthermore, assume that $\nabla f_n(\theta^\star) = 0$. Then, we obtain that*

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r) \setminus \mathbb{B}(\theta^\star, r_n)} \|F_n^{NGD}(\theta) - F^{NGD}(\theta)\| \leq c_4 r^{\gamma + 1 - \alpha} \varepsilon(n, \delta),$$

where $r_n := \left(\frac{6c_3\varepsilon(n,\delta)}{c_1}\right)^{\frac{1}{\alpha-\gamma}}$*, and* c_4 *is a universal constant depends on* $\eta, c_1, c_2, c_3, \alpha, \gamma$*.*

The proof of Lemma 2.2 can be found in Appendix B.2. We have a few remarks with Lemma 2.2. First, the assumption that $\nabla f_n(\theta^\star) = 0$ is to guarantee the local stability of $\nabla f_n(\theta)$ around $\nabla f(\theta)$ (Ho et al., 2020). This assumption is mild and satisfied by several models, such as low signal-to-noise regimes in generalized linear models, as demonstrated in Section 3, and Gaussian mixture models, as discussed in Section 3.2. Such assumption can also be removed when $\alpha = 0$, namely, the population loss function f is locally strongly convex and smooth (See Appendix A). Second, the assumption that $\alpha \geq \gamma + 1$ means that the signal is stronger than the noise in statistical models, which in turn leads to meaningful statistical rates. Third, the inner radius r_n in Lemma 2.2 corresponds to the final statistical radius, at the order $\mathcal{O}(\varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}})$. It means that we cannot go beyond that radius, otherwise, the empirical Hessian is not positive definite. Based on the contraction rate of the population NormGD operator in Lemma 2.1 and the uniform concentration of the sample NormGD operator around the population NormGD operator in Lemma 2.2, we have the following result on the statistical and computational complexities of the sample NormGD iterates around the true parameter $\theta^\star$.

Theorem 2.3. *Assume that Assumptions (W.1) and (W.2) and assumptions in Lemma 2.2 hold with* $\alpha \geq \gamma + 1$*. Assume that the sample size* n *is large enough such that* $\varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}} \leq \frac{(1-\kappa)r}{c_4\bar{C}^{\gamma+1-\alpha}}$ *where* κ *is defined in Lemma 2.1,* c_4 *is the universal constant in Lemma 2.2 and* $\bar{C} = (\frac{6c_3}{c_1})^{\frac{1}{\alpha-\gamma}}$*, and* r *is the local radius. Then, there exist universal constants* C_1*,* C_2 *such that with probability* $1-\delta$*, for* $t \geq C_1 \log(1/\varepsilon(n,\delta))$*, the following holds:*

$$\min_{k\in\{0,1,\cdots,t\}} \|\theta_n^k - \theta^\star\| \leq C_2 \cdot \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}}.$$

The proof of Theorem 2.3 follows the argument of part (b) of Theorem 2 in (Ho et al., 2020); therefore, it is omitted. A few comments with Theorem 2.3 are in order.

On the approximation of $\lambda_{\max}$: Computing the whole spectrum of a $d \times d$ matrix requires $\mathcal{O}(d^3)$ computation. But fortunately, we can compute the maximum eigenvalue in $\mathcal{O}(d^2)$ computation with the well-known power iteration (*a.k.a* power method, see Chapter 7.3, Golub & Van Loan, 1996)) which has broad applications in different areas (e.g. Hardt et al., 2016). Power iteration can compute the maximum eigenvalue up to ε error with at most $\mathcal{O}\left(\frac{\log\varepsilon}{\log(\lambda_2/\lambda_{\max})}\right)$ matrix vector products, where λ_2 is the second largest eigenvalue. Hence, when $\lambda_2/\lambda_{\max}$ is bounded away from 1, we can obtain a high-quality approximation of $\lambda_{\max}$ with a small number of computations. There can be some issues when $\lambda_2/\lambda_{\max}$ is close to 1. But in fact, we only require an approximation of $\lambda_{\max}$ within statistical accuracy defined in Assumption (W.2) (also see the proof in Appendix B.2). Hence, without loss of generality, we can assume $\lambda_2(\nabla^2 f_n(\theta)) \leq \lambda_{\max}(\nabla^2 f_n(\theta)) - c_3\|\theta-\theta^\star\|^\gamma\varepsilon(n,\delta)$, which means $\frac{\lambda_2(\nabla^2 f_n(\theta))}{\lambda_{\max}(\nabla^2 f_n(\theta))} \leq 1 - \frac{c_3}{c_2}\|\theta-\theta^\star\|^{\gamma-\alpha}$. Since $\alpha \geq \gamma+1$ and we only consider the case $\|\theta-\theta^\star\| \leq r$, there exists a universal constant $c_{\mathrm{PI}} < 1$ that does not depend on n, d, such that $\lambda_2/\lambda_{\max} \leq c_{\mathrm{PI}}$. Therefore, we can compute $\lambda_{\max}$ with a small number of iterations. However, as pointed out by (Kuczyński & Woźniakowski, 1992), the requirement for $\lambda_2/\lambda_{\max}$ to be bounded away from 1 can be relaxed by introducing a probabilistic framework through randomization in the algorithm.

Comparing to fixed step size gradient descent: Under the Assumptions (W.1) and (W.2), we have the following result on the statistical and computational complexities of fixed step size GD iterates.

Proposition 2.4. *Assume that Assumptions (W.1) and (W.2) hold with* $\alpha \geq \gamma+1$ *and* $\nabla f_n(\theta^\star) = 0$*. Suppose the sample size* n *is large enough so that* $\varepsilon(n,\delta) \leq C$ *for some universal constant* C*. Then there exist universal constant* C_1 *and* C_2*, such that for any fixed* $\tau \in \left(0, \frac{1}{\alpha-\gamma}\right)$*, as long as* $t \geq C_1\varepsilon(n,\delta)^{-\frac{\alpha}{\alpha-\gamma}}\log\frac{1}{\tau}$*, we have that*

$$\|\theta_{n,GD}^t - \theta^\star\| \leq C_2\varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}-\tau}.$$

The proof of Proposition 2.4 is similar to Proposition 1 in (Ren et al., 2022), and we omit the proof here. Therefore, the results in Theorem 2.3 indicate that the NormGD and fixed-step size GD iterates reach the same statistical radius $\varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}}$ within the true parameter $\theta^\star$. Nevertheless, the NormGD only takes $\mathcal{O}(\log(1/\varepsilon(n,\delta)))$ number of iterations while the fixed step size GD takes $\mathcal{O}(\varepsilon(n,\delta)^{-\frac{\alpha}{\alpha-\gamma}})$ number of iterations. If the dimension d is fixed, the total computational complexity of NormGD algorithm is at the order of $\mathcal{O}(n\cdot\log(1/\varepsilon(n,\delta)))$, which is much cheaper than that of fixed-step size GD, $\mathcal{O}(n\cdot\varepsilon(n,\delta)^{-\frac{\alpha}{\alpha-\gamma}})$, to reach the final statistical radius.

Comparing to Newton's method and BFGS. To the best of our knowledge, general theories for Newton's method and BFGS algorithms are currently lacking. However, in Section 3, we will compare the performance of NormGD to Newton's method and BFGS for specific models in terms of both computational complexity and statistical optimality.

3. Examples

In this section, we consider an application of our theories in the previous section to the generalized linear model and Gaussian mixture models.

3.1. Generalized Linear Model (GLM)

Generalized linear model (GLM) has been a widely used model in statistics and machine learning (Nelder & Wedderburn, 1972). It is a generalization of linear regression model where we use a link function to relate the covariates to the response variable. In particular, we assume that $(Y_1, X_1), \ldots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$ satisfy

$$Y_i = g(X_i^\top \theta^\star) + \varepsilon_i. \qquad \forall i \in [n] \tag{4}$$

Here, $g : \mathbb{R} \to \mathbb{R}$ is a given link function, $\theta^\star$ is a true but unknown parameter, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. noises from $\mathcal{N}(0, \sigma^2)$ where $\sigma > 0$ is a given variance parameter. We consider the random design setting where $X_1, \ldots, X_n$ are i.i.d. from $\mathcal{N}(0, I_d)$. A few comments with our model assumption. First, in our paper, we will not estimate the link function g. Second, the assumption that the noise follows the Gaussian distribution is just for the simplicity of calculations; similar proof argument still holds for sub-Gaussian noise. For the purpose of our theory, we consider the link function $g(r) := r^p$ for any $p \in \mathbb{N}$ and $p \geq 2$. When $p = 2$, the generalize linear model becomes the phase retrieval problem (Fienup, 1982; Shechtman et al., 2015; Candes et al., 2011; Netrapalli et al., 2015).

Least-square loss: We estimate the true parameter $\theta^\star$ via minimizing the least-square loss function:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}_n(\theta) := \frac{1}{2n} \sum_{i=1}^n (Y_i - (X_i^\top \theta)^p)^2. \tag{5}$$

By letting n goes to infinity, we obtain the population least-square loss function of GLM:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) := \frac{1}{2} \mathbb{E}_{X,Y}[(Y - (X^\top \theta)^p)^2],$$

where the outer expectation is taken with respect to $X \sim \mathcal{N}(0, I_d)$ and $Y = g(X^\top \theta^\star) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. It is clear that $\theta^\star$ is the global minimum of the population loss function $\mathcal{L}$. Furthermore, the function $\mathcal{L}$ is homogeneous, i.e., all directions have similar behaviors. In this section, we consider two regimes of the GLM for our study of sample NormGD iterates: Strong signal-to-noise ratio (SNR) regime and Low signal-to-noise ratio regime.

Strong signal-to-noise ratio regime: The strong SNR regime corresponds to the setting when $\theta^\star$ is bounded away from 0 and $\|\theta^\star\|$ is sufficiently large, i.e., $\|\theta^\star\| \geq C$ for some universal constant C. Under this setting, the population loss function $\mathcal{L}$ is locally strongly convex and smooth, i.e., it satisfies Assumption (S.1) under the homogeneous setting with all fast directions. Furthermore, in Appendix C.2 we prove that for Assumption (S.2), for any radius $r > 0$ there exist universal constants C_1, C_2, C_3 such that as long as $n \geq C_1(d \log(d/\delta))^{2p}$ the following bounds hold

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla \mathcal{L}_n(\theta) - \nabla \mathcal{L}(\theta)\| \leq C_2 \sqrt{\frac{d + \log(1/\delta)}{n}},$$
$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{\text{op}} \leq C_3 \sqrt{\frac{d + \log(1/\delta)}{n}} \tag{6}$$

with probability at least $1 - \delta$.

