Statistical and Computational Complexities of BFGS Quasi-Newton Method for Generalized Linear Models

Anonymous Authors[1]

Abstract

The gradient descent (GD) method has been used widely to solve parameter estimation in generalized linear models (GLMs), a generalization of linear models when the link function can be non-linear. While GD has optimal statistical and computational complexities for estimating the true parameter under the high signal-to-noise ratio (SNR) regime of the GLMs, it has sub-optimal complexities when the SNR is low, namely, the iterates of GD require polynomial number of iterations to reach the final statistical radius due to the local convexity of the least-square loss functions of the GLMs in this case. Even though Newton's method can be used to resolve the flat curvature of the loss functions in the low SNR case, its computational cost is prohibitive in high-dimensional settings as it is $\mathcal{O}(d^3)$. To address the shortcomings of GD and Newton's method, we propose the use of BFGS quasi-Newton method to solve parameter estimation of the GLMs, which has a per iteration cost of $\mathcal{O}(d^2)$. On the optimization side, when the SNR is low, we demonstrate that iterates of BFGS converge linearly to the optimal solution of the population least-square loss function, and the contraction coefficient of the BFGS algorithm is comparable to that of Newton's method. On the statistical side, we prove that the iterates of BFGS reach the final statistical radius of the low SNR GLMs after a logarithmic number of iterations, which is much lower than the polynomial number of iterations of GD.

1. Introduction

In supervised machine learning, we are given a set of n independent samples denoted by $X_1, \ldots, X_n$ with corresponding labels $Y_1, \ldots, Y_n$, that are drawn from some unknown distribution and our goal is to train a model that maps the feature vectors to their corresponding labels. We assume that the data is generated according to distribution $\mathcal{P}_{\theta^*}$ which is parameterized by a ground truth parameter θ^*. Our goal as the learner is to find θ^* by solving the empirical risk minimization (ERM) problem defined as

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; (X_i, Y_i)), \qquad (1)$$

where $\ell(\theta; (X_i, Y_i))$ is the loss function that measures the error between the predicted output of X_i using parameter θ and its true label Y_i. If we define θ_n^* as an optimal solution of the above optimization problem, i.e., $\theta_n^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathcal{L}_n(\theta)$, it can be considered as an approximation of the ground-truth solution θ^*, where θ^* is also a minimizer of the population loss defined as

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) := \mathbb{E}\left[\ell(\theta; (X, Y))\right]. \qquad (2)$$

If one can solve the empirical risk efficiently, the output model could be close to θ^*, when n is sufficiently large. Several works have studied the complexity of iterative methods for solving ERM or directly the population loss, for the case that the objective function is convex or strongly convex with respect to θ (Balakrishnan et al., 2017; Ho et al., 2020; Loh & Wainwright, 2015; Agarwal et al., 2012; Yuan & Zhang, 2013; Dwivedi et al., 2020b; Hardt et al., 2016; Candes et al., 2011). However, when we move beyond linear models, the underlying loss becomes non-convex and therefore the behavior of iterative methods could substantially change, and it is not even clear if they can reach a neighborhood of a global minimizer of the ERM problem.

The focus of this paper is on the generalized linear model (GLM) (Carroll et al., 1997; Netrapalli et al., 2015; Fienup, 1982; Shechtman et al., 2015; Feiyan Tian, 2021) where the labels and features are generated according to a polynomial link function and we have $Y_i = (X_i^\top \theta^*)^p + \zeta_i$, where ζ_i is an additive noise and $p \geq 2$ is an integer. Due to nonlinear structure of the generative model, even if we select a convex loss function ℓ, the ERM problem denoted to the considered GLM could be non-convex with respect to θ. Interestingly, depending on the norm of θ^*, the curvature of the ERM

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

problem and its corresponding population risk minimization problem could change substantially. More precisely, in the case that $\|\theta^*\|$ is sufficiently large, which we refer to this case as the high signal-to-noise ratio (SNR) regime, the underlying population loss of the problem of interest is locally strongly convex and smooth. On the other hand, in the regime that $\|\theta^*\|$ is close to zero, denoted by the low SNR regime, then the underlying problem is neither strongly convex nor smooth, and in fact, it is ill-conditioned.

These observations lead to the conclusion that in the high SNR setting, due to strong convexity and smoothness of the underlying problem, gradient descent (GD) reaches the final statistical radius exponentially fast and overall it only requires logarithmic number of iterations. However, in the low SNR case, as the problem becomes locally convex, GD converges at a sublinear rate to the final statistical radius and thus requires polynomial number of iterations in terms of the sample size. To resolve this issue, (Ren et al., 2022a) recommended the use of GD with Polyak step size to accelerate the convergence of GD in the low SNR case and showed that the number of iterations becomes logarithmic function of the sample size. However, as this method is still a first-order method, its overall complexity scales linearly by the condition number of the problem which depends on the condition number of the feature vectors covariance and the norm $\|\theta^*\|$. Moreover, implementation of Polyak step size requires access to the optimal objective function value. Even in the cases that we can approximate the optimal value sufficiently well, the Polyak iterates still have instability during training due to the influence of noise in the models.

An alternative approach is using Newton's method to handle the ill-conditioning issue in the low SNR case as well as eliminating the need to estimate the optimal function value. As we show in this paper, this idea indeed addresses the issue of poor curvature of the problem and leads to an exponentially fast rate with contraction factor $\frac{2p-2}{2p-1}$, when the sample size is infinite. Moreover, in the high SNR setting, Newton's method converges at a quadratic rate as the problem is strongly convex and smooth. Alas, these improvements come at the expense of increasing the computational complexity of each iteration to $\mathcal{O}(d^3)$ which is indeed more than the per iteration computational cost of GD that is $\mathcal{O}(d)$. These points lead to this question:

> *Is there a computationally-efficient method that performs well in both high and low SNR settings at a reasonable per iteration computational cost?*

Contributions. In this paper, we address this question and show that the BFGS method is capable of achieving these goals. BFGS is a quasi-Newton method that approximates the objective function curvature using gradient information and its per iteration cost is $\mathcal{O}(d^2)$. It is well-known that it enjoys a superlinear convergence rate that is independent of condition number in strongly convex and smooth settings, and hence, in the high SNR setting it outperforms GD. In the low SNR setting, where the Hessian at the optimal solution could be singular, we show that the BFGS method converges linearly and outperforms the sublinear convergence rate of GD. Next, we formally summarize our contributions.

- **Infinite sample, low SNR:** For the infinite sample case where we minimize the population loss, we show that in the low SNR case the iterates of BFGS converge to the ground truth θ^* at an exponentially fast rate that is independent of all problem parameters except the power of link function p. We further show that the linear convergence contraction coefficient of BFGS is comparable to that of Newton's method. This convergence result of BFGS is also of general interest as it provides the first global linear convergence of BFGS without line-search for a setting that is neither strictly nor strongly convex.

- **Finite sample, low SNR:** By leveraging the results developed for the population loss of the low SNR regime, we show that in the finite sample case, the BFGS iterates converge to the final statistical radius $\mathcal{O}(1/n^{1/(2p+2)})$ within the true parameter after a logarithmic number of iterations $\mathcal{O}(\log(n))$. It is substantially lower than the required number of iterations for fixed-step size GD, which is $\mathcal{O}(n^{(p-1)/p})$, to reach a similar statistical radius. This improvement is the direct outcome of the linear convergence of BFGS versus the sublinear convergence rate of GD in the low SNR case. Further, while the iteration complexity of BFGS is comparable to the logarithmic number of iterations of GD with Polyak step size, we show that BFGS removes the dependency of the overall complexity on the condition number of the problem as well as the need to estimate the optimal function value.

- **Experiments:** We conduct numerical experiments for both infinite and finite sample cases to compare the performance of GD (with constant stepsize and Polyak stepsize), Newton's method and BFGS. The provided empirical results are consistent with our theoretical findings and show the advantages of BFGS in the low SNR regime.

2. BFGS Algorithm

In this section, we review the basics of the BFGS quasi-Newton method, which is the main algorithm we analyze. Consider the case that we aim to minimize a differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$. The BFGS update is given by

$$\theta_{k+1} = \theta_k - \eta_k H_k \nabla f(\theta_k), \qquad \forall k \geq 0, \tag{3}$$

where η_k is the step size and $H_k \in \mathbb{R}^{d \times d}$ is a positive definite matrix that aims to approximate the true Hessian inverse $\nabla^2 f(\theta_k)^{-1}$. There are several approaches for approximating H_k leading to different quasi-Newton methods, (Conn et al., 1991; Broyden, 1965; Broyden et al., 1973; Gay, 1979; Davidon, 1959; Fletcher & Powell, 1963; Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970; Nocedal, 1980; Liu & Nocedal, 1989), but in this paper, we focus on the celebrated BFGS method, in which H_k is updated as

$$H_k = \left(I - \frac{s_{k-1}u_{k-1}^\top}{s_{k-1}^\top u_{k-1}}\right) H_{k-1} \left(I - \frac{u_{k-1}s_{k-1}^\top}{s_{k-1}^\top u_{k-1}}\right) + \frac{s_{k-1}s_{k-1}^\top}{s_{k-1}^\top u_{k-1}}, \quad \forall k \geq 1, \tag{4}$$

where the variable variation s_{k-1} and gradient displacement u_{k-1} are defined as

$$s_{k-1} := \theta_k - \theta_{k-1}, \quad u_{k-1} := \nabla f(\theta_k) - \nabla f(\theta_{k-1}), \quad \forall k \geq 1, \tag{5}$$

respectively. The logic behind the update (4) is to ensure that the Hessian inverse approximation H_k satisfies the secant condition $H_k u_{k-1} = s_{k-1}$, while it stays close to the previous approximation matrix H_{k-1}. The update in (4) only requires matrix-vector multiplications, and hence, the computational cost per iteration of BFGS is $O(d^2)$.

The main advantage of BFGS is its fast superlinear convergence rate under the assumption of strict convexity, i.e., $\lim_{k\to 1} \|\theta_k - \theta_{opt}\| / \|\theta_{k-1} - \theta_{opt}\| = 0$, where θ_{opt} is the optimal solution. Previous results on the superlinear convergence of quasi-Newton methods were all asymptotic, until the recent advancements on the non-asymptotic analysis of quasi-Newton methods (Rodomanov & Nesterov, 2021a;b;c; Jin & Mokhtari, 2020; Jin et al., 2022; Ye et al., 2021; Lin et al., 2021a;b). For instance, (Jin & Mokhtari, 2020) established a local superlinear convergence rate of $(1/\sqrt{k})^k$ for BFGS. However, all these superlinear convergence analyses require the objective function to be smooth and strictly or strongly convex. Alas, these conditions are not satisfied in the low SNR setting, since the Hessian at the optimal solution could be singular, and hence the loss function is neither strongly convex nor strictly convex; we further discuss this point in Section 3. This observation implies that we need novel convergence analyses to study the behavior of BFGS in the low SNR setting.

3. Generalized Linear Model with Polynomial Link Function

In this section, we formally present the generalized linear model (GLM) setting that we consider in our paper, and discuss the low and high SNR settings and optimization challenges corresponding to these cases. Consider the case that the feature vectors are denoted by $X \in \mathbb{R}^d$ and their corresponding labels are denoted by $Y \in \mathbb{R}$. Suppose that we have access to n sample points $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ that are i.i.d. samples from the following generalized linear model with polynomial link function of power p (Carroll et al., 1997), i.e.,

$$Y_i = (X_i^\top \theta^*)^p + \zeta_i, \tag{6}$$

where θ^* is a true but unknown parameter, $p \in \mathbb{N}$ is a given power, and $\zeta_1, \ldots, \zeta_n$ are independent Gaussian noises with zero mean and variance σ^2. The Gaussian assumption on the noise is for the simplicity of the discussion and similar results hold for the sub-Gaussian i.i.d. noise case. Furthermore, we assume the feature vectors are generated as $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix. Here we focus on the settings that $p \in \mathbb{N}^+$ and $p \geq 2$.

The above class of GLMs with polynomial link functions arise in several settings. When $p = 1$, the model in (6) is the standard linear regression model, and for the case that $p = 2$, the above setup corresponds to the phase retrieval model (Fienup, 1982; Shechtman et al., 2015; Candes et al., 2011; Netrapalli et al., 2015), which has found applications in optical imaging, x-ray tomography, and audio signal processing. Moreover, the analysis of GLMs with $p \geq 2$ also serves as the basis of the analysis on other popular statistical models. For example, as shown by Yi & Caramanis (2015); Wang et al. (2015); Xu et al. (2016); Balakrishnan et al. (2017); Daskalakis et al. (2017); Dwivedi et al. (2020a); Kwon et al. (2019); Dwivedi et al. (2020b); Kwon et al. (2021), the local landscape of log-likelihood for Gaussian mixture models and mixture linear regression models are identical to GLMs for $p = 2$.

