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

Qiujiang Jin QIUJIANG@AUSTIN.UTEXAS.EDU
Tongzheng Ren TONGZHENG@UTEXAS.EDU
Nhat Ho MINHNHAT@UTEXAS.EDU
Aryan Mokhtari MOKHTARI@AUSTIN.UTEXAS.EDU

Univerisity of Texas at Austin, Austin, TX, USA

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. The slow convergence of GD for the low SNR case is mainly due to the local convexity of the least-square loss functions of the GLMs. To address the shortcomings of GD, we propose to use the BFGS quasi-Newton method to solve parameter estimation of the GLMs. 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. 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. We also present numerical experiments that match our theoretical findings.

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 $\mathcal{P}_{\theta^*}$ parameterized by a ground truth parameter θ^*. Our goal as the learner is to find θ^* by solving the empirical risk minimization (ERM) problem:

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

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 approximate of the ground-truth solution θ^*, where θ^* is also a minimizer of the population loss defined as

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

If one can solve the empirical risk efficiently, the output model could be close to θ^*, when n is sufficiently large. There are several works on studying 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 [1, 2, 6, 11, 19, 20, 26, 37]. However, when we move beyond linear models, the underlying loss becomes non-convex and the behavior of iterative methods could change and they may not reach a neighborhood of a global minimizer of the ERM problem.

The focus of this paper is on the generalized linear model (GLM) [7, 12, 13, 28, 35] 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 and its corresponding population risk minimization problem could change substantially. More precisely, if we are in the setting that $\|\theta^*\|$ is sufficiently large, which we refer to this case as the high signal-to-noise ratio (SNR), the underlying population loss of the problem of interest is locally strongly convex and smooth. On the other hand, when we are 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 statistical radius exponentially fast. 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, in [30] the authors recommended the use of GD with Polyak step size to accelerate the convergence of GD in the low SNR setting, and they showed that the number of iterations becomes logarithmic function of the sample size. However, as this method is still a first-order method, its complexity scales linearly by the condition number of the problem which depends on the condition number of the feature vectors covariance as well as the norm $\|\theta^*\|$. Moreover, implementation of Polyak step size requires the knowledge of optimal objective function value. These points lead to the following question: *Can we find a method that performs well in both high and low SNR settings at a reasonable per iteration computational cost?*

Contributions. In this paper, we show that the answer to the above question is positive and the BFGS method is capable of achieving these goals. In particular, we show that in the low SNR regime, which is not strictly convex, the iterates generated by the BFGS method converges linearly (exponentially fast) and outperforms GD. We also discuss why in the high SNR regime the BFGS method converges to the ground-truth (in the population case) at a superlinear rate.

2. BFGS algorithm

In this section, we briefly review the basics of the BFGS quasi-Newton method, which is the main algorithm we analyze in this paper. Consider the case that we aim to minimize a differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$ with optimal solution $\hat{\theta}$. The iterative method is defined as

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

where $H_k \in \mathbb{R}^{d \times d}$ is the matrix and η_k is the step size. The main idea of quasi-Newton methods is to come up with a Hessian inverse approximation matrix H_k that is close to the exact Hessian inverse $\nabla^2 f(\theta_k)^{-1}$ using only first-order information. There are several approaches for approximating H_k leading to different quasi-Newton methods, [3–5, 8, 9, 14–16, 18, 25, 29, 34], but here, we focus on the BFGS method, in which H_k is updated as

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

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

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

respectively. The main logic behind the update in (4) is to ensure that the Hessian inverse approximation matrix H_k satisfies the secant condition $H_k u_{k-1} = s_{k-1}$.

The main advantage of BFGS is its fast superlinear convergence rate. For the past few decades, all the superlinear convergence results of quasi-Newton method are asymptotic. Recently, non-asymptotic superlinear convergence rates of quasi-Newton method have been established in [21–24, 31–33, 36]. However, all these superlinear convergence analyses require the objective function to be smooth and strictly or strongly convex. In this paper, as mentioned later, we will face settings in which the Hessian at the optimal solution could be singular, and, hence the above convergence guarantees do not hold and hence we need to establish new convergence guarantees for BFGS.