Low signal-to-noise ratio regime: The low SNR regime corresponds to the setting when the value of $\|\theta^\star\|$ is sufficiently small. To simplify the computation, we assume that $\theta^\star = 0$. Direct calculation shows that $\nabla \mathcal{L}_n(\theta^\star) = 0$. Furthermore, the population loss function becomes

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \frac{\sigma^2 + (2p-1)!!\|\theta - \theta^\star\|^{2p}}{2}. \tag{7}$$

Under this setting, the function $\mathcal{L}$ is no longer locally strong convex around $\theta^\star = 0$. Indeed, this function is homogeneous with all slow directions (see Appendix C.1 for the proof):

$$\lambda_{\max}(\nabla^2 \mathcal{L}(\theta)) \leq c_1 \|\theta - \theta^\star\|^{2p-2}, \tag{8}$$
$$\lambda_{\min}(\nabla^2 \mathcal{L}(\theta)) \geq c_2 \|\theta - \theta^\star\|^{2p-2}, \tag{9}$$

for all $\theta \in \mathbb{B}(\theta^\star, r)$ for some $r > 0$. Here, c_1, c_2 are some universal constants depending on r. Therefore, the homogeneous Assumption (W.1) is satisfied with $\alpha = 2p - 2$. Moving to Assumption (W.2), we demonstrate in Appendix C.2 that there exist universal constants C_1 and C_2 such that for $r > 0$ and $n \geq C_1(d \log(d/\delta))^{2p}$:

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{\text{op}} \leq C_2 (r^{p-2} + r^{2p-2}) \sqrt{(d + \log(1/\delta))/n} \tag{10}$$

with probability at least $1 - \delta$. Hence, Assumption (W.2) is satisfied with $\gamma = p - 2$. Based on the above results, Theorems 2.3 for homogeneous settings with all slow directions and A.3 for homogeneous settings with all fast directions lead to the following statistical and computational complexities of NormGD algorithm.

Corollary 3.1. *Given the generalized linear model (4) with $g(r) = r^p$ for some $p \in \mathbb{N}$ and $p \geq 2$, there exists universal constants $c, \tilde{c}_1, \tilde{c}_2, \bar{c}_1, \bar{c}_2$ such that when the sample size $n \geq c(d \log(d/\delta))^{2p}$ and the initialization $\theta_n^0 \in \mathbb{B}(\theta^\star, r)$ for some chosen radius $r > 0$, with probability $1 - \delta$ the sequence of sample NormGD iterates $\{\theta_n^t\}_{t \geq 0}$ satisfies the following bounds:*

(i) When $\|\theta^\star\| \geq C$ for some universal constant C,

$$\|\theta_n^t - \theta^\star\| \leq \tilde{c}_1 \sqrt{\frac{d + \log(1/\delta)}{n}},$$

holds for $t \geq \tilde{c}_2 \log(n/(d + \log(1/\delta)))$,

(ii) When $\theta^\star = 0$, we obtain

$$\min_{1\le k\le t} \|\theta_n^k - \theta^\star\| \le c_1' \left(\frac{d+\log(1/\delta)}{n}\right)^{1/2p},$$

holds for $t \ge c_2' \log\left(n/(d+\log(1/\delta))\right)$.

Comparison with other methods. For the strong SNR regime, the sample NormGD only takes logarithmic number of iterations $\log(n)$ to reach the optimal statistical radius $\mathcal{O}((d/n)^{1/2})$ around the true parameter. This guarantee is similar to that of the fixed step size GD iterates for solving the locally strongly convex and smooth loss function (Balakrishnan et al., 2017; Ho et al., 2020). For the low SNR regime, the sample NormGD iterates reach the final statistical radius $\mathcal{O}((d/n)^{\frac{1}{2p}})$ after logarithmic number of iterations in terms of n. In terms of the number of iterations, it is cheaper than that of the fixed step size GD algorithm, which takes at least $\mathcal{O}((n/d)^{\frac{p-1}{p}})$ number of iterations (See our discussion after Theorem 2.3). It indicates that the total computational complexity of NormGD algorithm, which is at the order of $\mathcal{O}((nd+d^2)\log(n/d))$, is smaller than that of fixed step size GD, which is $\mathcal{O}(n^{\frac{2p-1}{2p}} d^{\frac{1}{p}})$. Conversely, as demonstrated in (Jin et al., 2024), BFGS requires a logarithmic number of iterations to attain a statistical radius of $\mathcal{O}((d/n)^{\frac{1}{2p+2}})$ in the low SNR regime which can be improved to $\mathcal{O}((d/n)^{\frac{1}{2p}})$ in the univariate case. This is suboptimal when compared to the performance of NormGD and GD iterates. The statistical performance of Newton's method has not been conclusively established for the multivariate scenario. However, (Ho et al., 2020) demonstrated that in the univariate case, Newton's method can achieve the final statistical radius of $\mathcal{O}((d/n)^{\frac{1}{2p}})$. It is indeed not clear if the statistical behavior of Newton's method and BFGS can be improved to the optimal statistical accuracy of $\mathcal{O}((d/n)^{\frac{1}{2p}})$ in the multivariate case. Consequently, the overall computational complexity for these methods becomes $\mathcal{O}\left((nd+d^3)\cdot\log(n/d)\right)$ and $\mathcal{O}\left((nd+d^2)\log(n/d)\right)$, respectively. A summary of the results for the low SNR regime is presented in Table 1. Therefore, for the low SNR regime, the NormGD algorithm is more computationally efficient than the fixed step size GD, Newton's method, and BFGS for reaching a similar final statistical radius. Finally, we provide experiments to verify our theory on the statistical guarantee of the NormGD in Appendix F.

3.2. Gaussian Mixture Models (GMM)

We now consider Gaussian mixture models (GMM), one of the most popular statistical models for modeling heterogeneous data (Lindsay, 1995; McLachlan & Basford, 1988). Parameter estimation in these models plays an important role in capturing the heterogeneity of different subpopulations. The common approach to estimate the location and scale parameters in these model is via maximizing the log-likelihood function. The statistical guarantee of the maximum likelihood estimator (MLE) in Gaussian mixtures had been studied in (Chen, 1995; Ho & Nguyen, 2016). However, since the log-likelihood function is non-concave, we do not have closed-form expressions for the MLE. Therefore, in practice we utilize optimization algorithms to approximate the MLE. However, a complete picture about the statistical and computational complexities of these optimization algorithms has remained missing. In order to shed light on the behavior of NormGD algorithm for solving GMM, we consider a simplified yet important setting of this model, symmetric two-component location GMM. This model had been used in the literature to study the statistical behaviors of Expectation-Maximization (EM) algorithm (Balakrishnan et al., 2017; Dwivedi et al., 2020b). We assume that the data $X_1, X_2, \ldots, X_n$ are i.i.d. samples from $\frac{1}{2}\mathcal{N}(-\theta^\star, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta^\star, \sigma^2 I_d)$ where $\sigma > 0$ is given and $\theta^\star$ is true but unknown parameter. Our goal is to obtain an estimation of $\theta^\star$ via also using the symmetric two-component location Gaussian mixture:

$$\frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d). \tag{11}$$

As we mentioned earlier, we obtain an estimation of $\theta^\star$ via maximizing the sample log-likelihood function associated with model (11), which is given by:

$$\min_{\theta\in\mathbb{R}^d} \bar{\mathcal{L}}_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \log\Bigg(\frac{1}{2}\phi(X_i|\theta,\sigma^2 I_d) + \frac{1}{2}\phi(X_i|-\theta,\sigma^2 I_d)\Bigg). \tag{12}$$

Here, $\phi(\cdot|\theta,\sigma^2 I_d)$ is density function of Gaussian distribution with mean θ and covariance $\sigma^2 I_d$. Similar to GLM, we also consider two regimes of the true parameter: Strong signal-to-noise regime when $\|\theta^\star\|/\sigma$ is sufficiently large and Low signal-to-noise regime when $\|\theta^\star\|/\sigma$ is sufficiently small. To analyze the behaviors of sample NormGD iterates, we define the population version of the estimation (12) as follows:

$$\min_{\theta\in\mathbb{R}^d} \bar{\mathcal{L}}(\theta) := -\mathbb{E}\Bigg[\log\Bigg(\frac{1}{2}\phi(X|\theta,\sigma^2 I_d) + \frac{1}{2}\phi(X|-\theta,\sigma^2 I_d)\Bigg)\Bigg]. \tag{13}$$

where Here, the outer expectation is taken with respect to $X \sim \frac{1}{2}\mathcal{N}(-\theta^\star, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta^\star, \sigma^2 I_d)$. We can check that $\bar{\mathcal{L}}$ is also homogeneous in all directions. The strong SNR regime corresponds to the setting when $\bar{\mathcal{L}}$ is homogeneous with all fast directions while the low SNR regime is associated with the setting when $\bar{\mathcal{L}}$ is homogeneous with all slow

directions. In this section, we only focus on establishing the statistical behaviors of the NormGD algorithm under the low SNR regime while these behaviors under the strong SNR regime can be found in Appendix E.

Low signal-to-noise regime: Now we move to the low SNR regime, namely, when $\|\theta^\star\|/\sigma$ is sufficiently small. For the simplicity of computation we specifically assume that $\theta^\star = 0$. Under this setting, the true model becomes a single Gaussian distribution with mean 0 and covariance matrix $\sigma^2 I_d$ while the fitted model (11) has two components with similar weights and symmetric means. This setting is widely referred to as over-specified mixture model, namely, we fit the true mixture model with more components than needed, in statistics and machine learning (Chen, 1995; Rousseau & Mengersen, 2011). It is important in practice as the true number of components is rarely known and to avoid underfitting the true model, we tend to use a fitted model with more components than the true number of components. In Appendix D.1, we prove that the population loss function $\bar{\mathcal{L}}$ is homogeneous with all slow directions and satisfy the following properties:

$$\lambda_{\max}(\nabla^2 \bar{\mathcal{L}}(\theta)) \leq c_1 \|\theta - \theta^\star\|^2, \tag{14}$$

$$\lambda_{\min}(\nabla^2 \bar{\mathcal{L}}(\theta)) \geq c_2 \|\theta - \theta^\star\|^2, \tag{15}$$

for all $\theta \in \mathbb{B}(\theta^\star, \frac{\sigma}{2})$ where c_1 and c_2 are some universal constants. Therefore, the population loss function $\bar{\mathcal{L}}$ satisfies Assumption (W.1) with $\alpha = 2$. For the stability of second-order information, we prove in Appendix D.2 that there exist universal constants C_1 and C_2 such that for any $r > 0$, with probability $1 - \delta$

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla^2 \bar{\mathcal{L}}_n(\theta) - \nabla^2 \bar{\mathcal{L}}(\theta)\| \leq C_2 \sqrt{\frac{d \log(1/\delta)}{n}}, \tag{16}$$

as long as $n \geq C_1 d \log(1/\delta)$. The uniform concentration bound (16) shows that for the low SNR regime of two-component location Gaussian mixtures, the stability of second-order information in Assumption (W.2) is satisfied with $\gamma = 0$. Moreover, from Lemma 1 in (Dwivedi et al., 2020a) we know $\nabla \bar{\mathcal{L}}_n(\theta^\star) = 0$. Combining the results from the homogeneous behaviors of population loss function in equations (14), (15), and the uniform concentration bound in equation (16) to the result of Theorem 2.3, we obtain that the NormGD updates reach the final statistical radius $(d/n)^{1/4}$ after $\log(n)$ number of iterations. Now, we would like to formally state the statistical behaviors of the NormGD iterates for both the strong and low SNR regimes.