In the case that the polynomial link function parameter in the GLM model in (6) is $p = 2$, by adapting similar arguments from (Kwon et al., 2021) under the symmetric two-component location Gaussian mixture, there are essentially three regimes for estimating the true parameter: Low signal-to-noise ratio (SNR) regime: $\|\theta^*\|/\sigma \leq C_1 (d/n)^{1/4}$ where d is the dimension, n is the sample size, and C_1 is a universal constant; Middle SNR regime: $C_1 (d/n)^{1/4} \leq \|\theta^*\|/\sigma \leq C$ where C is a universal constant; High SNR regime: $\|\theta^*\|/\sigma \geq C$. The main idea is that with different θ^*, the optimization landscape of the parameter estimation problem changes. By generalizing the insights from the case $p = 2$, we define the following regimes for any $p \geq 2$:

- (i) Low SNR regime: $\|\theta^*\|/\sigma \leq C_1 (d/n)^{1/(2p)}$ where d is the dimension, n is the sample size, and C_1 is a universal constant;
- (ii) Middle SNR regime: $C_1 (d/n)^{1/(2p)} \leq \|\theta^*\|/\sigma \leq$

C where C is a universal constant;

- (iii) High SNR regime: $\|\theta^*\| / \sigma \geq C$.

Note that, the rate $(d/n)^{1/2p}$ that we use to define the SNR regimes is from the statistical rate of estimating the true parameter θ^* when θ^* approaches 0. Next, we provide insight into the landscape of the least-square loss function for each regime. In particular, given the GLM in (6), the sample least-square takes the following form:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (X_i^\top \theta)^p\right)^2. \tag{7}$$

To obtain insight about the landscape of the loss function $\mathcal{L}_n$, a useful approach is to consider an approximation of that function by its population version, which we refer to as population least-square loss function and is given by:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) := \mathbb{E}[\mathcal{L}_n(\theta)], \tag{8}$$

where the outer expectation is taken with respect to the data.

High SNR regime. In the setting that the ground truth parameter has a relatively large norm, i.e., $\|\theta^*\| \geq C$ for some constant $C > 0$ that only depends on σ, the population loss in (8) is locally strongly convex and smooth around θ^*. More precisely, when $\|\theta - \theta^*\|$ is small, we have

$$\begin{aligned}(X^\top \theta)^p - (X^\top \theta^*)^p \\ = p(X^\top \theta^*)^{p-1} X^\top(\theta - \theta^*) + o(\|\theta - \theta^*\|).\end{aligned}$$

Hence, in a neighborhood of the optimal solution, the objective in (8) can be approximated as

$$\begin{aligned}\mathcal{L}(\theta) = p^2(\theta - \theta^*)^\top \mathbb{E}_X\left[X(X^\top \theta^*)^{2p-2}X^\top\right](\theta - \theta^*) \\ + \sigma^2 + o(\|\theta - \theta^*\|^2).\end{aligned}$$

Indeed, if $\|\theta^*\| \geq C$ the above function behaves as a quadratic function that is smooth and strongly convex, assuming that $o(\|\theta - \theta^*\|^2)$ is negligible. As a result, the iterates of gradient descent (GD) converge to the solution at a linear rate and it requires $\kappa \log(1/\epsilon)$ to reach an ϵ-accurate solution, where κ depends on the conditioning of the covariance matrix Σ and the norm of θ^*. In this case, BFGS converges superlinearly to θ^* and the rate would be independent of κ, while the cost per iteration would be $O(d^2)$.

Low SNR regime. As mentioned above, in the high SNR case, GD has a fast linear rate. However, in the low SNR case where $\|\theta^*\|$ is small and $\|\theta^*\| \leq C_1(d/n)^{1/(2p)}$, the strong convexity parameter approaches zero when the sample size n goes to infinity and the problem becomes ill-conditioned. In this case, we deal with a function that is only convex and its gradient is not Lipschitz continuous. To better elaborate on this point, let us focus on the case that $\theta^* = 0$ as a special case of the low SNR setting. Considering the underlying distribution of X, which is $X \sim \mathcal{N}(0, \Sigma)$, for such a low SNR case, the population loss can be written as

$$\mathcal{L}(\theta) = \mathbb{E}_X\left[(X^\top \theta)^{2p}\right] + \sigma^2 = (2p-1)!!\|\Sigma^{1/2}\theta\|^{2p} + \sigma^2. \tag{9}$$

Since we focus on $p \geq 2$ it can be verified that $\mathcal{L}(\theta)$ is not strongly convex in a neighborhood of the solution $\theta^* = 0$. For this class of functions, it is well-known that GD with constant step size would converge at a sublinear rate, and hence GD iterates require polynomial number of iterations to reach the final statistical radius. In the next section, we study the behavior of BFGS for solving the low SNR setting and showcase its advantage compared to GD.

Middle SNR regime. Different from the low and high SNR regimes, the middle SNR regime is generally harder to analyze as the landscapes of both the population and sample least-square loss functions are complex. Adapting the insight from middle SNR regime of the symmetric two-component location Gaussian mixtures from (Kwon et al., 2021), for the middle SNR setting of the generalized linear model, the eigenvalues of the Hessian matrix of the population least-square loss function approach 0 and their vanishing rates depend on some increasing function of $\|\theta^*\|$. The optimal statistical rate of the optimization algorithms, such as gradient descent algorithm, for solving θ^* depends strictly on the tightness of these vanishing rates in terms of $\|\theta^*\|$, which are non-trivial to obtain. In fact, to the best of our knowledge, there is no result on the convergence of iterative methods (such as GD or its variants) for GLMs with a polynomial link function in the middle SNR. Hence, we leave the study of BFGS for this setting as a future work.

4. Convergence Analysis in the Low SNR Regime: Population Loss

In this section, we focus on the convergence properties of BFGS for the population loss in the low SNR case introduced in (9). This analysis provides an intuition for the analysis of the finite sample case that we discuss in Section 5, as we expect these two loss functions to be close to each other when the number of samples n is sufficiently large. Note that the loss function in (9) can be considered as a special case of the following convex optimization problem:

$$\min_{\theta \in \mathbb{R}^d} f(\theta) = \|A\theta - b\|^q, \tag{10}$$

where $A \in \mathbb{R}^{m \times d}$ is a matrix, $b \in \mathbb{R}^m$ is a given vector, and q satisfies $q \geq 4$. We should note that for $q \geq 4$, the considered objective is not strictly convex because the Hessian is singular when $A\theta = b$. Indeed, if we set $m = d$ and further let A be $\Sigma^{1/2}$ and choose $b = A\theta^* = 0$, then we recover the problem in (9) for $q = 2p$.

Notice that the problem in (10), which serves as a surrogate for the finite sample problem that we plan to study in the next section, has the same solution set as the quadratic problem of minimizing $\|A\theta - b\|^2$ with solution $(A^\top A)^{-1}A^\top b$ when $A^\top A$ is invertible. Given this point, one might suggest that instead of minimizing (10) we could directly solve the quadratic problem which is indeed much easier to solve. This point is valid, but the goal of this section is not to efficiently solve problem (10) itself. Our goal is to understand the convergence properties of the BFGS method for solving the problem in (10) with the hope that it will provide some intuitions for the convergence analysis of BFGS for the empirical loss (7) in the low SNR regime. As we will discuss in Section 5, the convergence analysis of BFGS on the population loss which is a special case of (10) is closely related to the one for the empirical loss in (7).

One remark is that population loss (9) holds for the restrictive assumption of $\theta^* = 0$, which is only a special case of the general low SNR regime of $\|\theta^*\| \le C_1(d/n)^{1/(2p)}$. Our ultimate goal is to analyze the convergence behavior of the BFGS method applied to the empirical loss (7) of GLM problems. The errors between gradients and Hessians of the population loss with $\theta^* = 0$ and $\|\theta^*\| \le C_1(d/n)^{1/(2p)}$ are upper bounded by the corresponding statistical errors between the population loss (8) and the empirical loss (7) in the low SNR regime, respectively. Therefore, the errors between iterations of applying BFGS to the population loss with $\theta^* = 0$ and $\|\theta^*\| \le C_1(d/n)^{1/(2p)}$ are negligible compared to the statistical errors. Instead of directly analyzing BFGS for the population loss (8) in the general low SNR regime, studying the convergence properties of BFGS for the population loss (9) with $\theta^* = 0$ can equivalently lay foundations for the convergence analysis of BFGS for the empirical loss (7) in the low SNR regime, which is the main target of this paper. The details can be found in the proof of Theorem 5.1.

Moreover, our results would be of general interest from an optimization point of view, since there is no global convergence theory (without line-search) for the BFGS method in the literature for the case that the objective function is not strictly convex, and our analysis provides the first result for such general settings. Before, stating our results, we first state the assumptions.

Assumption 4.1. There exists $\hat{\theta} \in \mathbb{R}^d$, such that $b = A\hat{\theta}$. In other words, b is in the range of matrix A.

This assumption implies that the problem in (10) is realizable, $\hat{\theta}$ is an optimal solution, and the optimal function value is zero. This assumption is satisfied in our considered low SNR setting in (9) as $\theta^* = 0$ which implies $b = 0$.

Assumption 4.2. The matrix $A^\top A \in \mathbb{R}^{d \times d}$ is invertible. This is equivalent to $A^\top A \succ 0$.

The above assumption is also satisfied for our considered setting as we assume that the covariance matrix for our input features is positive definite. Combining Assumptions 4.1 and 4.2, we conclude that $\hat{\theta}$ is the unique optimal solution of problem (10). Next, we state the convergence rate of BFGS for solving problem 10 under the disclosed assumptions.

Theorem 4.3. *Consider the BFGS method in (3)-(5). Suppose Assumptions 4.1 and 4.2 hold, and the initial Hessian inverse approximation matrix is selected as $H_0 = \nabla^2 f(\theta_0)^{-1}$, where $\theta_0 \in \mathbb{R}^d$ is an arbitrary initial point. If the step size is $\eta_k = 1$ for all $k \ge 0$, then the iterates of BFGS converge to the optimal solution $\hat{\theta}$ at a linear rate of*

$$\|\theta_k - \hat{\theta}\| = r_{k-1}\|\theta_{k-1} - \hat{\theta}\|, \quad \forall k \ge 1, \tag{11}$$

where the contraction factors $r_k \in [0, 1)$ satisfy

$$r_0 = \frac{q-2}{q-1}, \quad r_k = \frac{1 - r_{k-1}^{q-2}}{1 - r_{k-1}^{q-1}}, \quad \forall k \ge 1. \tag{12}$$

The proof of this theorem is available in Appendix A.1. Theorem 4.3 shows that the iterates of BFGS converge globally at a linear rate to the optimal solution of (10). This result is of interest as it illustrates the iterates generated by BFGS converge globally without any line search scheme and the stepsize is fixed as $\eta_k = 1$ for any $k \ge 0$. Moreover, the initial point θ_0 is arbitrary and there is no restriction on the distance between θ_0 and the optimal solution $\hat{\theta}$. This result is in contrast to most analyses of quasi-Newton methods which require the initial point θ_0 to be in a local neighborhood of $\hat{\theta}$ to guarantee the linear or superlinear convergence rate, without line-search.

Remark 4.4. The cost of computing the initial Hessian inverse approximation $H_0 = \nabla^2 f(\theta_0)^{-1}$ is $O(d^3)$, but this cost is only required for the first iteration, and it is not required for $k \ge 1$ as for those iterates we update the Hessian inverse approximation matrix H_k based on the update in (4) at a cost of $O(d^2)$.

Note that the result in Theorem 4.3 does not specify the exact complexity of BFGS for solving problem (10), as the contraction factors r_k are not explicitly given. In the following theorem, we show that for $q \ge 4$, the linear rate contraction factors $\{r_k\}_{k=0}^{\infty}$ also converge linearly to a fixed point contraction factor r^* determined by the parameter q.

Theorem 4.5. *Consider the linear convergence factors $\{r_k\}_{k=0}^{\infty}$ defined in (12) from Theorem 4.3. If $q \ge 4$, then the sequence $\{r_k\}_{k=0}^{\infty}$ converges to $r^* \in (0, 1)$ that is determined by the equation*

$$r_*^{q-1} + r_*^{q-2} = 1, \tag{13}$$

and the rate of convergence is linear with a contraction factor that is at most 1/2, i.e.,

$$|r_k - r_*| \le (1/2)^k |r_0 - r_*|, \quad \forall k \ge 0. \tag{14}$$

The proof is in Appendix A.2. Based on Theorem 4.5, the iterates of BFGS eventually converge to the optimal solution at the linear rate of r^* determined by (13). Specifically, the factors $\{r_k\}_{k=0}^{\infty}$ converge to the fixed point r^* at a linear rate with the contraction factor of $1/2$. Further, the linear convergence factors $\{r_k\}_{k=0}^{\infty}$ and their limit r^* are all only determined by q, and they are independent of the dimension d and the condition number κ_A of the matrix A. Hence, the performance of BFGS is not influenced by high-dimensional or ill-conditioned problems. This result is independently important from an optimization point of view, as it provides the first global linear convergence of BFGS without line-search for a setting that is not strictly or strongly convex, and interestingly the constant of linear convergence is independent of dimension or condition number.