Corollary 3.2. *Given the symmetric two-component mixture model* (11)*, we can find positive universal constants* $c, \bar{c}_1, \bar{c}_2, c_1', c_2'$ *such that with probability at least* $1-\delta$*, when* $n \geq cd \log(1/\delta)$ *the sequence of NormGD iterates* $\{\theta_n^t\}_{t \geq 0}$ *satisfies the following bounds:*

(i) When $\|\theta^\star\| \geq C$ *for some constant* C *and the initialization* $\theta_n^0 \in \mathbb{B}(\theta^\star, \frac{\|\theta\|^\star}{4})$*, we obtain that*

$$\|\theta_n^t - \theta^\star\| \leq \bar{c}_1 \sqrt{\frac{d \log(1/\delta)}{n}},$$

as long as $t \geq \bar{c}_2 \log\left(n/(d \log(1/\delta))\right)$*,*

(ii) Under the setting $\theta^\star = 0$ *and the initialization* $\theta_n^0 \in \mathbb{B}(\theta^\star, \frac{\sigma}{2})$*, we have*

$$\min_{1 \leq k \leq t} \|\theta_n^k - \theta^\star\| \leq c_1' \left(\frac{d \log(1/\delta)}{n}\right)^{1/4},$$

for $t \geq c_2' \log\left(n/(d \log(1/\delta))\right)$*.*

Comparison with other methods. For the low SNR regime, Newton's method and NormGD iterates reach the final statistical radius of $\mathcal{O}((d/n)^{1/4})$ after a logarithmic number of iterations in terms of n while the EM iterates reach the same radius after $\mathcal{O}((n/d)^{1/2})$ number of iterations (Dwivedi et al., 2020b). It demonstrates that the total computational complexity for NormGD is on the order of $\mathcal{O}((nd + d^2) \cdot \log n)$, which represents a significantly more efficient compared to the EM algorithm and Newton's method, both of which operate at the order of $\mathcal{O}(n^{3/2} d^{-1/2})$ and $\mathcal{O}((nd + d^3) \cdot \log n)$, respectively. It is worth mentioning that there is currently no established set of results for BFGS in the GMM setting. Finally, we provide experiments to verify our theory on the statistical guarantee of the NormGD in Appendix F.

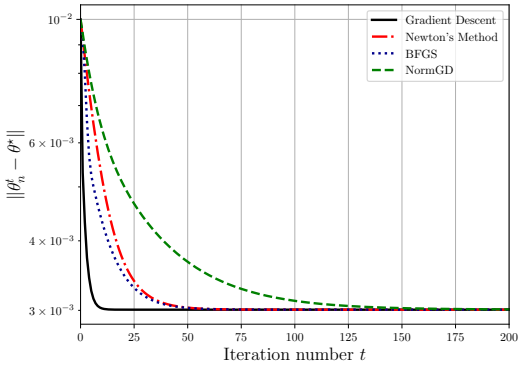
4. Experiments

In this section, we perform numerical experiments on the generalized linear model to empirically verify our theoretical results regarding the convergence rates and the statistical rates of the sample iterates. We conducted a comparative analysis between NormGD and fixed step-size gradient descent, Newton's method, and BFGS as baseline methods. Further information about the experimental setup and supplementary numerical experiments for the Gaussian mixture model can be found in Appendix F.

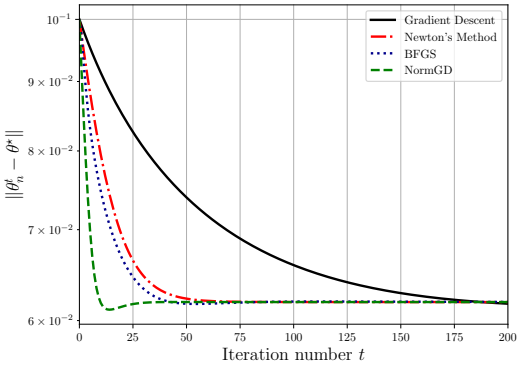
Convergence Rate. We run each algorithm in low and high SNR settings to verify our theoretical results for the generalized linear model. As shown in Figure 1a when in the strong signal-to-noise setting, all methods and our proposed Normalized Gradient Descent method (referred to as NormGD) converge linearly. However, once we shift to the low signal-to-noise setting, all second-order methods, that is, Newton's, BFGS, and NormGD, converge linearly, while GD converges sublinearly, as shown in Figure 1b.

Statistical Rate. To further verify our corollaries, especially how the statistical error scales with n, we plot the statistical error versus sample size in Figure 2. The experiments were

Figure 1: **Left**: All methods converge linearly in the high signal-to-noise setting; **Right**: all second-order methods converge linearly in the low signal-to-noise setting while GD converges sub-linearly.

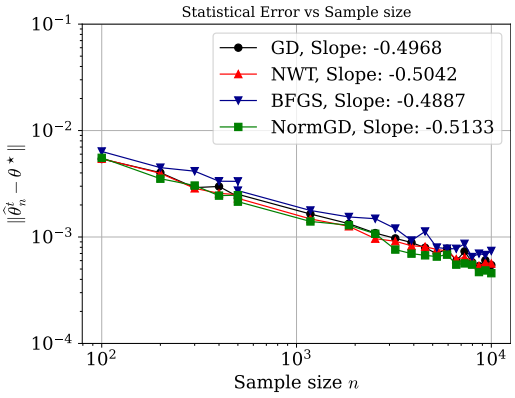


(a) Convergence rate of sample iterates in High SNR setting with $n = 1000$

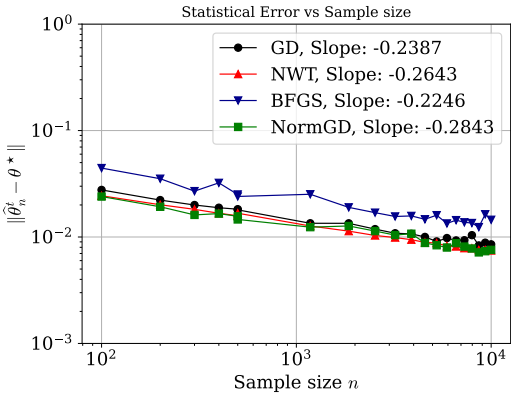


(b) Convergence rate of sample iterates in Low SNR setting with $n = 1000$

Figure 2: **Left**: (High SNR) The statistical error of all methods roughly scales with $n^{-0.5}$; **Right**: (Low SNR) the statistical error roughly scales with $n^{-0.25}$ for all methods.



(a) Statistical rate of sample iterates in High SNR setting with $d = 2$



(b) Statistical rate of sample iterates in Low SNR setting with $d = 2$

repeated for 10 times for 20 sample sizes from $n_{\min} = 100$ to $n_{\max} = 10000$ and the average of the statistical error is shown. The slope is computed as the linear regression coefficient of the log sample size versus the log statistical error. As in this log-log plot, in the strong signal-to-noise setting, the statistical error roughly scales with $n^{-0.5}$, while in the low signal-to-noise setting, the statistical error roughly scales with $n^{-0.25}$. This coincides with our theory as in Corollary 3.1.

5. Conclusion

In this paper, we show that by utilizing second-order information in the design of optimization algorithms, we are able to improve the computational complexity of these algorithms for solving parameter estimation in statistical models. In particular, we study the statistical and computational complexities of the NormGD algorithm, a variant of gradient descent algorithm whose step size is scaled by the maximum eigenvalue of the Hessian matrix of the loss function. We show that when the population loss function is homogeneous, the NormGD algorithm only needs a logarithmic number of iterations to reach the final statistical radius around the true parameter. In terms of iteration complexity and total computational complexity, it is cheaper than fixed step size GD algorithm, which requires a polynomial number of iterations to reach the similar statistical radius under the singular statistical model settings. It is worth mentioning that beyond homogeneous assumption even the theoretical guarantee for popular second order methods like Newton's method or BFGS has not been established. We leave a rigorous theory for NormGD and other methods beyond assumption (W.1) in future work. We wish to remark that there are potentially more efficient algorithms than NormGD by employing more structures of the Hessian matrix, such as using the trace of the Hessian matrix as the scaling factor of the GD algorithm. We leave a detailed development for such direction in future work.

Acknowledgments

NH acknowledges support from the NSF IFML 2019844 and the NSF AI Institute for Foundations of Machine Learning.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are potential societal consequences of our work; however, given the theoretical nature of the paper, we are unable to anticipate any negative consequences here.

References

Balakrishnan, S., Wainwright, M. J., and Yu, B. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

Candes, E. J., Eldar, Y., Strohmer, T., and Voroninski, V. Phase retrieval via matrix completion, 2011.

Chen, J. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995.

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I., and Yu, B. Sharp analysis of expectation-maximization for weakly identifiable models. *AISTATS*, 2020a.

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I., and Yu, B. Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics*, 44: 2726–2755, 2020b.

Fienup, J. R. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21(15):2758–2769, Aug 1982. doi: 10.1364/AO.21.002758. URL http://www.osapublishing.org/ao/abstract.cfm?URI=ao-21-15-2758.

Golub, G. H. and Van Loan, C. F. Matrix computations. johns hopkins studies in the mathematical sciences, 1996.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/hardt16.html.

Ho, N. and Nguyen, X. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016.

Ho, N., Khamaru, K., Dwivedi, R., Wainwright, M. J., Jordan, M. I., and Yu, B. Instability, computational efficiency and statistical accuracy. *Arxiv Preprint Arxiv: 2005.11411*, 2020.

Jin, Q., Ren, T., Ho, N., and Mokhtari, A. Statistical and computational complexities of BFGS quasi-newton method for generalized linear models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=PIL3YWXmx2.

Kuczyński, J. and Woźniakowski, H. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992. doi: 10.1137/0613066. URL https://doi.org/10.1137/0613066.

Kwon, J. Y., Ho, N., and Caramanis, C. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *AISTATS*, 2021.

Lee, L.-F. and Chesher, A. Specification testing when score test statistics are identically zero. *Journal of Econometrics*, 31(2):121–149, 1986. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(86)90045-X. URL https://www.sciencedirect.com/science/article/pii/030440768690045X.

Lindsay, B. *Mixture Models: Theory, Geometry and Applications*. In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995.

McLachlan, G. J. and Basford, K. E. *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs.* New York, 1988.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A), December 2018. ISSN 0090-5364. doi: 10.1214/17-aos1637. URL http://dx.doi.org/10.1214/17-AOS1637.

Mou, W., Ho, N., Wainwright, M. J., Bartlett, P., and Jordan, M. I. A diffusion process perspective on posterior contraction rates for parameters. *arXiv preprint arXiv:1909.00966*, 2019.

Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135, 1972.

Nesterov, Y. *Lectures on Convex Optimization*. Springer, 2018.

Netrapalli, P., Jain, P., and Sanghavi, S. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015. doi: 10.1109/TSP.2015.2448516.

Polyak, B. T. *Introduction to Optimization.* Optimization Software, Inc., New York, 1987.

Ren, T., Cui, F., Atsidakou, A., Sanghavi, S., and Ho, N. Towards statistical and computational complexities of Polyak step size gradient descent. In *AISTATS, 2022*, 2022.

Rousseau, J. and Mengersen, K. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:689–710, 2011.

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. doi: 10.1109/MSP.2014.2352673.

van der Vaart, A. W. and Wellner, J. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Yi, X. and Caramanis, C. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pp. 1567–1575, 2015.

Zhuo, J., Kwon, J., Ho, N., and Caramanis, C. On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv: 2102.02756*, 2021.

Supplementary Materials for "Improving Computational Complexity in Statistical Models with Local Curvature Information"

In the supplementary material, we collect proofs and results deferred from the main text. In Appendix A, we provide general theory for the statistical guarantee of NormGD for the homogeneous settings with all fast directions of the population loss function. In Appendix B, we provide proofs for the main results in the main text. We then provide proofs for the statistical and computational complexities of NormGD under generalized linear models and mixture models respectively in Appendices C and D. We study the statistical behaviors of the NormGD under the strong signal-to-noise regime of the symmetric two-component location Gaussian mixtures in Appendix E. Finally, we provide experiments to illustrate the statistical behaviors of the NormGD iterates under the generalized linear model and the Gaussian mixture model in Appendix F.

A. Homogeneous Settings with All Fast Directions

In this Appendix, we provide statistical guarantee for the NormGD iterates when the population loss function is homogeneous with all fast directions. Following the population to sample analysis in equation (3), we first consider the strong convexity and Lipschitz smoothness assumptions that characterize all fast directions.

(S.1) (Strong convexity and Lipschitz smoothness) For some radius $r > 0$, for all $\theta \in \mathbb{B}(\theta^\star, r)$ we have

$$\bar{c}_1 \leq \lambda_{\min}(\nabla^2 f(\theta)) \leq \lambda_{\max}(\nabla^2 f(\theta)) \leq \bar{c}_2,$$

where $\bar{c}_1 > 0$ and $\bar{c}_2 > 0$ are some universal constants depending on r.

The Assumption (S.1) is a special case of Assumption (W.1) when $\alpha = 0$. A simple example for the function f that satisfies Assumption (S.1) is $f(\theta) = \|\theta\|^2$.

Given the Assumption (S.1), we obtain the following result for the contraction of the population NormGD operator F around the true parameter $\theta^\star$.

Lemma A.1. *Assume Assumption (S.1) holds for some universal constants $\bar{c}_1, \bar{c}_2$. Then, if the step-size $\eta \leq \frac{\bar{c}_2^2}{2\bar{c}_1^2}$, then we have that*

$$\|F(\theta) - \theta^\star\| \leq \bar{\kappa}\|\theta - \theta^\star\|,$$

where $\bar{\kappa} < 1$ is a universal constant that only depends on $\eta, \bar{c}_1, \bar{c}_2$.

The proof of Lemma A.1 is a direct from the proof of Lemma 2.1 with $\alpha = 0$; therefore, its proof is omitted.

(S.2) (Stability of first and second-order information) For some fixed positive parameter $r > 0$, there exist a noise function $\varepsilon : \mathbb{N} \times (0, 1] \to \mathbb{R}^+$, and universal constants $\bar{c}_3, \bar{c}_4 > 0$ depends on r, such that

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla f_n(\theta) - \nabla f(\theta)\| \leq \bar{c}_3 \cdot \varepsilon(n, \delta),$$
$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla^2 f_n(\theta) - \nabla^2 f(\theta)\|_{\text{op}} \leq \bar{c}_4 \cdot \varepsilon(n, \delta).$$

for all $r \in (0, r)$ with probability $1 - \delta$.

We would like to remark that the assumption in the uniform concentration of $\nabla f_n(\theta)$ around $\nabla f(\theta)$ is standard for analyzing optimization algorithms for solving parameter estimation under locally strongly convex and smooth population loss function (Balakrishnan et al., 2017; Ho et al., 2020). The extra assumption on the uniform concentration of the empirical Hessian matrix $\nabla^2 f_n(\theta)$ around the population Hessian matrix $\nabla^2 f(\theta)$ is to ensure that $\lambda_{\max}(\nabla^2 f_n(\theta)$ in NormGD algorithm will stay close to $\lambda_{\max}(\nabla^2 f(\theta))$. These two conditions are sufficient to guarantee the stability of the sample NormGD operator F_n around the population NormGD operator in the following lemma.

Lemma A.2. *Assume that Assumption (S.2) holds, and n is sufficiently large such that $\bar{c}_1 > 2\bar{c}_3\varepsilon(n,\delta)$. Then, we obtain that*

$$\sup_{\theta\in\mathbb{B}(\theta^\star,r)} \|F_n(\theta) - F(\theta)\| \le \bar{c}_5\varepsilon(n,\delta),$$

and $\bar{c}_5$ is a universal constant depends on $\eta, \bar{c}_1, \bar{c}_2, \bar{c}_3, \bar{c}_4$.

Proof. With straightforward calculation, we have that

$$\begin{aligned}
\|F_n(\theta) - F(\theta)\| \le& \eta\left(\left\|\frac{\nabla f(\theta)(\lambda_{\max}(\nabla^2 f(\theta)) - \lambda_{\max}(\nabla^2 f_n(\theta)))}{\lambda_{\max}(\nabla^2 f_n(\theta))\lambda_{\max}(\nabla^2 f(\theta))}\right\| + \left\|\frac{\nabla f_n(\theta) - \nabla f(\theta)}{\lambda_{\max}(\nabla^2 f_n(\theta))}\right\|\right)\\
\le& \eta\left(\frac{\bar{c}_2\bar{c}_3\varepsilon(n,\delta)}{(\bar{c}_1 - \bar{c}_3\varepsilon(n,\delta))\bar{c}_1} + \frac{\bar{c}_4\varepsilon(n,\delta)}{\bar{c}_1 - \bar{c}_3\varepsilon(n,\delta)}\right)\\
\le& \eta\left(\frac{2\bar{c}_2\bar{c}_3 + 2\bar{c}_1\bar{c}_4}{\bar{c}_1{}^2}\right)\varepsilon(n,\delta).
\end{aligned}$$

Take $\bar{c}_5$ accordingly, we conclude the proof. □

Theorem A.3. *Assume Assumptions (S.1) and (S.2) hold, and n is sufficient large such that $\bar{c}_1 > 2\bar{c}_3\varepsilon(n,\delta)$ and $\bar{c}_5\varepsilon(n,\delta) \le (1-\bar{\kappa})r$ where $\bar{\kappa}$ is the constant defined in Lemma A.1. Then, there exist universal constants $\bar{C}_1$, $\bar{C}_2$ such that for $t \ge \bar{C}_1\log(1/\varepsilon(n,\delta))$, the following holds:*

$$\|\theta_n^t - \theta^\star\| \le \bar{C}_2 \cdot \varepsilon(n,\delta).$$

Proof. With the triangle inequality, we have that

$$\begin{aligned}
\|\theta_n^{t+1} - \theta^\star\| =& \|F_n(\theta_n^t) - \theta^\star\|\\
\le& \|F_n(\theta_n^t) - F(\theta_n^t)\| + \|F(\theta_n^t) - \theta^\star\|\\
\le& \sup_{\theta\in\mathbb{B}(\theta^\star,r)} \|F_n(\theta) - F(\theta)\| + \bar{\kappa}\|\theta_n^t - \theta^\star\|\\
\le& \bar{c}_5\varepsilon(n,\delta) + \bar{\kappa}r \le r.
\end{aligned}$$

Hence, we know $\|\theta_n^t - \theta^\star\| \le r$ for all $t \in \mathbb{N}$. Furthermore, by repeating the above argument T times, we can obtain that

$$\begin{aligned}
\|\theta_n^T - \theta^\star\| \le& \bar{c}_5\varepsilon(n,\delta)\left(\sum_{t=0}^{T-1}\bar{\kappa}^t\right) + \bar{\kappa}^T\|\theta_n^0 - \theta^\star\|\\
\le& \frac{\bar{c}_5}{1-\bar{\kappa}}\varepsilon(n,\delta) + \bar{\kappa}^T r.
\end{aligned}$$

By choosing $T \le \frac{\log(r)+\log(1/\varepsilon(n,\delta))}{\log(1/\bar{\kappa})}$, we know $\bar{\kappa}^T r \le \varepsilon(n,\delta)$, hence

$$\|\theta_n^T - \theta^\star\| \le \left(\frac{\bar{c}_5}{1-\bar{\kappa}} + 1\right)\varepsilon(n,\delta).$$

Take $\bar{C}_1$, $\bar{C}_2$ accordingly, we conclude the proof. □

B. Proofs of Main Results

In this Appendix, we provide proofs for the results in the main text.

B.1. Proof of Lemma 2.1

We start from the following lemma:

Lemma B.1. *Assume Assumption (W.1) holds, we have that*

$$f(\theta) - f(\theta^\star) \geq \frac{c_1\|\theta-\theta^\star\|^{\alpha+2}}{(\alpha+1)(\alpha+2)}.$$

Proof. Consider $g(\theta) = f(\theta) - \frac{c_1\|\theta-\theta^\star\|^{\alpha+2}}{(\alpha+1)(\alpha+2)}$. With Assumption (W.1), we know that

$$\nabla^2 g(\theta) = \nabla^2 f(\theta) - \frac{c_1}{(\alpha+1)(\alpha+2)}\left(\alpha(\alpha+2)\|\theta-\theta^\star\|^{\alpha-2}(\theta-\theta)^\star(\theta-\theta^\star)^\top + (\alpha+2)\|\theta-\theta^\star\|^\alpha I\right) \succeq 0,$$

as the operator norm of $\alpha(\alpha+2)\|\theta-\theta^\star\|^{\alpha-2}(\theta-\theta^\star)(\theta-\theta^\star)^\top + (\alpha+2)\|\theta\|^\alpha I$ is less than $(\alpha+1)(\alpha+2)\|\theta-\theta^\star\|^\alpha$. Meanwhile, we have that

$$\nabla g(\theta) = \nabla f(\theta) - \frac{c_1\|\theta-\theta^\star\|^\alpha}{\alpha+1}(\theta-\theta^\star).$$

As $\nabla f(\theta^\star) = 0$, we know $\nabla g(\theta^\star) = 0$, which means $\theta^\star$ is the minimizer of g. Hence,

$$f(\theta^\star) = g(\theta^\star) \leq g(\theta) = f(\theta) - \frac{c_1\|\theta-\theta^\star\|^{\alpha+2}}{(\alpha+1)(\alpha+2)},$$

which means

$$f(\theta) - f(\theta^\star) \geq \frac{c_1\|\theta-\theta^\star\|^{\alpha+2}}{(\alpha+1)(\alpha+2)}.$$

As a consequence, we obtain the conclusion of Lemma B.1. □

Now, we prove Lemma 2.1. Notice that

$$\begin{aligned}
\|F(\theta)-\theta^\star\|^2 =& \left\|\theta - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\nabla f(\theta) - \theta^\star\right\|^2 \\
=&\|\theta-\theta^\star\|^2 - \frac{2\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\langle\nabla f(\theta), \theta-\theta^\star\rangle + \frac{\eta^2}{\lambda^2_{\max}(\nabla^2(f(\theta)))}\|\nabla f(\theta)\|^2 \\
=&\|\theta-\theta^\star\|^2 - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\left(2\langle\nabla f(\theta), \theta-\theta^\star\rangle - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\|\nabla f(\theta)\|^2\right) \\
\leq&\|\theta-\theta^\star\|^2 - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\left(2(f(\theta)-f(\theta^\star)) - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\|\nabla f(\theta)\|^2\right),
\end{aligned}$$

where the last inequality is due to the convexity. With Assumption (W.1), we have that

$$\begin{aligned}
\|\nabla f(\theta)\| =& \left\|\int_0^1 \nabla^2 f(\theta^\star + t(\theta-\theta^\star))(\theta-\theta^\star)dt\right\| \\
\leq& \int_0^1 \left\|\nabla^2 f(\theta^\star + t(\theta-\theta^\star))(\theta-\theta^\star)\right\| dt \\
\leq& \int_0^1 \lambda_{\max}(\nabla^2 f(\theta^\star + t(\theta-\theta^\star)))\|\theta-\theta^\star\|dt \\
\leq& \int_0^1 c_2 t^\alpha \|(\theta-\theta^\star)\|^\alpha\|\theta-\theta^\star\|dt \\
\leq& \frac{c_2}{\alpha+1}\|\theta-\theta^\star\|^{\alpha+1}.
\end{aligned}$$

As $\eta \leq \frac{c_1^2}{2c_2^2} \leq \frac{c_1^2(\alpha+1)}{c_2^2(\alpha+2)}$, we have that

$$\frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\left(f(\theta) - f(\theta^\star) - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\|\nabla f(\theta)\|^2\right) \geq \frac{\eta}{c_2}\left(\frac{c_1}{(\alpha+1)(\alpha+2)} - \frac{\eta c_2^2}{c_1(\alpha+1)^2}\right)\|\theta - \theta^\star\|^2.$$

Hence, we find that

$$\|F(\theta) - \theta^\star\|^2 \leq \left(1 - \frac{\eta}{c_2}\left(\frac{c_1}{(\alpha+1)(\alpha+2)} - \frac{\eta c_2^2}{c_1(\alpha+1)^2}\right)\right)\|\theta - \theta^\star\|^2.$$

Take κ accordingly, we conclude the proof.