We illustrate the convergence of factors $\{r_k\}_{k=0}^{\infty}$ to the fixed point r^* in Figure 4 of the appendix B for $q = 4$ and $q = 100$. In plots (a) and (b), we observe that r_k becomes close to r^* after only 5 iterations. Hence, the linear convergence rate of BFGS is approximately r^* after only a few iterations. We further observe in plots (c) and (d) that the factors $\{r_k\}_{k=0}^{\infty}$ converge to the fixed point r^* at a linear rate upper bounded by $1/2$. Note that r^* is the solution of (13). These plots verify our results in Theorem 4.5.

4.1. Comparison with Newton's Method

Next, we compare the convergence results of BFGS for solving problem (10) with the one for Newton's method. The following theorem characterizes the global linear convergence of Newton's method with unit step size applied to the objective function in (10).

Theorem 4.6. Consider applying Newton's method to optimization problem (10) and suppose Assumptions 4.1 and 4.2 hold. Moreover, suppose the step size is $\eta_k = 1$ for any $k \geq 0$. Then, the iterates of Newton's method converge to the optimal solution $\hat{\theta}$ at a linear rate of

$$\|\theta_k - \hat{\theta}\| = \frac{q-2}{q-1}\|\theta_{k-1} - \hat{\theta}\|, \quad \forall k \geq 1. \tag{15}$$

Moreover, this linear convergence rate $\frac{q-2}{q-1}$ is smaller than the fixed point r^* defined in (13) of the BFGS quasi-Newton method, i.e., $\frac{q-2}{q-1} < r^* < \frac{2q-3}{2q-2}$ for all $q \geq 4$.

The proof is available in Appendix A.3. The convergence results of Newton's method are also global without any line search method, and the linear rate $\frac{q-2}{q-1}$ is independent of dimension d and condition number κ_A. Furthermore, the condition $\frac{q-2}{q-1} < r^*$ implies that iterates of Newton's method converge faster than BFGS, but the gap is not substantial as $r^* < \frac{2q-3}{2q-2}$. On the other hand, the computational cost per iteration of Newton's method is $O(d^3)$ which is worse than the $O(d^2)$ of BFGS.

Moving back to our main problem, one important implication of the above convergence results is that in the low SNr setting the iterates of BFGS converge linearly to the optimal solution of the population loss function, while the contraction coefficient of BFGS is comparable to that of Newton's method which is $(2p-2)/(2p-1)$. For instance, for $p = 2, 3, 5, 10$, the linear rate contraction factor of Newton's method are $0.667, 0.8, 0.889, 0.947$ and the approximate linear rate contraction factor of BFGS denoted by r^* are $0.755, 0.857, 0.922, 0.963$, respectively.

5. Convergence Analysis in the Low SNR Regime: Finite Sample Setting

Thus far, we have demonstrated that the BFGS iterates converge linearly to the true parameter when minimizing the population loss function $\mathcal{L}$ of the GLM in (9) in the low SNR regime. In this section, we study the statistical behavior of the BFGS iterates for the finite sample case by leveraging the insights developed in the previous section about the convergence rate of BFGS in the infinite sample case, i.e., population loss. More precisely, we focus on the application of BFGS for solving the least-square loss function $\mathcal{L}_n$ defined in (7) for the low SNR setting. The iterates of BFGS in this case follow the update rule

$$\theta^n_{k+1} = \theta^n_k - \eta_k H^n_k \nabla \mathcal{L}_n(\theta^n_k), \tag{16}$$

where H^n_k is updated using the gradient information of the loss $\mathcal{L}_n$ by the BFGS rule.

We next show that the BFGS iterates (16) $\{\theta^n_k\}_{k \geq 0}$ converge to the final statistical radius within a logarithmic number of iterations under the low SNR regime of the GLMs. To prove this claim, we track the difference between the iterates $\{\theta^n_k\}_{k \geq 0}$ generated based on the empirical loss and the iterates $\{\theta_k\}_{k \geq 0}$ generated according to the population loss. Assuming that they both start from the same initialization θ_0, with the concentration of the gradient $\|\nabla \mathcal{L}_n(\theta) - \nabla \mathcal{L}(\theta)\|$ and the Hessian $\|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{op}$ from Mou et al. (2019); Ren et al. (2022b) we control the deviation between these two sequences. Using this bound and the convergence results of the iterates generated based on the population loss discussed in the previous section, we prove the following result for the finite sample setting.

Theorem 5.1. Consider the low SNR regime of the GLM in (6) namely, $\|\theta^*\| \leq C_1 (d/n)^{1/(2p)}$. Apply the BFGS method to the empirical loss (7) with the initial Hessian inverse approximation matrix as

$$H_0 = \nabla^2 f(\theta^n_0)^{-1},$$

where $\theta^n_0 \in \mathbb{B}(\theta^*, r)$ for some $r > 0$ and step size $\eta_k = 1$. For any failure probability $\delta \in (0, 1)$, if the number of samples is $n \geq C_2 (d \log(d/\delta))^{2p}$, and the number of iterations satisfies $T \geq \log(n/d)(\log(1/\delta))$, then with probability $1-\delta$, we have

$$\min_{t\in[T]} \|\theta_t^n - \theta^*\| \leq C_3 \left(\frac{d\log(1/\delta)}{n}\right)^{\frac{1}{2p+2}}, \qquad (17)$$

where C_1, C_2, and C_3 are universal constants independent of n and d, and C_3 has a polynomial dependency on p.

The proof is available in Appendix A.4. Theorem 5.1 shows that BFGS achieves the statistical accuracy $\mathcal{O}(\log n)$ iterations, which is faster than the sublinear convergence $\mathcal{O}(n^{\frac{p-1}{p}})$ of GD shown in (Ren et al., 2022a). A few comments about Theorem 5.1 are in order.

Comparing to GD, GD with Polyak step size , and Newton's method: Theorem 5.1 indicates that under the low SNR regime, the BFGS iterates reach the final statistical radius $\mathcal{O}(n^{-1/(2p+2)})$ within the true parameter θ^* after $\mathcal{O}(\log(n))$ number of iterations. The statistical radius is slightly worse than the optimal statistical radius $\mathcal{O}(n^{-1/(2p)})$. However, we conjecture that this is due to the proof technique and BFGS can still reach the optimal $\mathcal{O}(n^{-1/(2p)})$ in practice. In our experiments, in the next section, we observe that when $d = 4$ and $p = 2$, the statistical radius of BFGS is closer to the optimal radius of $\mathcal{O}(n^{-1/4})$ instead of $\mathcal{O}(n^{-1/6})$ suggested by our analysis. We leave an improvement of the statistical analysis as the future work. On the other hand, the overall computational complexity of BFGS, which is $\mathcal{O}(\log(n))$, is indeed better than the polynomial number of iterations of GD, which is at the order of $\mathcal{O}(n^{(p-1)/p})$ (Corollary 3 in (Ho et al., 2020)).

Moreover, the complexity of BFGS is better than the one for GD with Polyak step size which is $\mathcal{O}(\kappa \log(n))$ iterations (Corollary 1 in (Ren et al., 2022a)), where κ is the condition number of the covariance matrix. Note that while the iteration complexity of BFGS is comparable to that of GD with Polyak step size in terms of the sample size, the BFGS overcomes the need to approximate the optimal value of the sample least-square loss $\mathcal{L}_n$, which can be unstable in practice, and also removes the dependency on the condition number that appears in the complexity bound of GD with Polyak step size. Finally, the iteration complexity of BFGS is comparable to the $\mathcal{O}(\log(n))$ of Newton's method (Corollary 3 in (Ho et al., 2020)), while per iteration cost of BFGS is lower than Newton's method.

On the minimum number of iterations: The results in Theorem 5.1 involve the minimum number of iterations, namely, this result holds for some $1 \leq t \leq T$. It suggests that the BFGS iterates may diverge after they reach the final statistical radius. As highlighted in (Ho et al., 2020), such instability behavior of BFGS is inherent to fast and unstable methods. While it may sound limited, this issue can be handled via an early stopping scheme using the cross-validation approaches. We illustrate such early stopping of the BFGS iterates for the low SNR regime in Figure 2.

6. Numerical Experiments

We divide our experiments into two sections where the first one focuses on the behavior of iterative methods on the population loss of GLM with polynomial link functions and the second one focuses on the finite sample setting.

Experiments for the population loss function. In this section, we compare the performance of Newton's method, BFGS, GD with constant step size, and GD with Polyak step size applied to (10) which corresponds to the population loss. We choose different values of parameter m, dimension d and the exponential parameter q in (10). We generate a random matrix $A \in \mathbb{R}^{m \times d}$ and a random vector $\hat{\theta} \in \mathbb{R}^d$, and compute the vector $b = A\hat{\theta} \in \mathbb{R}^d$. The initial point $\theta_0 \in \mathbb{R}^d$ is also generated randomly. The GD constant step size η is tuned by hand to achieve the best performance of GD on each problem. We present the logarithmic scale of $\|\theta_k - \hat{\theta}\|$ versus the number of iteration k for different algorithms. All the values of different parameters m, d, q and η as well as the numerical results of our experiments are shown in Figure 1.

We observe that GD with constant step converges slowly due to its sub-linear convergence rate. The performance of GD with Polyak step size is also poor when dimension is large or the parameter q is huge. This is due to the fact that as dimension increases the problem becomes more ill-conditioned and hence the linear convergence contraction factor approaches 1. We observe that both Newton's method and BFGS generate iterations with linear convergence rates, and their linear convergence rates are only affected by the parameter q, i.e., the dimension d has no impact over the performance of BFGS and Newton's method. Although the convergence speed of Newton's method is faster than BFGS, their gap is not significantly large as we expected from our theoretical results in Section 4.

Experiments for the empirical loss function. We next study the statistical and computational complexities of BFGS on the empirical loss. In our experiments, we first consider the case that $d = 4$ and the power of the link function is $p = 2$, namely, we consider the multivariate setting of the phase retrieval problem. The data is generated by first sampling the inputs according to $\{X_i\}_{i=1}^n \sim \mathcal{N}(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_4^2))$ where $\sigma_k = (0.5)^{k-1}$, and then generating their labels based on $Y_i = (X_i^\top \theta^*)^2 + \varepsilon_i$ where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. samples from $\mathcal{N}(0, 0.01)$. In the low SNR regime, we set $\theta^* = 0$, and in the high SNR regime we select θ^* uniformly at random from the unit sphere. Further, for GD, we set the step size as $\eta = 0.1$, while for Newton's method and BFGS, we use the unit stepsize $\eta = 1$.

(a) $d = 10; q = 4$. (b) $d = 10; q = 10$. (c) $d = 10^3; q = 4$. (d) $d = 10^3; q = 10$.

Figure 1. Convergence of Newton's method, BFGS, GD with constant step size and GD with Polyak step size for different values of d and q. In plot (a), $m = 100$ and $= 10^{4}$. In plot (b), $m = 100$ and $= 10^{8}$. In plot (c), $m = 2000$ and $= 10^{12}$. In plot (d), $m = 2000$ and $= 10^{15}$.

(a) High SNR regime. (b) Low SNR regime. (c) High SNR regime. (d) Low SNR regime.

Figure 2. Illustration of different methods with high SNR regime in (a) and low SNR regime in (b). Illustration of the statistical radius of BFGS with high SNR regime in (c) and low SNR regime in (d).

In plots (a) and (b) of Figure 2, we consider the setting that the sample size is $n = 10^4$, and we run GD, GD with Polyak step size, BFGS, and Newton's method to find the optimal solution of the sample least-square loss L_n. Furthermore, for both Newton's method and the BFGS algorithm, due to their instability, we also perform cross-validation to choose their early stopping. In particular, we split the data into training and the test sets. The training set consists of 90% of the data while the test set has 10% of the data. The yellow points in plots (a) and (b) of Figure 2 show the iterates of BFGS and Newton, respectively, with the minimum validation loss. As we observe, under the low SNR regime, the iterates of GD with Polyak step size, BFGS and Newton's method converge geometrically fast to the final statistical radius while those of the GD converge slowly to that radius. Under the high SNR regime, the iterates of all of these methods converge geometrically fast to the final statistical radius. The faster convergence of GD with Polyak step size over GD is due to the optimality of step size of, while the faster convergence of BFGS and Newton's method over GD is due to their independence on the problem condition number. Finally, in plots (c) and (d) of Figure 2, we run BFGS when the sample size is ranging from 10^2 to 10^4 to empirically verify the statistical radius of these methods. As indicated in the plots of that figure, under the high SNR regime, the BFGS has statistical radius is $O(n^{1/2})$, while under the low SNR regime, its statistical radius becomes $O(n^{1/4})$.