B.2. Proof of Lemma 2.2

Notice that

$$\begin{aligned}
\|F_n(\theta) - F(\theta)\| =& \left\|\frac{\eta}{\lambda_{\max}(\nabla^2 f_n(\theta))}\nabla f_n(\theta) - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\nabla f(\theta)\right\| \\
=& \eta\left\|\frac{\nabla f_n(\theta)\lambda_{\max}(\nabla^2 f(\theta)) - \nabla f(\theta)\lambda_{\max}(\nabla^2 f_n(\theta))}{\lambda_{\max}(\nabla^2 f_n(\theta))\lambda_{\max}(\nabla^2 f(\theta))}\right\| \\
\leq& \eta\left(\left\|\frac{\nabla f(\theta)(\lambda_{\max}(\nabla^2 f(\theta)) - \lambda_{\max}(\nabla^2 f_n(\theta)))}{\lambda_{\max}(\nabla^2 f_n(\theta))\lambda_{\max}(\nabla^2 f(\theta))}\right\| + \left\|\frac{\nabla f_n(\theta) - \nabla f(\theta)}{\lambda_{\max}(\nabla^2 f_n(\theta))}\right\|\right).
\end{aligned}$$

For the term $\|\nabla f_n(\theta) - \nabla f(\theta)\|$, we have that

$$\begin{aligned}
\|\nabla f_n(\theta) - \nabla f(\theta)\| \leq& \|\nabla f_n(\theta^\star) - \nabla f(\theta^\star)\| \\
&+ \left\|\int_0^1 (\nabla^2 f_n(\theta^\star + t(\theta - \theta^\star)) - \nabla^2 f(\theta^\star + t(\theta - \theta^\star)))(\theta - \theta^\star)dt\right\| \\
\leq& \int_0^1 \|(\nabla^2 f_n(\theta^\star + t(\theta - \theta^\star)) - \nabla^2 f(\theta^\star + t(\theta - \theta^\star)))(\theta - \theta^\star)\|dt \\
\leq& \int_0^1 \|\nabla^2 f_n(\theta^\star + t(\theta - \theta^\star)) - \nabla^2 f(\theta^\star + t(\theta - \theta^\star))\|_{\mathrm{op}}\|\theta - \theta^\star\|dt \\
\leq& \int_0^1 c_3 t^\gamma \varepsilon(n,\delta)\|\theta - \theta^\star\|^{\gamma+1}dt \\
=& \frac{c_3\|\theta - \theta^\star\|^{\gamma+1}\varepsilon(n,\delta)}{\gamma+1}.
\end{aligned}$$

Meanwhile, it's straightforward to show that

$$|\lambda_{\max}(\nabla^2 f_n(\theta)) - \lambda_{\max}(\nabla f(\theta))| \leq 3c_3 r^\gamma \varepsilon(n,\delta).$$

Hence, we have that

$$\begin{aligned}
\|F_n(\theta) - F(\theta)\| \leq& \eta\left(\left\|\frac{\nabla f(\theta)(\lambda_{\max}(\nabla^2 f(\theta)) - \lambda_{\max}(\nabla^2 f_n(\theta)))}{\lambda_{\max}(\nabla^2 f_n(\theta))\lambda_{\max}(\nabla^2 f(\theta))}\right\| + \left\|\frac{\nabla f_n(\theta) - \nabla f(\theta)}{\lambda_{\max}(\nabla^2 f_n(\theta))}\right\|\right) \\
\leq& \eta\left(\frac{3c_2c_3r^{\gamma+1-\alpha}\varepsilon(n,\delta)}{(\alpha+1)(c_1r^\alpha - 3c_3r^\gamma\varepsilon(n,\delta))c_1r^\alpha} + \frac{c_3r^\gamma\varepsilon(n,\delta)}{(\gamma+1)(c_1r^\alpha - 3c_3r^{\gamma+1}\varepsilon(n,\delta))}\right).
\end{aligned}$$

As $r \geq \left(\frac{6c_3\varepsilon(n,\delta)}{c_1}\right)^{1/(\alpha-\gamma)}$, we can further have

$$\|F_n(\theta) - F(\theta)\| \leq \eta\left(\frac{6c_2c_3}{(\alpha+1)c_1^2} + \frac{2c_3}{(\gamma+1)c_1}\right)r^{\gamma+1-\alpha}\varepsilon(n,\delta).$$

Taking c_4 accordingly, we conclude the proof.

C. Proof of Generalized Linear Models

In this appendix, we provide the proof for the NormGD in generalized linear models.

C.1. Homogeneous assumptions

Based on the formulation of the population loss function $\mathcal{L}$ in equation (7), we have

$$\nabla\mathcal{L}(\theta) = 2p(2p-1)!!(\theta-\theta^\star)\|\theta-\theta^\star\|^{2p-2},$$
$$\nabla^2\mathcal{L}(\theta) = (2p(2p-1)!!)\|\theta-\theta^\star\|^{2p-4}\left(\|\theta-\theta^\star\|^2 I_d + (2p-4)(\theta-\theta^\star)(\theta-\theta^\star)^\top\right).$$

Notice that, $\theta-\theta^\star$ is an eigenvector of $\|\theta-\theta^\star\|^2 I_d + (2p-4)(\theta-\theta^\star)(\theta-\theta^\star)^\top$ with eigenvalue $(2p-3)\|\theta-\theta^\star\|^2$, and any vector that is orthogonal to $\theta-\theta^\star$ (which forms a $d-1$ dimensional subspace) is an eigenvector of $\|\theta-\theta^\star\|^2 I_d + (2p-4)(\theta-\theta^\star)(\theta-\theta^\star)^\top$ with eigenvalue $\|\theta-\theta^\star\|^2$. Hence, we have that

$$\lambda_{\max}(\|\theta-\theta^\star\|^2 I_d + (2p-4)(\theta-\theta^\star)(\theta-\theta^\star)^\top) = (2p-3)\|\theta-\theta^\star\|^2,$$
$$\lambda_{\min}(\|\theta-\theta^\star\|^2 I_d + (2p-4)(\theta-\theta^\star)(\theta-\theta^\star)^\top) = \|\theta-\theta^\star\|^2,$$

which shows that $\mathcal{L}(\theta)$ satisfies the homogeneous assumption.

C.2. Uniform concentration bound

The proof for the first concentration bound (6) is in Appendix D.1 of (Ren et al., 2022); therefore, it is omitted. We focus on proving the second uniform concentration bounds (6) and (10) for the Hessian matrix $\nabla^2\mathcal{L}_n(\theta)$ around the Hessian matrix $\nabla^2\mathcal{L}(\theta)$ under both the strong and low signal-to-noise regimes. Indeed, we would like to show the following uniform concentration bound that captures both the bounds (6) and (10).

Lemma C.1. *There exist universal constants C_1 and C_2 such that as long as $n \geq C_1(d\log(d/\delta))^{2p}$ we obtain that*

$$\sup_{\theta\in\mathbb{B}(\theta^\star,r)} \|\nabla^2\mathcal{L}_n(\theta) - \nabla^2\mathcal{L}(\theta)\|_{op} \leq C_2\left((r+\|\theta^\star\|)^{p-2} + (r+\|\theta^\star\|)^{2p-2}\right)\sqrt{\frac{d+\log(1/\delta)}{n}}. \tag{17}$$

Proof of Lemma C.1. Direct calculation shows that

$$\nabla^2\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \left(p(2p-1)(X_i^\top\theta)^{2p-2} - p(p-1)Y_i(X_i^\top\theta)^{p-2}\right)X_iX_i^\top,$$
$$\nabla^2\mathcal{L}(\theta) = \mathbb{E}\left[p(2p-1)(X^\top\theta)^{2p-2} - p(p-1)(X^\top\theta^\star)^p(X^\top\theta)^{p-2}XX^\top\right].$$

Therefore, we obtain

$$\begin{aligned}\nabla^2\mathcal{L}_n(\theta) - \nabla^2\mathcal{L}(\theta) =& p(2p-1)\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{2p-2}X_iX_i^\top - \mathbb{E}\left[(X^\top\theta)^{2p-2}XX^\top\right]\right)\\ &- p(p-1)\left(\frac{1}{n}\sum_{i=1}^n Y_i(X_i^\top\theta)^{p-2}X_iX_i^\top - \mathbb{E}\left[(X^\top\theta^\star)^p(X^\top\theta)^{p-2}XX^\top\right]\right).\end{aligned}$$

Using the triangle inequality with the operator norm, the above equation leads to

$$\sup_{\theta\in\mathbb{B}(\theta^\star,r)} \|\nabla^2\mathcal{L}_n(\theta) - \nabla^2\mathcal{L}(\theta)\|_{\text{op}} \leq C(A_1+A_2+A_3), \tag{18}$$

where C is some universal constant and A_1, A_2, A_3 are defined as follows:

$$\begin{aligned}A_1 &= \sup_{\theta\in\mathbb{B}(\theta^\star,r)}\left\|\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{2p-2}X_iX_i^\top - \mathbb{E}\left[(X^\top\theta)^{2p-2}XX^\top\right]\right\|_{\text{op}},\\ A_2 &= \sup_{\theta\in\mathbb{B}(\theta^\star,r)}\left\|\frac{1}{n}\sum_{i=1}^n (Y_i - (X_i^\top\theta^\star)^p)(X_i^\top\theta)^{p-2}X_iX_i^\top\right\|_{\text{op}},\\ A_3 &= \sup_{\theta\in\mathbb{B}(\theta^\star,r)}\left\|\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^\star)^p(X_i^\top\theta)^{p-2}X_iX_i^\top - \mathbb{E}\left[(X^\top\theta^\star)^p(X^\top\theta)^{p-2}XX^\top\right]\right\|_{\text{op}}.\end{aligned} \tag{19}$$

With variational characterization of the operator norm and upper bound the norm of any $\theta \in \mathbb{B}(\theta^\star, r)$ with $r + \|\theta^\star\|$, we have

$$\begin{aligned} A_1 &\leq (r + \|\theta^\star\|)^{2p-2} T_1, \\ A_2 &\leq (r + \|\theta^\star\|)^{p-2} T_2, \\ A_3 &\leq (r + \|\theta^\star\|)^{p-2} T_3, \end{aligned}$$

where the terms T_1, T_2, T_3 are defined as follows:

$$\begin{aligned} T_1 &:= \sup_{u \in \mathbb{S}^{d-1}, \theta \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (X_i^\top \theta)^{2p-2} (X_i^\top u)^2 - \mathbb{E}\left[(X^\top \theta)^{2p-2} (X^\top u)^2 \right] \right| \\ T_2 &:= \sup_{u \in \mathbb{S}^{d-1}, \theta \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (Y_i - (X_i^\top \theta^\star)^p)(X_i^\top \theta)^{p-2} (X_i^\top u)^2 \right| \\ T_3 &:= \sup_{u \in \mathbb{S}^{d-1}, \theta \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (X_i^\top \theta^\star)^p (X_i^\top \theta)^{p-2} (X_i^\top u)^2 - \mathbb{E}\left[(X^\top \theta^\star)^p (X^\top \theta)^{p-2} (X^\top u)^2 \right] \right|, \end{aligned}$$

where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. We know consider the high probability bound of each individual terms following the proof strategy in (Ren et al., 2022).

Bound for T_2: Assume U is a $1/8$-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$ with at most 17^d elements, the standard discretization arguments (e.g. Chapter 6 in (Wainwright, 2019)) show

$$\sup_{u \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (Y_i - (X_i^\top \theta^\star)^p)(X_i^\top \theta)^{p-2} (X_i^\top u)^2 \right| \leq 2 \sup_{u \in U} \left| \frac{1}{n} \sum_{i=1}^n (Y_i - (X_i^\top \theta^\star)^p)(X_i^\top \theta)^{p-2} (X_i^\top u)^2 \right|.$$

With a symmetrization argument, we know for any even integer $q \geq 2$,

$$\begin{aligned} &\mathbb{E}\left(\sup_{u \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (Y_i - (X_i^\top \theta^\star)^p)(X_i^\top \theta)^{p-2} (X_i^\top u)^2 \right| \right)^q \\ &\qquad \leq \mathbb{E}\left(\sup_{u \in \mathbb{S}^{d-1}} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^\star)^p)(X_i^\top \theta)^{p-2} (X_i^\top u)^2 \right| \right)^q, \end{aligned}$$

where $\{\varepsilon_i\}_{i \in [n]}$ is a set of i.i.d. Rademacher random variables. Furthermore, for a compact set Ω, we define

$$\mathcal{R}(\Omega) := \sup_{\theta \in \Omega, p' \in [1,p]} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^\star)^p)(X_i^\top \theta)^{p'-2} (X_i^\top u)^2 \right|,$$

where $\mathcal{N}(t)$ is a t-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$. Then we have

$$\mathcal{R}(\mathbb{S}^{d-1}) \leq 2\mathcal{R}(\mathcal{N}(t)) + 3^{p-2} t \mathcal{R}(\mathbb{S}^{d-1}).$$

By taking $t = 3^{-p+1}$, we obtain that $\mathcal{R}(\mathbb{S}^{d-1}) \leq 3\mathcal{R}(\mathcal{N}(3^{-p+1}))$. Furthermore, with the union bound, for any $q \geq 1$ we have that

$$\sup_{\theta \in \mathbb{S}^{d-1}, p' \in [2,p]} \mathbb{E}\left[\left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^\star)^p((X_i^\top \theta)^{p'-2} (X_i^\top u)^2 \right|^q \right] \leq \frac{\mathbb{E}[\mathcal{R}^q(\mathcal{N}(3^{-p+1}))]}{p|\mathcal{N}(3^{-p+1})|}.$$

Therefore, we only need to consider $\mathbb{E}\left[\left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^\star)^p)(X_i^\top \theta)^{p'-2} (X_i^\top u)^2 \right|^q \right]$. An application of Khintchine's inequality (Boucheron et al., 2013) demonstrates that we can find a universal constant C such that for all $p' \in [2, p]$, we have

$$\begin{aligned} &\mathbb{E}\left[\left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^\star)^p)(X_i^\top \theta)^{p'-2} (X_i^\top u)^2 \right|^q \right] \\ &\qquad \leq \mathbb{E}\left[\left(\frac{Cq}{n^2} \sum_{i=1}^n (Y_i - (X_i^\top \theta^\star)^p)^2 (X_i^\top \theta)^{2(p'-2)} (X_i^\top u)^4 \right)^{q/2} \right]. \end{aligned}$$

From the assumptions on Y_i and X_i for all $i \in [n]$, we have

$$\begin{aligned}\mathbb{E}\left[(Y_i - (X_i^\top\theta^\star)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\right] \leq &(2p')^{p'},\\ \mathbb{E}\left[\left((Y - (X_i^\top\theta^\star)^p)^2(X_i^\top\theta)^{2(p'-4)}(X_i^\top u)^4\right)^{q/2}\right] \leq &(2p'q)^{p'q}.\end{aligned}$$

From Lemma 2 in (Mou et al., 2019), we have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n\left((Y - (X_i^\top\theta^\star)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\right)^{q/2} - \mathbb{E}\left[(Y_i - (X_i^\top\theta^\star)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\right]\right|\\ &\quad\leq (8p')^{p'}\sqrt{\frac{\log 4/\delta}{n}} + (2p'\log(n/\delta))^{p'}\frac{\log 4/\delta}{n}\end{aligned}$$

with probability at least $1-\delta$. Hence, we find that

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n(Y_i - (X_i^\top\theta^\star)^p)^2(X_i^\top\theta)^{2(p'-1)}(X_i^\top u)^4\right)^{q/2}\right]\\ &\quad\leq(4p')^{p'q} + 2^{q/2}\int_0^\infty \mathbb{P}\Bigg[\Bigg|\sum_{i=1}^n\left((Y_i - (X_i^\top\theta^\star)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\right)^{q/2}\\ &\qquad - \mathbb{E}\left[(Y_i - (X_i^\top\theta^\star)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\right]\Bigg| \geq \lambda\Bigg]d\lambda^{q/2}\\ &\quad\leq(4p')^{p'q} + C'p'q\left((32p')^{p'q/2}n^{-q/4})\Gamma(q/4)\right.\\ &\qquad\left.+(8p')^{(p'+1)q/2}n^{-q/2}\left((\log n)^{(p+1)q/2} + \Gamma((p'+1)q/2)\right)\right),\end{aligned}$$

where C' is a universal constant and $\Gamma(\cdot)$ is the Gamma function. As a result, we have that

$$\begin{aligned}&\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{i=1}^n(Y_i - (X_i^\top\theta^\star)^p)(X_i^\top\theta)^{p-2}X_i^\top u - \mathbb{E}[(X^\top\theta^\star)^p(X^\top\theta)^{p-1}(X^\top u)^4]\right)\right|^q\right]\\ &\quad\leq p\cdot 3^{dp+2d+q}\left(\frac{Cq}{n}\right)^{q/2}\left((4p)^{pq} + 2C'pq\,(32p)^{pq}\,n^{-q/4}\Gamma(q/4)\right.\\ &\quad\left.+(8p)^{(p+1)q/2}n^{-q/2}\left((\log n)^{(p+1)q/2} + \Gamma((p+1)q/2)\right)\right),\end{aligned}$$

for any $u \in U$. Eventually, with union bound, we obtain

$$\begin{aligned}&\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n(Y_i - (X_i^\top\theta^\star)^p)(X_i^\top\theta)^{p-2}X_iX_i^\top\right\|^q\right]\right)^{1/q}\\ &\quad\leq 2\cdot(17)^{d/q}\cdot 3^{\frac{dp+2d}{q}+1}\left[\sqrt{\frac{C_pq}{n}} + \left(\frac{C_pq}{n}\right)^{3/4} + \frac{C_p}{n}(\log n + q)^{(p+1)/2}\right],\end{aligned}$$

where C_p is a universal constant that only depends on p. Take $q = d(p+3) + \log(1/\delta)$ and use the Markov inequality, we get the following bound on the second term T_2 with probability $1-\delta$:

$$T_2 \leq c_1\left(\sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{1}{n}\left(d + \log\left(\frac{n}{\delta}\right)\right)^{\frac{p+1}{2}}\right). \tag{20}$$

Bounds for T_1 and T_3: Using the same argument as that of T_2, we obtain the following high probability bounds for T_1 and T_3:

$$T_1 \leq c_2\left(\sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{1}{n}\left(d+\log\left(\frac{n}{\delta}\right)\right)^{\frac{2p+1}{2}}\right)., \tag{21}$$

$$T_3 \leq c_3\left(\sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{1}{n}\left(d+\log\left(\frac{n}{\delta}\right)\right)^{\frac{2p+1}{2}}\right). \tag{22}$$

with probability $1-\delta$ where c_2 and c_3 are some universal constants. Plugging the bounds (20), (21), and (22) to the bounds (18) and (19), and use the condition that $n \geq C_1(d\log(d/\delta))^{2p}$ we obtain the conclusion of the lemma. □

D. Proof of Gaussian Mixture Models

In this appendix, we provide the proof for the NormGD in Gaussian mixture models.

D.1. Homogeneous Assumptions

The proof for the claim (14) is direct from Appendix A.2.2 from (Ren et al., 2022). Therefore, we only focus on proving the claim (15). Indeed, direct calculation shows that

$$\nabla^2\bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^2}\left(I_d - \frac{1}{\sigma^2}\mathbb{E}\left(XX^\top \text{sech}^2\left(\frac{X^\top\theta}{\sigma^2}\right)\right)\right).$$

We can simplify the computation of $\nabla^2\bar{\mathcal{L}}$ via a change of coordinates. In particular, we choose an orthogonal matrix Q such that $Q\theta = \|\theta\|e_1$. Here, $e_1 = (1,0,\ldots,0)$ is the first canonical basis in dimension d. We then denote $W = \frac{QX}{\sigma}$. Since $X \sim \mathcal{N}(0,\sigma^2 I_d)$, we have $W = (W_1,\cdots,W_d) \sim \mathcal{N}(0,I_d)$. Therefore, we can rewrite $\nabla^2\bar{\mathcal{L}}$ as follows:

$$\nabla^2\bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^2}\left(I_d - \mathbb{E}_W\left(WW^\top \text{sech}^2\left(\frac{W_1\|\theta\|}{\sigma}\right)\right)\right) = \frac{1}{\sigma^2}\left(I_d - B\right).$$

It is clear that the matrix B is diagonal matrix and satisfies that $B_{11} = \mathbb{E}_{W_1}\left[W_1^2\text{sech}^2\left(\frac{W_1\|\theta\|}{\sigma}\right)\right]$, $B_{ii} = \mathbb{E}_{W_1}\left[\text{sech}^2\left(\frac{W_1\|\theta\|}{\sigma}\right)\right]$ for all $2 \leq i \leq d$. An application of $\text{sech}^2(x) \leq 1 - x^2 + \frac{2}{3}x^4$ for all $x \in \mathbb{R}$ shows that

$$B_{11} \leq \mathbb{E}_{W_1}\left[W_1^2\left(1 - \frac{W_1^2\|\theta\|^2}{\sigma^2} + \frac{2W_1^4\|\theta\|^4}{3\sigma^4}\right)\right] = 1 - \frac{3\|\theta\|^2}{\sigma^2} + \frac{10\|\theta\|^4}{\sigma^4},$$

$$B_{ii} \leq \mathbb{E}_{W_1}\left[\left(1 - \frac{W_1^2\|\theta\|^2}{\sigma^2} + \frac{2W_1^4\|\theta\|^4}{3\sigma^4}\right)\right] = 1 - \frac{\|\theta\|^2}{\sigma^2} + \frac{2\|\theta\|^4}{\sigma^4},$$

for all $2 \leq i \leq d$. When $\|\theta\| \leq \frac{\sigma}{2}$, we have that $\frac{\|\theta\|^2}{\sigma^4} \leq \frac{1}{4}$, and hence

$$B_{11} \leq 1 - \frac{3\|\theta\|^2}{\sigma^2} + \frac{10\|\theta\|^4}{\sigma^4} \leq 1 - \frac{\|\theta\|^2}{2\sigma^2},$$

$$B_{ii} \leq 1 - \frac{\|\theta\|^2}{\sigma^2} + \frac{2\|\theta\|^4}{\sigma^4} \leq 1 - \frac{\|\theta\|^2}{2\sigma^2}, \quad \forall\, 2 \leq i \leq d.$$

Hence, as long as $\|\theta\| \leq \sigma/2$ we have that

$$\lambda_{\min}(\nabla^2\bar{L}(\theta)) \geq \frac{\|\theta\|^2}{2\sigma^2},$$

which concludes the proof.