We present additional numerical experiments regarding the linear contraction factors, experiments on the empirical loss with high dimensions and empirical results for the middle SNR regimes in the appendix B.

7. Conclusions

In this paper, we analyzed the convergence rates of BFGS on both population and empirical loss functions of the generalized linear model in the low SNR regime. We showed that in this case, BFGS outperforms GD and performs similar to Newton's method in terms of iteration complexity, while it requires a lower per iteration computational complexity compared to Newton's method. We also provided experiments for both infinite and finite sample loss functions and showed that our empirical results are consistent with our theoretical findings. Perhaps one limitation of the BFGS method is that its computational cost is still not linear in the dimension and scales as $O(d^2)$. One future research direction is to analyze some other iterative methods such as limited memory-BFGS (L-BFGS) which may be able to achieve a fast linear convergence rate in the low SRN setting, while its computational cost per iteration is $O(d)$.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40 (5):2452–2482, 2012.

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.

Broyden, C. G. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19 (92):577–593, 1965.

Broyden, C. G. The convergence of single-rank quasi-Newton methods. *Mathematics of Computation*, 24(110): 365–382, 1970.

Broyden, C. G., Jr., J. E. D., Broyden, and More, J. J. On the local and superlinear convergence of quasi-Newton methods. *IMA J. Appl. Math*, 12(3):223–245, June 1973.

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

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92:477–489, 1997.

Conn, A. R., Gould, N. I. M., and Toint, P. L. Convergence of quasi-Newton matrices generated by the symmetric rank one update. *Mathematical programming*, 50(1-3): 177–195, 1991.

Daskalakis, C., Tzamos, C., and Zampetakis, M. Ten steps of EM suffice for mixtures of two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

Davidon, W. Variable metric method for minimization. Technical report, Argonne National Lab., Lemont, Ill., 1959.

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.

Feiyan Tian, Lei Liu, X. C. Generalized memory approximate message passing. *https://arxiv.org/abs/2110.06069*, 2021.

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.

Fletcher, R. A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322, 1970.

Fletcher, R. and Powell, M. J. A rapidly convergent descent method for minimization. *The computer journal*, 6(2): 163–168, 1963.

Gay, D. M. Some convergence properties of Broyden's method. *SIAM Journal on Numerical Analysis*, 16(4): 623–630, 1979.

Goebel, K. and Kirk, W. A. *Topics in Metric Fixed Point Theory*. Cambridge University Press, 1990.

Goldfarb, D. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24 (109):23–26, 1970.

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., 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. and Mokhtari, A. Non-asymptotic superlinear convergence of standard quasi-newton methods. *arXiv preprint arXiv:2003.13607*, 2020.

Jin, Q., Koppel, A., Rajawat, K., and Mokhtari, A. Sharpened quasi-newton methods: Faster superlinear rate and larger local convergence neighborhood. *The 39th International Conference on Machine Learning (ICML 2022)*, 2022.

Kwon, J., Qian, W., Caramanis, C., Chen, Y., , and Damek, D. Global convergence of the EM algorithm for mixtures of two component linear regression. In *Conference on Learning Theory (COLT)*, 2019.

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.

Lin, D., Ye, H., and Zhang, Z. Explicit superlinear convergence of broyden's method in nonlinear equations. arXiv preprint arXiv:2109.01974, 2021a.

Lin, D., Ye, H., and Zhang, Z. Greedy and random quasi-newton methods with faster explicit superlinear convergence. Advances in Neural Information Processing Systems 34, 2021b.

Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. Mathematical programming, 45(1-3):503–528, 1989.

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

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.

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.

Nocedal, J. Updating quasi-Newton matrices with limited storage. Mathematics of computation, 35(151):773–782, 1980.

Ren, T., Cui, F., Atsidakou, A., Sanghavi, S., and Ho, N. Towards statistical and computational complexities of Polyak step size gradient descent. Artificial Intelligence and Statistics Conference, 2022a.

Ren, T., Zhuo, J., Sanghavi, S., and Ho, N. Improving computational complexity in statistical models with second-order information. arXiv preprint arXiv:2202.04219, 2022b.

Rodomanov, A. and Nesterov, Y. Greedy quasi-newton methods with explicit superlinear convergence. SIAM Journal on Optimization, 31(1):785–811, 2021a.

Rodomanov, A. and Nesterov, Y. Rates of superlinear convergence for classical quasi-newton methods. Mathematical Programming, pp. 1–32, 2021b.

Rodomanov, A. and Nesterov, Y. New results on superlinear convergence of classical quasi-newton methods. Journal of Optimization Theory and Applications, 188(3):744–769, 2021c.

Shanno, D. F. Conditioning of quasi-Newton methods for function minimization. Mathematics of computation, 24 (111):647–656, 1970.

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.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. High-dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. In Advances in Neural Information Processing Systems, 28, 2015.

Xu, J., Hsu, D., and Maleki, A. Global analysis of expectation maximization for mixtures of two Gaussians. In Advances in Neural Information Processing Systems, 29, 2016.

Ye, H., Lin, D., Zhang, Z., and Chang, X. Explicit superlinear convergence rates of the sr1 algorithm. arXiv preprint arXiv:2105.07162, 2021.

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

Yuan, X.-T. and Zhang, T. Truncated power method for sparse eigenvalue problems. Journal of Machine Learning Research, 14(Apr):899–925, 2013.

A. Proofs

Lemma A.1. *Consider the objective function* (10) *satisfying Assumption* 4.1 *and* 4.2*. Then, the inverse matrix of its Hessian $\nabla^2 f(\theta)$ can be expressed as*

$$\nabla^2 f(\theta)^{-1} = \frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}} - \frac{(q-2)(\theta-\hat{\theta})(\theta-\hat{\theta})^\top}{q(q-1)\|A\theta - b\|^{q}}. \tag{18}$$

Proof. Notice that the Hessian of objective function (10) can be expressed as

$$\nabla^2 f(\theta) = q\|A\theta - b\|^{q-2}A^\top A + q(q-2)\|A\theta - b\|^{q-4}A^\top(A\theta - b)(A\theta - b)^\top A. \tag{19}$$

We use the Sherman–Morrison formula. Suppose that $X \in \mathbb{R}^{d\times d}$ is an invertible matrix and $a, b \in \mathbb{R}^d$ are two vectors satisfying that $1 + b^\top X^{-1}a \neq 0$. Then, the matrix $X + ab^\top$ is invertible and

$$(X + ab^\top)^{-1} = X^{-1} - \frac{X^{-1}ab^\top X^{-1}}{1 + b^\top X^{-1}a}. \tag{20}$$

Applying the Sherman–Morrison formula with $X = q\|A\theta - b\|^{q-2}A^\top A$, $a = q(q-2)\|A\theta - b\|^{q-4}A^\top(A\theta - b)$ and $b = A^\top(A\theta - b)$. Notice that $A^\top A$ is invertible, hence X is invertible and

$$\begin{aligned}
& 1 + b^\top X^{-1}a \\
= \;& 1 + (A\theta - b)^\top A\frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}}q(q-2)\|A\theta - b\|^{q-4}A^\top(A\theta - b) \\
= \;& 1 + (q-2)(A\theta - b)^\top A\frac{(A^\top A)^{-1}A^\top A(\theta - \hat{\theta})}{\|A\theta - b\|^2} \\
= \;& 1 + (q-2)\frac{(A\theta - b)^\top(A\theta - b)}{\|A\theta - b\|^2} \\
= \;& q - 1 \neq 0. \qquad (\text{Since } q \geq 4.)
\end{aligned} \tag{21}$$

Therefore, we obtain that

$$\begin{aligned}
& \nabla^2 f(\theta)^{-1} \\
= \;& \frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}} - \frac{\frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}}q(q-2)\|A\theta - b\|^{q-4}A^\top(A\theta - b)(A^\top(A\theta - b))^\top\frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}}}{q-1} \\
= \;& \frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}} - \frac{(q-2)}{q(q-1)\|A\theta - b\|^{q}}(A^\top A)^{-1}AA^\top(\theta - \hat{\theta})(\theta - \hat{\theta})^\top AA^\top(A^\top A)^{-1} \\
= \;& \frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}} - \frac{(q-2)(\theta-\hat{\theta})(\theta-\hat{\theta})^\top}{q(q-1)\|A\theta - b\|^{q}}.
\end{aligned} \tag{22}$$

As a consequence, we obtain the conclusion of the lemma. □

Lemma A.2. *Banach's Fixed-Point Theorem. Consider the differentiable function $f : D \subseteq \mathbb{R} \to D \subseteq \mathbb{R}$. Suppose that there exists $C \in (0, 1)$ such that $|f'(x)| \leq C$ for any $x \in D$. Now let $x_0 \in D$ be arbitrary and define the sequence $\{x_k\}_{k=0}^{\infty}$ as*

$$x_{k+1} = f(x_k), \quad \forall k \geq 0. \tag{23}$$

Then, the sequence $\{x_k\}_{k=0}^{\infty}$ converges to the unique fixed point x^* defined as

$$x^* = f(x^*), \tag{24}$$

with linear convergence rate of

$$|x_k - x^*| \leq C^k|x_0 - x^*|, \quad \forall k \geq 0. \tag{25}$$

Proof. Check (Goebel & Kirk, 1990). □

A.1. Proof of Theorem 4.3

We use induction to prove the convergence results (11) and (12). Note that $b = A\hat{\theta}$ by Assumption 4.1 and the gradient and Hessian of the objective function in (10) are explicitly given by

$$\nabla f(\theta) = q\|A\theta - b\|^{q-2}A^\top(A\theta - b) = q\|A\theta - b\|^{q-2}A^\top A(\theta - \hat{\theta}); \tag{26}$$

$$\nabla^2 f(\theta) = q\|A\theta - b\|^{q-2}A^\top A + q(q-2)\|A\theta - b\|^{q-4}A^\top(A\theta - b)(A\theta - b)^\top A. \tag{27}$$

Applying Lemma A.1, we can obtain that

$$\nabla^2 f(\theta)^{-1} = \frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}} - \frac{(q-2)(\theta - \hat{\theta})(\theta - \hat{\theta})^\top}{q(q-1)\|A\theta - b\|^{q}}. \tag{28}$$

First, we consider the initial iteration

$$\theta_1 = \theta_0 - H_0\nabla f(\theta_0) = \theta_0 - \nabla^2 f(\theta_0)^{-1}\nabla f(\theta_0); \tag{29}$$

$$\theta_1 - \hat{\theta} = \theta_0 - \hat{\theta} - \nabla^2 f(\theta_0)^{-1}\nabla f(\theta_0). \tag{30}$$

Notice that $b = A\hat{\theta}$ by Assumption 4.1 and

$$\begin{aligned} &\nabla^2 f(\theta_0)^{-1}\nabla f(\theta_0) \\ &= \left[\frac{(A^\top A)^{-1}}{q\|A\theta_0 - b\|^{q-2}} - \frac{(q-2)(\theta_0 - \hat{\theta})(\theta_0 - \hat{\theta})^\top}{q(q-1)\|A\theta_0 - b\|^{q}}\right] q\|A\theta_0 - b\|^{q-2}A^\top A(\theta_0 - \hat{\theta}) \\ &= \theta_0 - \hat{\theta} - \frac{q-2}{q-1}\frac{(\theta_0 - \hat{\theta})^\top A^\top A(\theta_0 - \hat{\theta})}{\|A\theta_0 - b\|^2}(\theta_0 - \hat{\theta}) \\ &= \theta_0 - \hat{\theta} - \frac{q-2}{q-1}\frac{(A\theta_0 - b)^\top(A\theta_0 - b)}{\|A\theta_0 - b\|^2}(\theta_0 - \hat{\theta}) \\ &= \theta_0 - \hat{\theta} - \frac{q-2}{q-1}(\theta_0 - \hat{\theta}). \end{aligned} \tag{31}$$

Therefore, we obtain that

$$\theta_1 - \hat{\theta} = \theta_0 - \hat{\theta} - \nabla^2 f(\theta_0)^{-1}\nabla f(\theta_0) = \frac{q-2}{q-1}(\theta_0 - \hat{\theta}). \tag{32}$$