D.2. Uniform Concentration Bounds for Mixture Models

See Corollary 4 in (Balakrishnan et al., 2017) for the proof of the uniform concentration result between $\nabla\bar{\mathcal{L}}_n(\theta)$ and $\nabla\bar{\mathcal{L}}(\theta)$ in equation (24) for the strong signal-to-noise regime. Now, we prove the uniform concentration bounds between $\nabla^2\bar{\mathcal{L}}_n(\theta)$ and $\nabla\bar{\mathcal{L}}(\theta)$ in equations (24) and (16) for both the strong signal-to-noise and low signal-to-noise regimes. It is sufficient to prove the following lemma.

Lemma D.1. *There exist universal constants C_1 and C_2 such that as long as $n \geq C_1 d \log(1/\delta)$ we obtain that*

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{op} \leq C_2(\|\theta^\star\| + \sigma^2)\sqrt{\frac{d + \log(1/\delta)}{n}}. \tag{23}$$

Proof of Lemma D.1. For the sample log-likelihood function of the Gaussian mixture model, direct calculation shows that

$$\bar{\mathcal{L}}_n(\theta) = \frac{\|\theta\|^2 + \frac{1}{n}\sum_{i=1}^n \|X_i\|^2}{2\sigma^2} - \frac{1}{n}\sum_{i=1}^n \log\left(\exp\left(-\frac{X_i^\top \theta}{\sigma^2}\right) + \exp\left(\frac{X_i^\top \theta}{\sigma^2}\right)\right) - \log(2(\sqrt{2\pi})^d \sigma^d).$$

Therefore, we find that

$$\nabla \bar{\mathcal{L}}_n(\theta) = \frac{\theta}{\sigma^2} - \frac{1}{n\sigma^2}\sum_{i=1}^n X_i \tanh\left(\frac{X_i^\top \theta}{\sigma^2}\right),$$

$$\nabla^2 \bar{\mathcal{L}}_n(\theta) = \frac{1}{\sigma^2}\left(I_d - \frac{1}{n\sigma^2}\sum_{i=1}^n X_i X_i^\top \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right)\right),$$

where $\text{sech}^2(x) = \frac{4}{(\exp(-x)+\exp(x))^2}$ for all $x \in \mathbb{R}$.

For the population log-likelihood function, we have

$$\nabla^2 \bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^2}\left(I_d - \frac{1}{\sigma^2}\mathbb{E}\left(XX^\top \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right).$$

Therefore, we obtain that

$$\nabla^2 \bar{\mathcal{L}}_n(\theta) - \nabla^2 \bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^4}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left(XX^\top \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right).$$

Use the variational characterization of operator norm, it's sufficient to consider

$$T = \sup_{u \in \mathbb{S}^{d-1}, \theta \in \mathbb{B}(\theta^\star, r)} \left|\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left((X^\top u)^2 \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right|.$$

With a standard discretization argument (e.g. Chapter 6 in (Wainwright, 2019)), let U be a $1/8$-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$ whose cardinality can be upper bounded by 17^d, we have that

$$\begin{aligned}
&\sup_{u \in \mathbb{S}^{d-1}, \theta \in \mathbb{B}(\theta^\star, r)} \left|\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left((X^\top u)^2 \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right| \\
\leq &2 \sup_{u \in U, \theta \in \mathbb{B}(\theta^\star, r)} \left|\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left((X^\top u)^2 \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right|.
\end{aligned}$$

With a symmetrization argument on probability (van der Vaart & Wellner, 1996), we have that

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \left|\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left((X^\top u)^2 \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right| \geq t\right] \\
\leq &c_1 \mathbb{P}\left[\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right)\right| \geq c_2 t\right],
\end{aligned}$$

where $\{\varepsilon_i\}$ is a set of i.i.d Rademacher random variable and c_1 and c_2 are two positive universal constants. Define

$$Z := \sup_{\theta \in \mathbb{B}(\theta^\star, r)} \left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right)\right|.$$

For $x \in \mathbb{R}$, define $(x)_+ = \max(x, 0)$, $(x)_- = \min(x, 0)$. Furthermore, for random variable X, we denote

$$\|X\|_q = \left(\mathbb{E}[X^q]\right)^{1/q}.$$

With Theorem 15.5 in (Boucheron et al., 2013), there exists an absolute constant C, such that for all $q \geq 2$,

$$\left\|(Z - \mathbb{E}[Z])_+\right\|_q \leq \sqrt{Cq\|V^+\|_{q/2}},$$

where is the Efron-Stein estimate that is defined as

$$V^+ = \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')_+^2)|X\right]$$

where Z_i' is obtained by replacing the variable X_i by an independent copy X_i' (see Section 6.9 of (Boucheron et al., 2013) for the details). With the proof idea of Theorem 15.14 in (Boucheron et al., 2013), V^+ can be bounded as

$$\begin{aligned} V^+ \leq & \sup_{\theta \in \mathbb{B}(\theta^\star, r)} \frac{1}{n}\mathbb{E}\left[(X^\top u)^4 \text{sech}^4\left(\frac{X^\top \theta}{\sigma^2}\right)\right] + \sup_{\theta \in \mathbb{B}(\theta^\star, r)} \frac{4}{n^2}\sum_{i=1}^{n}(X_i^\top u)^4 \text{sech}^4\left(\frac{X_i^\top \theta}{\sigma^2}\right) \\ \leq & \frac{1}{n}\mathbb{E}\left[(X^\top u)^4\right] + \frac{4}{n^2}\sum_{i=1}^{n}(X_i^\top u)^4. \end{aligned}$$

Here we use the fact that $0 \leq \text{sech}^2(x) \leq 1$ for all x in the last step. Notice that, $X_i^\top u \sim \frac{1}{2}\mathcal{N}(u^\top \theta^\star, \sigma^2) + \frac{1}{2}\mathcal{N}(-u^\top \theta^\star, \sigma^2)$. We can verify that there exists absolute constant c, such that

$$\mathbb{E}\left[(X_i^\top u)^{2p}\right] = \mathbb{E}_{X \sim \mathcal{N}(u^\top \theta^\star, \sigma^2)}\left[(X^\top u)^{2p}\right] \leq (2cp)^p(\|\theta^\star\|^2 + \sigma^2)^p.$$

Apply Lemma 2 in (Mou et al., 2019) with $Y_i = (X_i^\top u)^4$, we have that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^\top u)^4 \leq c(\|\theta^\star\|^2 + \sigma^2)^2\left(1 + \sqrt{\frac{\log 1/\delta}{n}} + \frac{\sqrt{\log n/\delta}\log 1/\delta}{n}\right),$$

with probability at least $1 - \delta$ for some absolute constant c. As $n \geq C_1(d + \log 1/\delta)$, we can conclude that

$$V^+ \leq \frac{c'(\|\theta^\star\|^2 + \sigma^2)^2}{n}$$

for some universal constant c'. Furthermore, as $Z \geq 0$, we have that

$$\left\|(Z - \mathbb{E}[Z])_-\right\|_q \leq \mathbb{E}[Z].$$

Hence, with Minkowski's inequality, we have that

$$\|Z\|_q \leq 2\mathbb{E}[Z] + \sqrt{\frac{cq(\|\theta^\star\|^2 + \sigma^2)^2}{n}},$$

for some absolute constant c. We now bound $\mathbb{E}[Z]$. Consider the following function class

$$\mathcal{G} := \left\{g_\theta : X \to (X^\top u)^2 \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right) \middle| \theta \in \mathbb{R}^d\right\}.$$

Clearly, the function class $\mathcal{G}$ has an envelop function $\bar{G}(X) = (X^\top u)^2$. Meanwhile, as the function sech^2 is monotonic in $(-\infty, 0)$ and $(0, \infty)$ and θ here effects only in the form $X^\top \theta$. Following some algebra we know the VC subgraph dimension of $\mathcal{G}$ is at most $d + 2$. Hence, the L_2-covering number of $\mathcal{G}$ can be bounded by

$$\bar{N}(t) := \sup_Q \left|\mathcal{N}(\mathcal{G}, \|\cdot\|_{L_2(Q)}, t\|\bar{G}\|_{L_2(Q)})\right| \leq (1/t)^{c(d+1)}$$

for any $t > 0$, where c is an absolute constant. With Dudley's entropy integral bound (e.g. (Wainwright, 2019, Theorem 5.22)), we have

$$\begin{aligned}\mathbb{E}[Z] \leq & c\sqrt{\frac{\sum_{i=1}^n (X_i^\top u)^4}{n^2}} \int_0^1 \sqrt{1+\bar{N}(t)}dt \\ \leq & c\sqrt{\frac{d(\|\theta^\star\|^2+\sigma^2)^2}{n}}\end{aligned}$$

for some absolute constant c.

Take $q = \log 1/\delta$, use the Markov equality and an union bound over u, we know there exist universal constants C_1 and C_2, such that the following inequality

$$\sup_{\theta\in\mathbb{B}(\theta^\star,r)} \left| \frac{1}{n}\sum_{i=1}^n X_i X_i^\top \operatorname{sech}^2\left(\frac{X_i^\top\theta}{\sigma^2}\right) - \mathbb{E}\left(X_i X_i^\top \operatorname{sech}^2\left(\frac{X^\top\theta}{\sigma^2}\right)\right)\right| \leq C_2(\|\theta^\star\|^2+\sigma^2)\sqrt{\frac{d+\log 1/\delta}{n}}$$

holds with probability at least $1-\delta$ as long as $n \geq C_1(d+\log 1/\delta)$. As a consequence, we obtain the conclusion of the lemma. □

E. Gaussian mixture models: Strong Signal-to-Noise Regime

In this appendix, we study the behavior of the NormGD under the strong signal-to-noise regime of the symmetric two-component Gaussian mixture models (11).