Condition (11) holds for $k = 1$ with $r_0 = \frac{q-2}{q-1}$. Now we assume that condition (11) holds for $k = t$ where $t \geq 1$, i.e.,

$$\theta_t - \hat{\theta} = r_{t-1}(\theta_{t-1} - \hat{\theta}). \tag{33}$$

Considering the condition $b = A\hat{\theta}$ in Assumption 4.1 and the condition in (33), we further have

$$A\theta_t - b = A(\theta_t - \hat{\theta}) = r_{t-1}A(\theta_{t-1} - \hat{\theta}) = r_{t-1}(A\theta_{t-1} - b); \tag{34}$$

which implies that

$$\nabla f(\theta_t) = q r_{t-1}^{q-1}\|A(\theta_{t-1} - \hat{\theta})\|^{q-2}A^\top A(\theta_{t-1} - \hat{\theta}). \tag{35}$$

We further show that the variable displacement and gradient difference can be written as

$$s_{t-1} = \theta_t - \theta_{t-1} = \theta_t - \hat{\theta} - \theta_{t-1} + \hat{\theta} = (r_{t-1} - 1)(\theta_{t-1} - \hat{\theta}); \tag{36}$$

and

$$u_{t-1} = \nabla f(\theta_t) - \nabla f(\theta_{t-1}) = q(r_{t-1}^{q-1} - 1)\|A(\theta_{t-1} - \hat{\theta})\|^{q-2}A^\top A(\theta_{t-1} - \hat{\theta}). \tag{37}$$

Considering these expressions, we can show that the rank-1 matrix in the update of BFGS $u_{t-1}s_{t-1}^\top$ is given by

$$u_{t-1}s_{t-1}^\top = q(r_{t-1}^{q-1} - 1)(r_{t-1} - 1)\|A(\theta_{t-1} - \hat{\theta})\|^{q-2}A^\top A(\theta_{t-1} - \hat{\theta})(\theta_{t-1} - \hat{\theta})^\top; \tag{38}$$

and the inner product $s_{t-1}^\top u_{t-1}$ can be written as

$$\begin{aligned} s_{t-1}^\top u_{t-1} &= q(r_{t-1}^{q-1} - 1)(r_{t-1} - 1)\|A(\theta_{t-1} - \hat{\theta})\|^{q-2}(\theta_{t-1} - \hat{\theta})^\top A^\top A(\theta_{t-1} - \hat{\theta}) \\ &= q(r_{t-1}^{q-1} - 1)(r_{t-1} - 1)\|A(\theta_{t-1} - \hat{\theta})\|^{q}. \end{aligned} \tag{39}$$

These two expressions allow us to simplify the matrix $\frac{u_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}$ in the update of BFGS as

$$I - \frac{u_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}} = I - \frac{A^\top A(\theta_{t-1} - \hat{\theta})(\theta_{t-1} - \hat{\theta})^\top}{\|A(\theta_{t-1} - \hat{\theta})\|^2}. \tag{40}$$

An important property of the above matrix is that its null space is the set of the vectors that are parallel to u_{t-1}. Considering the expression for u_{t-1}, any vector that is parallel to $A^\top A(\theta_{t-1} - \hat{\theta})$ is in the null space of the above matrix. We can easily observe that the gradient defined in (35) satisfies this condition and therefore

$$\begin{aligned} &\left(I - \frac{u_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\right)\nabla f(\theta_t) \\ &= q r_{t-1}^{q-1}\|A(\theta_{t-1} - \hat{\theta})\|^{q-2}\left(I - \frac{A^\top A(\theta_{t-1} - \hat{\theta})(\theta_{t-1} - \hat{\theta})^\top}{\|A(\theta_{t-1} - \hat{\theta})\|^2}\right)A^\top A(\theta_{t-1} - \hat{\theta}) \\ &= q r_{t-1}^{q-1}\|A(\theta_{t-1} - \hat{\theta})\|^{q-2}\left(A^\top A(\theta_{t-1} - \hat{\theta}) - \frac{A^\top A(\theta_{t-1} - \hat{\theta})\|A(\theta_{t-1} - \hat{\theta})\|^2}{\|A(\theta_{t-1} - \hat{\theta})\|^2}\right) \\ &= 0. \end{aligned} \tag{41}$$

This important observation shows that if the condition (33) holds, then the BFGS descent direction $H_t\nabla f(\theta_t)$ can be simplified as

$$\begin{aligned} &H_t\nabla f(\theta_t) \\ &= \left(I - \frac{s_{t-1}u_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\right)H_{t-1}\left(I - \frac{u_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\right)\nabla f(\theta_t) + \frac{s_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\nabla f(\theta_t) \\ &= \frac{s_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\nabla f(\theta_t) \\ &= \frac{(r_{t-1} - 1)^2(\theta_{t-1} - \hat{\theta})(\theta_{t-1} - \hat{\theta})^\top}{q(r_{t-1}^{q-1} - 1)(r_{t-1} - 1)\|A(\theta_{t-1} - \hat{\theta})\|^{q}} q r_{t-1}^{q-1}\|A(\theta_{t-1} - \hat{\theta})\|^{q-2}A^\top A(\theta_{t-1} - \hat{\theta}) \\ &= \frac{1 - r_{t-1}}{1 - r_{t-1}^{q-1}} r_{t-1}^{q-1}(\theta_{t-1} - \hat{\theta})\frac{\|A(\theta_{t-1} - \hat{\theta})\|^{q-2}(\theta_{t-1} - \hat{\theta})^\top A^\top A(\theta_{t-1} - \hat{\theta})}{\|A(\theta_{t-1} - \hat{\theta})\|^{q}} \\ &= \frac{1 - r_{t-1}}{1 - r_{t-1}^{q-1}} r_{t-1}^{q-1}(\theta_{t-1} - \hat{\theta}). \end{aligned} \tag{42}$$

This simplification implies that for the new iterate θ_{t+1}, we have

$$\begin{aligned} \theta_{t+1} - \hat{\theta} &= \theta_t - H_t\nabla f(\theta_t) - \hat{\theta} = \theta_t - \hat{\theta} - \frac{1 - r_{t-1}}{1 - r_{t-1}^{q-1}} r_{t-1}^{q-1}\frac{(\theta_t - \hat{\theta})}{r_{t-1}} \\ &= \frac{1 - r_{t-1}^{q-2}}{1 - r_{t-1}^{q-1}}(\theta_t - \hat{\theta}) = r_t(\theta_t - \hat{\theta}), \end{aligned} \tag{43}$$

where

$$r_t = \frac{1 - r_{t-1}^{q-2}}{1 - r_{t-1}^{q-1}}. \tag{44}$$

Therefore, we prove that condition (11) holds for $k = t + 1$. By induction, we prove the linear convergence results in (11) and (12).

One property of this convergence results is that the error vectors $\{\theta_k - \hat{\theta}\}_{k=0}^{\infty}$ are parallel to each other with the same direction as shown in (11). This indicates that the iterations $\{\theta_k\}_{k=0}^{\infty}$ converge to the optimal solution $\hat{\theta}$ along the same straight line defined by $\theta_0 - \hat{\theta}$. Only the length of each vector $\theta_k - \hat{\theta}$ reduces to zero with the linear convergence rates $\{r_k\}_{k=0}^{\infty}$ specified in (12) and the direction remains all the same.

A.2. Proof of Theorem 4.5

Recall that we have

$$r_0 = \frac{q-2}{q-1}; \qquad r_k = \frac{1 - r_{k-1}^{q-2}}{1 - r_{k-1}^{q-1}}; \qquad \forall k \geq 1. \tag{45}$$

Consider that $q \geq 4$ and define the function $g(r)$ as

$$g(r) := \frac{1 - r^{q-2}}{1 - r^{q-1}}; \qquad r \in [0, 1]. \tag{46}$$

Suppose that $r^* \in (0, 1)$ satisfying that $r^* = g(r^*)$, which is equivalent to

$$r^{q-1} + r^{q-2} = 1. \tag{47}$$

Notice that

$$g'(r) = \frac{(q-1)r^{q-2} - r^{2q-4} - (q-2)r^{q-3}}{(1 - r^{q-1})^2}; \tag{48}$$

and

$$\begin{aligned}
&(q-1)r^{q-2} - r^{2q-4} - (q-2)r^{q-3} \\
= \; & r^{q-3}[(q-1)(r-1) - (r^{q-1} - 1)] \\
= \; & r^{q-3}(r-1)\left(q - 1 - \frac{r^{q-1} - 1}{r - 1}\right) \\
= \; & r^{q-3}(r-1)\left(q - 1 - \sum_{i=0}^{q-2} r^i\right).
\end{aligned} \tag{49}$$

Since $r \in [0, 1]$, we have that

$$r^{q-3} \geq 0; \quad r - 1 \leq 0; \quad \sum_{i=0}^{q-2} r^i \leq \sum_{i=0}^{q-2} 1 = q - 1. \tag{50}$$

Therefore, we obtain that

$$(q-1)r^{q-2} - r^{2q-4} - (q-2)r^{q-3} \leq 0; \tag{51}$$

and

$$|g'(r)| = \frac{r^{2q-4} + (q-2)r^{q-3} - (q-1)r^{q-2}}{(1 - r^{q-1})^2}. \tag{52}$$

Our target is to prove that for any $r \in [0, 1]$,

$$|g'(r)| \leq \frac{1}{2}. \tag{53}$$

First, we present the plots of $|g'(r)|$ for $r \in [0, 1]$ with $4 \leq q \leq 11$ in Figure 3. We observe that for $4 \leq q \leq 11$, $|g'(r)| \leq 1/2$ always holds.

Next, we prove that for $q \geq 12$ and any $r \in [0, 1]$, we have

$$|g'(r)| = -\frac{(q-1)r^{q-2} - r^{2q-4} - (q-2)r^{q-3}}{(1 - r^{q-1})^2} \leq \frac{1}{2}; \tag{54}$$

which is equivalent to

$$r^{2q-2} - 2r^{2q-4} - 2r^{q-1} + 2(q-1)r^{q-2} - 2(q-2)r^{q-3} + 1 \geq 0; \qquad \forall r \in [0, 1]. \tag{55}$$

(a) $q = 4$. (b) $q = 5$. (c) $q = 6$. (d) $q = 7$.

(e) $q = 8$. (f) $q = 9$. (g) $q = 10$. (h) $q = 11$.

Figure 3. Plots of $|g'(r)|$ with $r \in [0,1]$ for $4 \le q \le 11$.

Define the function $h(r)$ as

$$h(r) := r^{2q-2} - 2r^{2q-4} - 2r^{q-1} + 2(q-1)r^{q-2} - 2(q-2)r^{q-3} + 1. \tag{56}$$

We obtain that

$$\frac{dh(r)}{dr} = 2r^{q-4}h^{(1)}(r), \tag{57}$$

where

$$h^{(1)}(r) := (q-1)r^{q+1} - 2(q-2)r^{q-1} - (q-1)r^2 + (q-1)(q-2)r - (q-2)(q-3). \tag{58}$$

Hence, we have that

$$\frac{dh^{(1)}(r)}{dr} = (q-1)h^{(2)}(r), \tag{59}$$

where

$$h^{(2)}(r) := (q+1)r^q - 2(q-2)r^{q-2} - 2r + q - 2. \tag{60}$$

Therefore, we obtain that

$$\frac{dh^{(2)}(r)}{dr} = h^{(3)}(r) := (q+1)qr^{q-1} - 2(q-2)^2r^{q-3} - 2, \tag{61}$$

and

$$\frac{dh^{(3)}(r)}{dr} = r^{q-4}h^{(4)}(r), \tag{62}$$

where

$$h^{(4)}(r) := q(q+1)(q-1)r^2 - 2(q-2)^2(q-3). \tag{63}$$

Now we define the function $l(q)$ as

$$\begin{aligned} l(q) &:= 2(q-2)^2(q-3) - q(q+1)(q-1) \\ &= q^3 - 14q^2 + 33q - 24 \\ &= q^2(q-14) + 33(q-1) + 9. \end{aligned} \tag{64}$$

We observe that for $q \geq 14$, we have $l(q) > 0$ and we calculate that $l(12) = 84 > 0$ and $l(13) = 236 > 0$. Hence, we obtain that $l(q) > 0$ for all $q \geq 12$, which indicates that for all $r \in [0,1]$,

$$r^2 - 1 < \frac{2(q-2)^2(q-3)}{q(q+1)(q-1)}; \tag{65}$$

$$q(q+1)(q-1)r^2 - 2(q-2)^2(q-3) < 0. \tag{66}$$

Therefore, for all $r \in [0,1]$, $h^{(4)}(r)$ defined in (63) satisfies that $h^{(4)}(r) < 0$ and from (62) we know that $\frac{dh^{(3)}(r)}{dr} < 0$. Hence, $h^{(3)}(r)$ defined in (61) is decreasing function and $h^{(3)}(r) <= h^{(3)}(0) = -2 < 0$. We know that $\frac{dh^{(2)}(r)}{dr} = h^{(3)}(r) < 0$, which implies that $h^{(2)}(r)$ defined in (60) is decreasing function. So we have that $h^{(2)}(r) \geq h^{(2)}(1) = 1 > 0$. From (59) we know that $\frac{dh^{(1)}(r)}{dr} > 0$ and $h^{(1)}(r)$ defined in (58) is increasing function for $r \in [0,1]$. Hence, we get that $h^{(1)}(r) \leq h^{(1)}(1) = 0$ and from (57) we obtain that $h(r)$ defined in (56) is decreasing function for $r \in [0,1]$. Therefore, we have that $h(r) \geq h(1) = 0$ and condition (55) holds for all $r \in [0,1]$, which is equivalent to $|g'(r)| \leq 1/2$.

In summary, we proved that for any $q \geq 12$, we have $|g'(r)| \leq 1/2$. Combining this with the results from Figure 3, we obtain that $|g'(r)| \leq 1/2$ holds for all $q \geq 4$. Applying Banach's Fixed-Point Theorem from Lemma A.2, we prove the final conclusion (14).

A.3. Proof of Theorem 4.6

Notice that the gradient and the Hessian of the objective function (10) can be expressed as

$$\nabla f(\theta) = q\|A\theta - b\|^{q-2}A^\top(A\theta - b) = q\|A\theta - b\|^{q-2}A^\top A(\theta - \hat{\theta}); \tag{67}$$

$$\nabla^2 f(\theta) = q\|A\theta - b\|^{q-2}A^\top A + q(q-2)\|A\theta - b\|^{q-4}A^\top(A\theta - b)(A\theta - b)^\top A. \tag{68}$$

Applying Lemma A.1, we can obtain that

$$\nabla^2 f(\theta)^{-1} = \frac{(A^\top A)^{-1}}{q\|A\theta - b\|^{q-2}} - \frac{(q-2)(\theta - \hat{\theta})(\theta - \hat{\theta})^\top}{q(q-1)\|A\theta - b\|^{q}}. \tag{69}$$

Hence, we have that for any $k \geq 1$,

$$\theta_k = \theta_{k-1} - \nabla^2 f(\theta_{k-1})^{-1}\nabla f(\theta_{k-1}); \tag{70}$$

$$\theta_k - \hat{\theta} = \theta_{k-1} - \hat{\theta} - \nabla^2 f(\theta_{k-1})^{-1}\nabla f(\theta_{k-1}). \tag{71}$$

Notice that $b = A\hat{\theta}$ by Assumption 4.1 and

$$\begin{aligned}
&\nabla^2 f(\theta_{k-1})^{-1}\nabla f(\theta_{k-1}) \\
&= \left[\frac{(A^\top A)^{-1}}{q\|A\theta_{k-1} - b\|^{q-2}} - \frac{(q-2)(\theta_{k-1} - \hat{\theta})(\theta_{k-1} - \hat{\theta})^\top}{q(q-1)\|A\theta_{k-1} - b\|^{q}}\right] q\|A\theta - b\|^{q-2}A^\top A(\theta_{k-1} - \hat{\theta}) \\
&= \theta_{k-1} - \hat{\theta} - \frac{q-2}{q-1}\frac{(\theta_0 - \hat{\theta})^\top A^\top A(\theta_{k-1} - \hat{\theta})}{\|A\theta_{k-1} - b\|^2}(\theta_{k-1} - \hat{\theta}) \\
&= \theta_{k-1} - \hat{\theta} - \frac{q-2}{q-1}\frac{(A\theta_{k-1} - b)^\top(A\theta_{k-1} - b)}{\|A\theta_{k-1} - b\|^2}(\theta_{k-1} - \hat{\theta}) \\
&= \theta_{k-1} - \hat{\theta} - \frac{q-2}{q-1}(\theta_{k-1} - \hat{\theta}).
\end{aligned} \tag{72}$$

Therefore, we prove the conclusion that for any $k \geq 1$,

$$\theta_k - \hat{\theta} = \theta_{k-1} - \hat{\theta} - \nabla^2 f(\theta_{k-1})^{-1}\nabla f(\theta_{k-1}) = \frac{q-2}{q-1}(\theta_{k-1} - \hat{\theta}). \tag{73}$$

We observe that the iterations generated by Newton's method also satisfy the parallel property, i.e., all vectors $\{\theta_k - \hat{\theta}\}_{k=0}^{\infty}$ are parallel to each other with the same direction.

Notice that the function $h(r) = r^{q-1} + r^{q-2}$ is strictly increasing and $h(\frac{q-2}{q-1}) < 1$, $h(r^*) = 1$ as well as $h(\frac{2q-3}{2q-2}) > 1$. Hence, we know that $\frac{q-2}{q-1} < r^* < \frac{2q-3}{2q-2}$.

A.4. Proof of Theorem 5.1

In the following proof, we use L and L_n to denote the population objective and empirical objective, and θ_t^n, θ_t, H_t^n and H_t to denote the corresponding iterations and Hessian approximating matrices in the population update and empirical update accordingly. For the ease of presentation, we use C_p to denote any constant that is independent of d, n, and C_p can be varied case by case to simply the proof. From Mou et al. (2019) and Ren et al. (2022b), we have that, as long as $n = \Omega((d \log d/\delta)^{2p})$, we have the following two uniform concentration results:

$$\begin{aligned} \sup_{\theta \in B(\theta^*, r)} \|\nabla L_n(\theta)) - \nabla L(\theta))\| &\leq C_p(\|\theta - \theta^*\| + r)^{p-1}\sqrt{d \log(1/\delta)/n}; \\ \sup_{\theta \in B(\theta^*, r)} \|\nabla^2 L_n(\theta)) - \nabla^2 L(\theta))\| &\leq C_p(\|\theta - \theta^*\| + r)^{p-2}\sqrt{d \log(1/\delta)/n}. \end{aligned} \tag{74}$$

Notice that we have

$$L(\theta) = E[(Y - (X^\top \theta)^p)^2] = E[((X^\top \theta^*)^p + \varepsilon - (X^\top \theta)^p)^2] \tag{75}$$

$$= E[((X^\top \theta^*)^p - (X^\top \theta)^p)^2] + E[2\varepsilon((X^\top \theta^*)^p - (X^\top \theta)^p)^2] + E[\varepsilon^2] \tag{76}$$

$$= E[((X^\top \theta^*)^p - (X^\top \theta)^p)^2] + \sigma^2; \tag{77}$$

where we use the fact that ε is independent of X and $E[\varepsilon] = 0$, $E[\varepsilon^2] = \sigma^2$. Hence, we have that

$$\nabla L(\theta) = 2pE[((X^\top \theta^*)^p - (X^\top \theta)^p)(X^\top \theta)^{p-1}X]. \tag{78}$$

$$\nabla^2 L(\theta) = 2p^2 E[(X^\top \theta)^{2p-2} X X^\top] + 2p(p-1)E[((X^\top \theta^*)^p - (X^\top \theta)^p)(X^\top \theta)^{p-2})XX^\top]. \tag{79}$$

We denote L^0 as the population loss function with respect to the assumption of $\theta^* = 0$. Therefore, we have that

$$L^0(\theta) = E[((X^\top \theta)^p)^2] + \sigma^2. \tag{80}$$

$$\nabla L^0(\theta) = 2pE[(X^\top \theta)^p)(X^\top \theta)^{p-1}X]. \tag{81}$$

$$\nabla^2 L^0(\theta) = 2p^2 E[(X^\top \theta)^{2p-2} X X^\top] - 2p(p-1)E[(X^\top \theta^*)^p (X^\top \theta)^{p-2})XX^\top]. \tag{82}$$

Hence, we have

$$\|\nabla L(\theta) - \nabla L^0(\theta)\| = C_p E[(X^\top \theta^*)^p (X^\top \theta)^{p-1} X] \leq C_p E[\|X\|^{2p}]\|\theta^*\|^p \|\theta\|^{p-1}. \tag{83}$$

Recall that X is a Gaussian or sub-Gaussian random variable with $E[\|X\|^{2p}] < +\infty$. For any $\theta \in B(\theta^*, r)$, $\|\theta\| \leq \|\theta - \theta^*\| + r$ and in the low SNR regime, we have $\|\theta^*\| \leq C_1(\frac{d}{n})^{\frac{1}{2p}}$. Hence, we have

$$\sup_{\theta \in B(\theta^*, r)} \|\nabla L(\theta) - \nabla L^0(\theta)\| \leq C_p(\|\theta - \theta^*\| + r)^{p-1}\sqrt{d \log(1/\delta)/n}. \tag{84}$$

Similarly, we have that

$$\|\nabla^2 L(\theta) - \nabla^2 L^0(\theta)\| = C_p E[(X^\top \theta^*)^p (X^\top \theta)^{p-2} X X^\top] \leq C_p E[\|X\|^{2p}]\|\theta^*\|^p \|\theta\|^{p-2}. \tag{85}$$

$$\sup_{\theta \in B(\theta^*, r)} \|\nabla^2 L(\theta) - \nabla^2 L^0(\theta)\| \leq C_p(\|\theta - \theta^*\| + r)^{p-2}\sqrt{d \log(1/\delta)/n}. \tag{86}$$

Leveraging (74), (84) and (86) we obtain that

$$\sup_{\theta \in B(\theta^*, r)} \|\nabla L_n(\theta)) - \nabla L^0(\theta))\| \leq C_p(\|\theta - \theta^*\| + r)^{p-1}\sqrt{d \log(1/\delta)/n}; \tag{87}$$

$$\sup_{\theta \in B(\theta^*, r)} \|\nabla^2 L_n(\theta)) - \nabla^2 L^0(\theta))\| \leq C_p(\|\theta - \theta^*\| + r)^{p-2}\sqrt{d \log(1/\delta)/n}. \tag{88}$$

This explains the remark of assumption $\theta^* = 0$ in section 4. The errors between gradients and Hessians of the population loss with $\theta^* = 0$ and $\|\theta - \theta^*\| \geq C_1(d/n)^{1/(2p)}$ are upper bounded by the corresponding statistical errors between the population loss (8) and the empirical loss (7) in the low SNR regime, respectively. Hence, $\mathcal{L}'$ and $\mathcal{L}$ can be treated as equivalent. In the following proof, we replace all $\mathcal{L}'$ with $\mathcal{L}$ in the general low SNR regime. We can assume that the results of Theorem 4.3 and 4.5 also hold for the $\mathcal{L}$, i.e., there exists linear convergence rates $\{r_t\}_{t=0}^{\infty}$ with

$$\|\theta_{t+1} - \theta^*\| \leq r_t \|\theta_t - \theta^*\|, \quad \forall t \geq 0.$$

Recall that $\mathcal{L}'(\theta) = \mathbb{E}_X[(X^\top \theta)^{2p}] = C_p\|\Sigma^{1/2}\theta\|^{2p}$. Since $\mathcal{L}'$ is equivalent to $\mathcal{L}$, we can also assume that

$$\lambda_{\max}(\nabla^2\mathcal{L}(\theta)) \leq C_p\|\theta - \theta^*\|^{2p-2}; \quad \lambda_{\min}(\nabla^2\mathcal{L}(\theta)) \geq C_p\|\theta - \theta^*\|^{2p-2}; \quad \|\nabla\mathcal{L}(\theta)\| \geq C_p\|\theta - \theta^*\|^{2p-1}. \tag{89}$$

We assume that $\|\theta - \theta^*\| \geq \|\theta^n - \theta^*\|$ for any θ. Otherwise, we have that $\|\theta - \theta^*\| < \|\theta^n - \theta^*\| \leq C_1(\frac{d}{n})^{\frac{1}{2p}}$ by the definition of the low SNR regime, which indicates that θ has already achieved the optimal statistical radius. Hence, for any $\theta \in \mathcal{B}(\theta^*; r)$, we have that $\|\theta^n - \theta^*\| \leq \|\theta - \theta^*\| \leq r$ and thus, we have the following conditions

$$\sup_{\theta \in \mathcal{B}(\theta^*; r)} \|\nabla\mathcal{L}_n(\theta) - \nabla\mathcal{L}(\theta)\| \leq C_p r^{p-1}\sqrt{d\log(1/\delta)/n}; \tag{90}$$

$$\sup_{\theta \in \mathcal{B}(\theta^*; r)} \|\nabla^2\mathcal{L}_n(\theta) - \nabla^2\mathcal{L}(\theta)\| \leq C_p r^{p-2}\sqrt{d\log(1/\delta)/n}. \tag{91}$$