Strong signal-to-noise regime: Recall that we obtain an estimation of $\theta^\star$ via maximizing the sample log-likelihood function associated with model (11). For the strong signal-to-noise regime, we assume that $\|\theta^\star\| \geq C\sigma$ for some universal constant C. Since the function $\bar{\mathcal{L}}$ is locally strongly convex and smooth as long as $\theta \in \mathbb{B}(\theta^\star, \frac{\|\theta^\star\|}{4})$ (see Corollary 1 in (Balakrishnan et al., 2017)), the Assumption (S.1) under the homogeneous setting with all fast directions is satisfied. Furthermore, as long as we choose the radius $r \leq \frac{\|\theta^\star\|}{4}$ and the sample size $n \geq C_1 d\log(1/\delta)$ for some universal constant C_1, with probability at least $1-\delta$ there exist universal constants C_2 and C_3 such that

$$\begin{aligned}&\sup_{\theta\in\mathbb{B}(\theta^\star,r)} \|\nabla\bar{\mathcal{L}}_n(\theta) - \nabla\bar{\mathcal{L}}(\theta)\| \leq C_2\sqrt{d\log(1/\delta)/n}, \\ &\sup_{\theta\in\mathbb{B}(\theta^\star,r)} \|\nabla^2\bar{\mathcal{L}}_n(\theta) - \nabla^2\bar{\mathcal{L}}(\theta)\|_{\text{op}} \leq C_3\sqrt{d\log(1/\delta)/n}.\end{aligned} \tag{24}$$

The proof of claims (24) is in Appendix D.2. In light of Theorem A.3 in Appendix A for homogeneous settings with all fast directions, the NormGD iterates converge to the final statistical radius $(d/n)^{1/2}$ after $\log(n)$ iterations (see Corollary 3.2 for a formal statement of this result).

Comparison with other methods. In the strong signal-to-noise case, Newton's method and NormGD reach the final statistical radius $(d/n)^{1/2}$ around the true parameter $\theta^\star$ after $\log(n)$ number of iterations, while the fixed step size GD algorithm, which is also the EM algorithm for the symmetric two-component mixture, requires $(d/n)^{1/2}$ number of iterations to reach the same statistical radius of $(d/n)^{1/2}$ (Ho et al., 2020).

F. Experimental Setup and Additional Experiments

In this appendix, we initially present the specifics of the experimental setup. Additionally, we include further experiments involving Gaussian mixture models.

F.1. Experimental Setup

Initialization. We initialize the starting point by uniformly sampling a point from a unit sphere centered at the true parameter $\theta^\star$, i.e., $\theta^0 := \theta^\star + \alpha\cdot\rho^0$ where $\rho^0 \sim \text{Uniform}(\mathbb{S}^{d-1})$. Throughout the experiments, for strong signal-to-noise ratio, we set $\theta^\star = \sqrt{d}\cdot\mathbf{1}_d$, and in low signal-to-noise ratio setting, we simply set $\theta^\star = \mathbf{0}_d$.

Baselines. In our simulation, we conducted a comparative analysis between NormGD and fixed step-size gradient descent, Newton's method, and BFGS as the baseline methods. For BFGS algorithm the iterative method is defined as

$$\theta^{t+1} = \theta^t - \eta H_t \nabla f(\theta^t), \qquad \forall t \geq 0, \tag{25}$$

where $H_t \in \mathbb{R}^{d\times d}$ and η is the step size. There are several approaches for approximating H_t leading to different quasi-Newton methods, but in this paper, we focus on the method considered in (Jin et al., 2024), in which H_t is updated as

$$H_t = \left(I - \frac{s_{t-1}u_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\right) H_{t-1} \left(I - \frac{u_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\right) + \frac{s_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}, \qquad \forall t \geq 1, \tag{26}$$

where $s_{t-1} := \theta^t - \theta^{t-1}$ and $u_{t-1} := \nabla f(\theta^t) - \nabla f(\theta^{t-1})$ for all $t \geq 1$. It is worth mentioning that in all experiments we initialize $H_0 := (\nabla^2 f(\theta^0))^{-1}$ to help with the stability of BFGS.

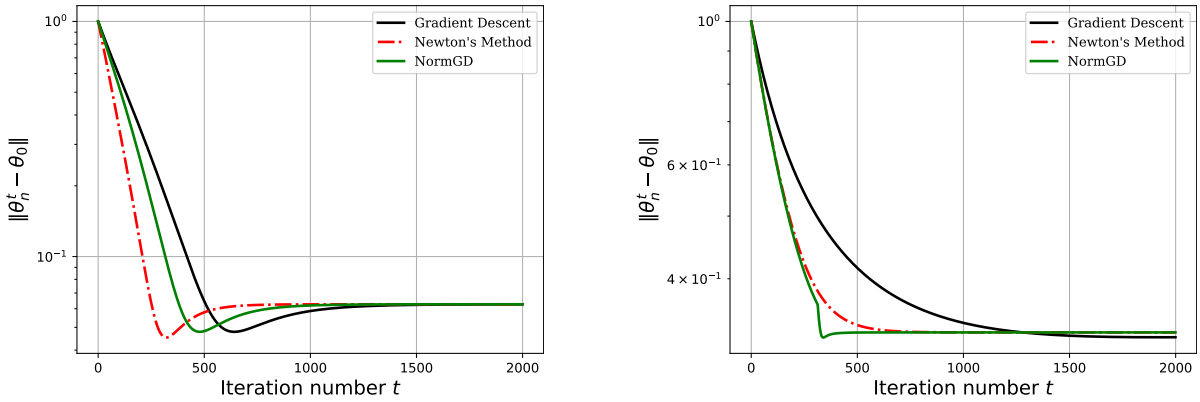
Synthetic Data. In the experiments for GLM described in Section 4, we created a dataset of samples $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ under the following criteria: X_i follows a normal distribution with zero mean and covariance matrix I_d, and $Y_i = g(X_i^\top \theta^\star) + \varepsilon_i$. Here, the function $g(r)$ is defined as $g(r) := r^2$, and the noise terms $\{\varepsilon_i\}_{i=1}^n$ are independent and identically distributed following a zero mean normal distribution with variance σ^2. Throughout the simulations for the generalized linear model, we maintained a sample size of $n = 1000$, a data dimension of $d = 2$, a noise variance of $\sigma^2 = 0.01$, and a learning rate of $\eta = 0.1$.

In the GMM experiments, we generated a set of samples $\{X_i\}_{i=1}^n \subset \mathbb{R}^d$ from the Gaussian mixture of two-component symmetric location: $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(\theta^\star, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(-\theta^\star, \sigma^2 I_d)$. To do so, we first generate n i.i.d. Rademacher random variables $\{\sigma_i\}_{i=1}^n$ uniformly chosen from $\{\pm 1\}$. Then, we generate i.i.d. samples from the conditional distribution of $X_i|\sigma_i \sim \mathcal{N}(\sigma_i\theta^\star, \sigma^2 I_d)$. Throughout the simulations for the Gaussian mixture model, we set the sample size $n = 1000$, the dimension of the data of $d = 5$, and the variance of the Gaussian component $\sigma^2 = 1$.

F.2. Gaussian Mixture Model

In this Appendix, we conduct numerical experiments on the Gaussian mixture model to empirically validate our theoretical findings concerning the convergence rates and the statistical rates of the sample iterates.

Figure 3: *Convergence Rate for the Gaussian Mixture Model (GMM)*. **Left**: All methods converge linearly in the high signal-to-noise setting; **Right**: NormGD and Newton's method in the low signal-to-noise setting converges linearly while GD converges sub-linearly.



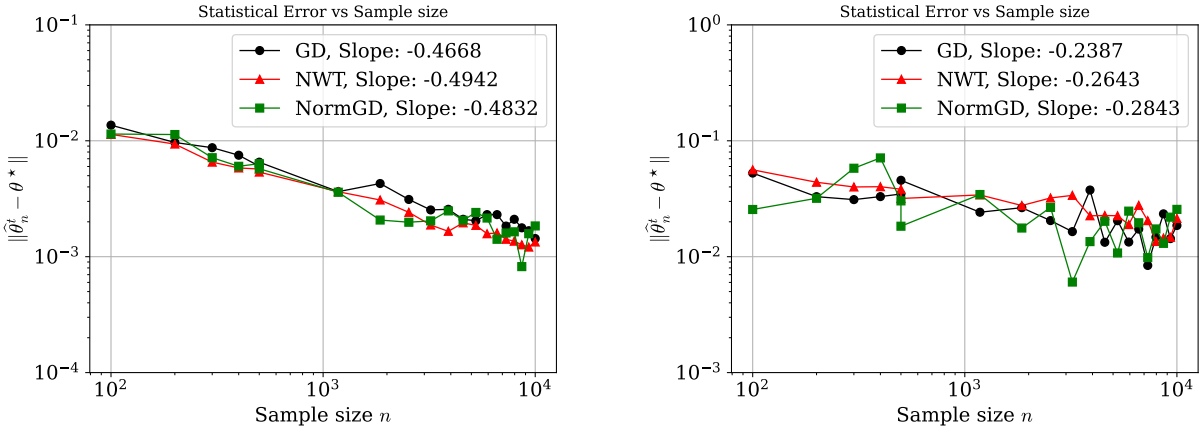
(a) Convergence rate of sample iterates in High SNR setting with $n = 1000$

(b) Convergence rate of sample iterates in Low SNR setting with $n = 1000$

Convergence Rate. To verify our theory in Corollary 3.2, we performed a simulation on Gaussian Mixture Model (GMM), and the results are shown in Figure 3. We fixed the learning rate at $\eta = 0.01$. As shown in Figure 3, we have depicted the convergence rates of Gradient Descent, Newton's method, and NormGD for strong and low signal-to-noise ratio settings. Notably, we do not include a comparison with BFGS due to its instability in both settings. In Figure 3a, we show the linear convergence of sample iterations in the strong signal-to-noise setting. Meanwhile, in Figure 3b, we demonstrate the

sublinear convergence of fixed step-size gradient descent and the linear convergence of Newton's and NormGD in the low SNR setting. Moreover, NormGD converges to a smaller statistical radius compared to the other methods, as shown in Figure 3b.

Figure 4: *Statistical rate for the Gaussian Mixture Model (GMM)*. **Left**: (High SNR) The statistical error of all methods roughly scales with $n^{-0.5}$; **Right**: (Low SNR) the statistical error roughly scales with $n^{-0.25}$ for all methods.



(a) Statistical rate of sample iterates in High SNR setting with $d = 5$

(b) Statistical rate of sample iterates in Low SNR setting with $d = 5$

Statistical Rate. We further validate our corollaries, particularly in terms of how the statistical error scales with the sample size n, we generated a graph showing the statistical error versus the sample size in Figure 4. These experiments were repeated 10 times for 20 sample sizes from $n_{\min} = 100$ to $n_{\max} = 10000$, and the plot displays the average statistical error. The slope was calculated as the linear regression coefficient for the logarithm of the statistical error versus the logarithm of the sample size. In this log-log plot, in the strong signal-to-noise setting, the statistical error appears to scale roughly with $n^{-0.5}$, while in the low signal-to-noise setting, it roughly scales with $n^{-0.25}$. These observations align with our theoretical predictions, as discussed in Corollary 3.2.