In this proof, we use induction and apply conditions (89), (90) and (91) to prove the final results and start with the base case. Assume $\theta_0^n = \theta_0$, and $\|\theta_0 - \theta^*\|^p \geq C_p\varepsilon(n, \delta)$. For the first step, we have that

$$\begin{aligned} \|\theta_1^n - \theta_1\| &= \|(\nabla^2\mathcal{L}_n(\theta_0))^{-1}\nabla\mathcal{L}_n(\theta_0) - (\nabla^2\mathcal{L}(\theta_0))^{-1}\nabla\mathcal{L}(\theta_0)\| \\ &\leq \|\nabla^2\mathcal{L}_n(\theta_0)^{-1} - (\nabla^2\mathcal{L}(\theta_0))^{-1}\|\|\nabla\mathcal{L}_n(\theta_0)\| + \|\nabla^2\mathcal{L}(\theta_0)^{-1}\|\|\nabla\mathcal{L}_n(\theta_0) - \nabla\mathcal{L}(\theta_0)\| \end{aligned} \tag{92}$$

We observe that for invertible matrices A and B,

$$(A^{-1} - B^{-1}) = A^{-1}(B - A)B^{-1}. \tag{93}$$

We can bound

$$\begin{aligned} &\|\nabla^2\mathcal{L}_n(\theta_0)^{-1} - \nabla^2\mathcal{L}(\theta_0)^{-1}\| \\ &\leq \|(\nabla^2\mathcal{L}_n(\theta_0))^{-1}\|\|\nabla^2\mathcal{L}_n(\theta_0) - \nabla^2\mathcal{L}(\theta_0)\|\|(\nabla^2\mathcal{L}(\theta_0))^{-1}\| \\ &\leq C_p\|\theta_0 - \theta^*\|^{2-3p}\sqrt{d\log(1/\delta)/n}, \end{aligned} \tag{94}$$

which leads to the bound

$$\|\theta_1^n - \theta_1\| \leq C_p\|\theta_0 - \theta^*\|^{1-p}\sqrt{d\log(1/\delta)/n}. \tag{95}$$

We now show that $\forall t < T = O(\log(n/d))$, we have

$$\|\theta_t^n - \theta_t\| \leq c_t\|\theta_{t-1} - \theta^*\|^{1-p}\sqrt{d\log(1/\delta)/n}, \tag{96}$$

where $c_t = \Theta(\exp(t))$. Before we start, we assume $\forall t < T$, we have

$$c_t\|\theta_t - \theta^*\|^{-p}\sqrt{d\log(1/\delta)/n} \leq \frac{1}{C_p}. \tag{97}$$

Otherwise, our results simply follow. We now prove (96) by induction. Notice that from (95), we know that (96) holds for $t = 1$. Assume for $k \leq t$, we have $\|\theta_k^n - \theta_k\| \leq c_k\|\theta_k - \theta^*\|^{1-p}\sqrt{d\log(1/\delta)/n}$. Note that

$$\|\theta_{t+1}^n - \theta_{t+1}\| \leq \|\theta_t^n - \theta_t\| + \|H_t^n\nabla\mathcal{L}_n(\theta_t^n) - H_t\nabla\mathcal{L}(\theta_t)\|. \tag{98}$$

We apply the update of

$$H_t^n = \left(I - \frac{s_{t-1}^n (u_{t-1}^n)^\top}{(u_{t-1}^n)^\top s_{t-1}^n}\right) H_{t-1}^n \left(I - \frac{u_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right) + \frac{s_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}; \tag{99}$$

$$H_t = \left(I - \frac{s_{t-1} (u_{t-1})^\top}{(u_{t-1})^\top s_{t-1}}\right) H_{t-1} \left(I - \frac{u_{t-1} (s_{t-1})^\top}{(s_{t-1})^\top u_{t-1}}\right) + \frac{s_{t-1} (s_{t-1})^\top}{(s_{t-1})^\top u_{t-1}}; \tag{100}$$

and

$$\left(I - \frac{u_{t-1} s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\right) \nabla L(\theta_t) = 0\,; \tag{101}$$

which follows equation (41) from the population analysis. Then, we have the following decomposition:

$$\begin{aligned}
\|H_t^n \nabla L_n(\theta_t^n) - H_t \nabla f(\theta_t)\| &= \left\| H_t^n \nabla L_n(\theta_t^n) - \frac{s_{t-1} s_{t-1}^\top}{s_{t-1}^\top u_{t-1}} \nabla L(\theta_t) \right\| \\
&\le \left\| \left(I - \frac{s_{t-1}^n (u_{t-1}^n)^\top}{(u_{t-1}^n)^\top s_{t-1}^n}\right) H_{t-1}^n \left(I - \frac{u_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right) \nabla L_n(\theta_t^n) \right\| + \left\| \frac{s_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n} \nabla L_n(\theta_t^n) - \frac{s_{t-1} s_{t-1}^\top}{s_{t-1}^\top u_{t-1}} \nabla L(\theta_t) \right\|;
\end{aligned} \tag{102}$$

Now we bound these two terms separately. The first term can be bounded as

$$\begin{aligned}
&\left\| \left(I - \frac{s_{t-1}^n (u_{t-1}^n)^\top}{(u_{t-1}^n)^\top s_{t-1}^n}\right) H_{t-1}^n \left(I - \frac{u_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right) \nabla L_n(\theta_t^n) \right\| \\
&\le \left\| I - \frac{s_{t-1}^n (u_{t-1}^n)^\top}{(u_{t-1}^n)^\top s_{t-1}^n} \right\| \|H_{t-1}^n\| \left\| \left(I - \frac{u_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right) \nabla L_n(\theta_t^n) \right\| \\
&\le \|H_{t-1}^n\| \left\| \left(I - \frac{u_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right) \nabla L_n(\theta_t^n) \right\|;
\end{aligned} \tag{103}$$

and

$$\begin{aligned}
&\left\| \left(I - \frac{u_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right) \nabla L_n(\theta_t^n) \right\| = \left\| \left(I - \frac{u_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right) \nabla L_n(\theta_t^n) - \left(I - \frac{u_{t-1} s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\right) \nabla L(\theta_t) \right\| \\
&\le \|\nabla L_n(\theta_t^n) - \nabla L(\theta_t)\| + \left\| \frac{u_{t-1}^n (s_{t-1}^n)^\top \nabla L_n(\theta_t^n)}{(s_{t-1}^n)^\top u_{t-1}^n} - \frac{u_{t-1}^n s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}} \right\| + \left\| \frac{u_{t-1}^n s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}} - \frac{u_{t-1} s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}} \right\| \\
&\le \|\nabla L_n(\theta_t^n) - \nabla L(\theta_t)\| + \|u_{t-1}^n\| \left| \frac{(s_{t-1}^n)^\top \nabla L_n(\theta_t^n)}{(s_{t-1}^n)^\top u_{t-1}^n} - \frac{s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}} \right| + \|u_{t-1}^n - u_{t-1}\| \left| \frac{s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}} \right|;
\end{aligned} \tag{104}$$

where we use the fact that $\left\| \frac{s_{t-1}^n (u_{t-1}^n)^\top}{(u_{t-1}^n)^\top s_{t-1}^n} \right\| \le 1$ and $\left(I - \frac{u_{t-1} s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\right) \nabla L(\theta_t) = 0$. The second term can be bounded as

$$\begin{aligned}
&\left\| \frac{s_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n} \nabla L_n(\theta_t^n) - \frac{s_{t-1} s_{t-1}^\top}{s_{t-1}^\top u_{t-1}} \nabla L(\theta_t) \right\| \\
&\le \left\| \frac{s_{t-1}^n (s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n} \nabla L_n(\theta_t^n) - \frac{s_{t-1}^n s_{t-1}^\top}{s_{t-1}^\top u_{t-1}} \nabla L(\theta_t) \right\| + \left\| \frac{s_{t-1}^n s_{t-1}^\top}{s_{t-1}^\top u_{t-1}} \nabla L(\theta_t) - \frac{s_{t-1} s_{t-1}^\top}{s_{t-1}^\top u_{t-1}} \nabla L(\theta_t) \right\| \\
&\le \|s_{t-1}^n\| \left| \frac{(s_{t-1}^n)^\top \nabla L_n(\theta_t^n)}{(s_{t-1}^n)^\top u_{t-1}^n} - \frac{s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}} \right| + \|s_{t-1} - s_{t-1}^n\| \left| \frac{s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}} \right|.
\end{aligned} \tag{105}$$

We start from the bound on several basic terms.

- Bounds related to s^n_{t-1}:

$$\begin{aligned}\|s^n_{t-1} - s_{t-1}\| &\le \|\theta^n_t - \theta_t\| + \|\theta^n_{t-1} - \theta_{t-1}\| \\ &\le (c_t\|\theta_{t-1} - \theta^*\|^{1-p} + c_{t-1}\|\theta_{t-2} - \theta^*\|^{1-p})\sqrt{d\log(1/\delta)/n} \\ &\le (c_t + c_{t-1} r_{t-1}^{p-1})\|\theta_{t-1} - \theta^*\|^{1-p}\sqrt{d\log(1/\delta)/n};\end{aligned} \tag{106}$$

which also gives

$$\begin{aligned}\|s^n_{t-1}\| &\le \|s_{t-1}\| + \|s^n_{t-1} - s_{t-1}\| \\ &= (1 - r_{t-1})\|\theta_{t-1} - \theta^*\| + (c_t\|\theta_{t-1} - \theta^*\|^{1-p} + c_{t-1}\|\theta_{t-2} - \theta^*\|^{1-p})\sqrt{d\log(1/\delta)/n} \\ &\le (1 - r_{t-1})\|\theta_{t-1} - \theta^*\| + \frac{1 + 1/r_{t-2}}{C_p}\|\theta_{t-1} - \theta^*\| \\ &\le \|\theta_{t-1} - \theta^*\|;\end{aligned} \tag{107}$$

where we use the fact that $c_t\|\theta_{t-1} - \theta^*\|^{-p}\sqrt{d\log(1/\delta)/n} \le \frac{1}{C_p}$ and we can choose sufficiently large C_p to make the last inequality holds.

- Bounds related to $\nabla L_n(\theta^t_n)$:

$$\begin{aligned}&\|\nabla L_n(\theta^n_t) - \nabla L(\theta_t)\| \\ \le&\ \|\nabla L_n(\theta^n_t) - \nabla L(\theta^n_t)\| + \|\nabla L(\theta^n_t) - \nabla L(\theta_t)\| \\ \le&\ C_p\|\theta^n_t - \theta^*\|^{p-1}\sqrt{d\log(1/\delta)/n} + C_p\|\theta^n_t - \theta_t\|(\|\theta^n_t - \theta_t\| + \|\theta_t - \theta^*\|)^{2p-2} \\ \le&\ C_p(\|\theta^n_t - \theta_t\| + \|\theta_t - \theta^*\|)^{p-1}\sqrt{d\log(1/\delta)/n}\left(1 + c_t\|\theta_{t-1} - \theta^*\|^{1-p}(\|\theta^n_t - \theta_t\| + \|\theta_t - \theta^*\|)^{p-1}\right) \\ \le&\ (C_p + C_p c_t)\|\theta_{t-1} - \theta^*\|^{p-1}\sqrt{d\log(1/\delta)/n};\end{aligned} \tag{108}$$

which also gives

$$\|\nabla L_n(\theta^n_t)\| \le \|\nabla L(\theta_t)\| + \|\nabla L_n(\theta^n_t) - \nabla L(\theta_t)\| \le C_p\|\theta_t - \theta^*\|^{2p-1}; \tag{109}$$

where we still use the fact that $c_t\|\theta_{t-1} - \theta^*\|^{-p}\sqrt{d\log(1/\delta)/n} \le \frac{1}{C_p}$ and we can choose sufficiently large C_p to make the noise term negligible.

- Bounds related to u^n_{t-1}:

$$\begin{aligned}\|u^n_{t-1} - u_{t-1}\| &\le \|\nabla L_n(\theta^n_t) - \nabla L(\theta_t)\| + \|\nabla L_n(\theta^n_{t-1}) - \nabla L(\theta_{t-1})\| \\ &\le (C_p + C_p c_t + C_p c_{t-1})\|\theta_{t-1} - \theta^*\|^{p-1}\sqrt{d\log(1/\delta)/n}.\end{aligned} \tag{110}$$

With the same technique, we can show that

$$\|u^n_{t-1}\| \le C_p\|\theta_{t-1} - \theta^*\|^{2p-1}. \tag{111}$$

- Bound of $|(s^n_{t-1})^\top \nabla L_n(\theta^n_t) - s^\top_{t-1}\nabla L(\theta_t)|$:

$$\begin{aligned}&|(s^n_{t-1})^\top \nabla L_n(\theta^n_t) - s^\top_{t-1}\nabla L(\theta_t)| \\ \le&\ \|s^n_{t-1} - s_{t-1}\|\|\nabla L_n(\theta^n_t)\| + \|s_{t-1}\|\|\nabla L_n(\theta^n_t) - \nabla L(\theta_t)\| \\ \le&\ (C_p + C_p c_t + C_p c_{t-1})\|\theta_{t-1} - \theta^*\|^{p}\sqrt{d\log(1/\delta)/n}.\end{aligned} \tag{112}$$

And with the same technique, we have

$$C_p\|\theta_{t-1} - \theta^*\|^{2p} \le (s^n_{t-1})^\top \nabla L_n(\theta^n_t) \le C_p\|\theta_{t-1} - \theta^*\|^{2p}.$$

- Bound of $|(s_{t-1}^n)^\top u_{t-1}^n - s_{t-1}^\top u_{t-1}|$:

$$
\begin{aligned}
&|(s_{t-1}^n)^\top (u_{t-1}^n) - s_{t-1}^\top u_{t-1}| \\
&\le \|s_{t-1}^n - s_{t-1}\|\|u_{t-1}^n\| + \|s_{t-1}\|\|u_{t-1}^n - u_{t-1}\| \\
&\le (C_p + C_p c_t + C_p c_{t-1})\|\theta_{t-1} - \theta^*\|^p \sqrt{d\log(1/\delta)/n},
\end{aligned} \tag{113}
$$

which also gives

$$
C_p\|\theta_{t-1} - \theta^*\|^{2p} \le (s_{t-1}^n)^\top u_{t-1}^n \le C_p\|\theta_{t-1} - \theta^*\|^{2p}. \tag{114}
$$

- Bound for $\frac{(s_{t-1}^n)^\top \nabla L_n(\theta_t^n)}{(s_{t-1}^n)^\top u_{t-1}^n} - \frac{s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}}$:

$$
\begin{aligned}
&\frac{(s_{t-1}^n)^\top \nabla L_n(\theta_t^n)}{(s_{t-1}^n)^\top u_{t-1}^n} - \frac{s_{t-1}^\top \nabla L(\theta_t)}{s_{t-1}^\top u_{t-1}} \\
&\le \frac{\left((s_{t-1}^n)^\top \nabla L_n(\theta_t^n) - s_{t-1}^\top \nabla L(\theta_t)\right) s_{t-1}^\top u_{t-1} + s_{t-1}^\top \nabla L(\theta_t)\left(s_{t-1}^\top u_{t-1} - (s_{t-1}^n)^\top u_{t-1}^n\right)}{(s_{t-1}^n)^\top u_{t-1}^n s_{t-1}^\top u_{t-1}} \\
&\le (C_p + C_p c_t + C_p c_{t-1})\|\theta_{t-1} - \theta^*\|^{-p}\sqrt{d\log(1/\delta)/n}.
\end{aligned} \tag{115}
$$

- Straightforward computation shows that $s_{t-1}^\top \nabla L(\theta_t) = C_p\|\theta_t - \theta^*\|^{2p}$, $s_{t-1}^\top u_{t-1} = C_p\|\theta - \theta^*\|^{2p}$.

To summarize, we have that

$$
\left\|\left(I - \frac{u_{t-1}^n(s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right)\nabla L_n(\theta_t^n)\right\| \le (C_p + C_p c_t + C_p c_{t-1})\|\theta_{t-1} - \theta^*\|^{p-1}\sqrt{d\log(1/\delta)/n}; \tag{116}
$$

$$
\left\|\frac{s_{t-1}^n(s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\nabla L_n(\theta_t^n) - \frac{s_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}\nabla L(\theta_t)\right\| \le (C_p + C_p c_t + C_p c_{t-1})\|\theta_{t-1} - \theta^*\|^{1-p}\sqrt{d\log(1/\delta)/n}. \tag{117}
$$

Now we bound $\|H_t^n\|$. Recall the update of H_t^n

$$
H_t^n = \left(I - \frac{s_{t-1}^n(u_{t-1}^n)^\top}{(u_{t-1}^n)^\top s_{t-1}^n}\right) H_{t-1}^n \left(I - \frac{u_{t-1}^n(s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right) + \frac{s_{t-1}^n(s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}. \tag{118}
$$

With the property of $\left\|I - \frac{s_{t-1}^n(u_{t-1}^n)^\top}{(u_{t-1}^n)^\top s_{t-1}^n}\right\| \le 1$ and $\left\|I - \frac{u_{t-1}^n(s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right\| \le 1$, we have that

$$
\begin{aligned}
\|H_t^n\| &\le \left\|I - \frac{s_{t-1}^n(u_{t-1}^n)^\top}{(u_{t-1}^n)^\top s_{t-1}^n}\right\|\|H_{t-1}^n\|\left\|I - \frac{u_{t-1}^n(s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right\| + \left\|\frac{s_{t-1}^n(s_{t-1}^n)^\top}{(s_{t-1}^n)^\top u_{t-1}^n}\right\| \\
&\le \|H_{t-1}^n\| + \frac{\|s_{t-1}^n\|^2}{(s_{t-1}^n)^\top u_{t-1}^n} \\
&\le \|\nabla^2 L_n(\theta_0)^{-1}\| + \sum_{i=0}^{t-1}\frac{\|s_i^n\|^2}{(s_i^n)^\top (u_i^n)}.
\end{aligned} \tag{119}
$$

From the previous computation, we know

$$
\frac{\|s_i^n\|^2}{(s_i^n)^\top (u_i^n)} \le C_p\|\theta_i - \theta^*\|^{2-2p}. \tag{120}
$$

From Theorem 4.5 of the population analysis, we know that there exists $0 < r_l < r_h < 1$, such that for all the linear convergence rates $\{r_t\}_{t=0}^{\infty}$, we have $r_l \le r_t \le r_h$ for all $t \ge 0$ and $\|\theta_{t+1} - \theta^*\| = r_t\|\theta_t - \theta^*\|$ for all $t \ge 0$. Hence, we have that for all $0 \le i \le t-1$,

$$
\|\theta_{t-1} - \theta^*\| = \prod_{j=i}^{t-2} r_j\|\theta_i - \theta^*\| \le r_h^{t-1-i}\|\theta_i - \theta^*\|; \tag{121}
$$

$$\|\theta_i - \theta\| \le r_h^{i+1-t}\|\theta_{t-1} - \theta\|; \tag{122}$$

$$\|\theta_i - \theta\|^{2-2p} \le r_h^{(2p-2)(t-1-i)}\|\theta_{t-1} - \theta\|^{2-2p}; \tag{123}$$

$$\sum_{i=0}^{t-1}\|\theta_i - \theta\|^{2-2p} \le \sum_{i=0}^{t-1} r_h^{(2p-2)(t-1-i)}\|\theta_{t-1} - \theta\|^{2-2p} \le \frac{1}{1-r_h^{2p-2}}\|\theta_{t-1} - \theta\|^{2-2p}. \tag{124}$$

Hence, we obtain that

$$\begin{aligned}\|H_t^n\| &\le \|\nabla^2 L_n(\theta_0)^{-1}\| + \sum_{i=0}^{t-1}\frac{\|s_i^n\|^2}{(s_i^n)^\top (u_i^n)} \\ &\le \frac{1}{\lambda_{\min}(\nabla^2 L_n(\theta_0))} + C_p\sum_{i=0}^{t-1}\frac{\|s_i^n\|^2}{(s_i^n)^\top (u_i^n)} \\ &\le \frac{1}{2p\|\theta_0 - \theta\|^{2p-2}} + C_p\frac{1}{1-r_h^{2p-2}}\|\theta_{t-1} - \theta\|^{2-2p} \\ &\le \left(\frac{r_h^{(2p-2)(t-1)}}{2p} + \frac{C_p}{1-r_h^{2p-2}}\right)\|\theta_{t-1} - \theta\|^{2-2p}.\end{aligned} \tag{125}$$

As long as C_p is large enough, we obtain that

$$\|H_t^n\| \le C_p\|\theta_{t-1} - \theta\|^{2-2p}; \tag{126}$$

and

$$\|H_{t-1}^n\| \le C_p\|\theta_{t-2} - \theta\|^{2-2p} \le C_p\frac{1}{r_{t-2}^{2-2p}}\|\theta_{t-1} - \theta\|^{2-2p} \le C_p r_h^{2p-2}\|\theta_{t-1} - \theta\|^{2-2p}. \tag{127}$$

Combining the previous results of (102), (103), (116), (117) and (127), we have that

$$\|H_t^n\nabla L_n(\theta_t^n) - H_t\nabla L(\theta_t)\| \le (C_p + C_p c_t + C_p c_{t-1})\|\theta_{t-1} - \theta\|^{1-p}\sqrt{d\log(1/\delta)/n}. \tag{128}$$

Notice that by induction, we observe that

$$\begin{aligned}\|\theta_t^n - \theta_t\| &\le c_t\|\theta_{t-1} - \theta\|^{1-p}\sqrt{d\log(1/\delta)/n} \\ &= c_t\frac{1}{r_{t-1}^{1-p}}\|\theta_t - \theta\|^{1-p}\sqrt{d\log(1/\delta)/n} \\ &\le c_t r_h^{p-1}\|\theta_t - \theta\|^{1-p}\sqrt{d\log(1/\delta)/n} \\ &\le (C_p + C_p c_t + C_p c_{t-1})\|\theta_t - \theta\|^{1-p}\sqrt{d\log(1/\delta)/n}.\end{aligned} \tag{129}$$

Therefore for C_p large enough, we have that

$$\begin{aligned}\|\theta_{t+1}^n - \theta_{t+1}\| &\le \|\theta_t^n - \theta_t\| + \|H_t^n\nabla L_n(\theta_t^n) - H_t\nabla L(\theta_t)\| \\ &\le (C_p + C_p c_t + C_p c_{t-1})\|\theta_t - \theta\|^{1-p}\sqrt{d\log(1/\delta)/n}.\end{aligned} \tag{130}$$

We define that

$$c_{t+1} = C_p + C_p c_t + C_p c_{t-1}; \tag{131}$$

and we prove that

$$\|\theta_{t+1}^n - \theta_{t+1}\| \le c_{t+1}\|\theta_t - \theta\|^{1-p}\sqrt{d\log(1/\delta)/n}. \tag{132}$$

With the standard recursion, we know that $c_t \quad (C_p)^t$ for C_p large enough. Therefore, using induction we proved that (96) holds:

$$k \,{}^n_t \quad {}_t k \quad c_t k \,{}_{t \;\; 1} \quad k^1 \;{}^p \sqrt{p} \, \overline{d \log(1= \;)=n}; \tag{133}$$

where $c_t = \;(\exp(\;\; t))$. Notice that

$$k \,{}^n_t \quad k \;\; k \;{}^n_t \quad {}^t k + k \;{}_t \quad k \quad (C_p)^t k \;{}_t \quad k^1 \;{}^p \sqrt{p} \, \overline{d \log(1= \;)=n} + k \;{}_t \quad k: \tag{134}$$

The optimal T with minimum $k \,{}^n_T \quad k$ should satisfy that

$$C_p^T k \;{}_T \quad k \;{}^p \sqrt{p} \, \overline{d \log(1= \;)=n} = C_p; \tag{135}$$

for which we obtain $T = \frac{C \log(n=d \log(1= \;))}{2(p+1)}$ for some $C = \text{polylog}(p)$, and $k \,{}^n_T \quad k \quad C_p(d \log(1= \;)=n)^{1=(2p+2)}$, that nishes the proof.

B. Additional Experiments Results

B.1. Experiments for Linear Factors

(a) $q = 4$. (b) $q = 100$. (c) $q = 4$. (d) $q = 100$.

Figure 4. Convergence of factors $f r_k g^1_{k=0}$ to r .

B.2. Additional Experiments for the Empirical Loss

(a) High SNR regime. (b) Low SNR regime. (c) High SNR regime. (d) Low SNR regime.

Figure 5. Convergence and statistical results for $d = 50$. Convergence of different methods for high SNR regime are shown in (a) and low SNR regime in (b). Statistical radius of BFGS in high SNR regime and low SNR regime are shown in (c) and (d).

To show that BFGS can also be applied to high dimension scenarios, we conduct additional experiments on the generalized linear model with input $d = 50; 100; 500$ and the power of link function $p = 2$. The inputs are generated by $f X_i g^n_{i=1}$ $N(0; \text{diag}(\;{}^2_1; \;\; ; \;\;{}^2_d))$ where $\;_k = (0:96)^{k \;\; 1}$, and the remaining setting and hyper-parameters are set identical to the low dimension scenarios. The results are shown in Figure 5 and 6. As the results show, the performance of BFGS in high dimensional scenarios are nearly identical to the low dimensional scenarios.