3. Generalized linear model with polynomial link function

In this section, we 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 $(Y_1, X_1), \ldots, (Y_n, X_n)$ that are i.i.d. samples from the following GLM with polynomial link function of power p, 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 noises with zero mean and variance σ^2. Further, we assume that the feature vectors are such that $X \in \mathbb{R}^d \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{d \times d}$ is a positive definite matrix. In this setting, when $p = 1$, the model in (6) is the standard linear model, while when $p = 2$, it corresponds to the phase retrieval model [13]. Here, we focus on the settings that $p \geq 2$. For the parameter estimation, there exist three regimes of GLMs (6): (1) 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; (2) Middle SNR regime: $C_1(d/n)^{1/(2p)} \leq \|\theta^*\|/\sigma \leq C_2$ where C_2 is a universal constant; and (3) High SNR regime: $\|\theta^*\| \geq C_2$.

It can be verified that the ERM problem corresponding to the above model for $p \geq 2$ with quadratic loss ℓ is a non-convex function with respect to θ and finding a global minimizer of that could be a challenging task. On the other hand, it is locally strongly convex when we are in the high signal-to-noise ratio (SNR) case. In the low SNR setting, however, the problem becomes convex and the strong convexity condition does not hold. To showcase this issue, let us focus first on the population loss, which is the limit of the ERM when the sample size goes to infinity. Note that the population loss in the considered generalized linear model (6) with a quadratic loss function ℓ is given by $\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \mathbb{E}_{X,Y}[(Y - (X^\top \theta)^p)^2]$, which based on the assumptions on the generalized linear model setting and the distribution of the noise can be simplified as

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \mathbb{E}_X \Big[\Big((X^\top \theta^*)^p - (X^\top \theta)^p \Big)^2 \Big] + \sigma^2. \tag{7}$$

Indeed, the ground truth parameter θ^* is an optimal solution of the above. Next, we discuss the structure of the objective function in (7) for low and high SNR settings.

High signal-to-noise regime. In the setting that the ground truth parameter has a relative large norm, i.e., $\|\theta^*\|/\sigma \geq C$ for some universal constant $C > 0$, the population loss function in (7) is locally strongly convex and smooth around θ^*. More precisely, when $\|\theta - \theta^*\|$ is small, we have $(X^\top\theta^*)^p - (X^\top\theta)^p = p(X^\top\theta^*)^{p-1}X^\top(\theta - \theta^*) + o(\|\theta - \theta^*\|)$. Hence, in a neighborhood of the optimal solution, the objective in (7) can be approximated as

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

Indeed, if $\|\theta^*\| \geq C\sigma^2$ the above objective function behaves as a quadratic function that is smooth and strongly convex, assuming that $o(\|\theta - \theta^*\|^2)$ is negligible. Since the population loss in the high SNR case is approximately a strongly convex smooth quadratic function, the iterates of gradient descent (GD) converge to the solution at a linear rate and hence 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 the optimal solution and the rate would be independent of κ, however the cost per iteration would be $\mathcal{O}(d^2)$.

Low signal-to-noise 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 close to zero, the strong convexity parameter approaches zero 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 $\theta^* = 0$. 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

$$\mathcal{L}(\theta) = \mathbb{E}_X[(X^\top\theta)^{2p}] + \sigma^2 = (2p-1)!!\|\Sigma^{1/2}\theta\|^{2p} + \sigma^2. \tag{8}$$

Since we focus on $p \geq 2$ it can be verified that this objective function is not strongly convex in a neighborhood of $\theta^* = 0$. For this problem, GD with constant step size converges at a sublinear rate, and hence, GD iterates require polynomial number of iterations to reach the final statistical radius.

4. Convergence analysis for the low signal-to-noise case

In this section, we focus on the convergence properties of BFGS for solving the population loss in the case of low signal-to-noise introduced in (8). This analysis provides an intuition for the analysis of the finite sample case that we discuss in Section B, 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 (8) can be considered as a special case of the following convex optimization problem:

$$\min_{\theta\in\mathbb{R}^d} f(\theta) = \|A\theta - b\|^q, \tag{9}$$

where $A \in \mathbb{R}^{m\times d}$ is a matrix, $b \in \mathbb{R}^m$ is a given vector, and q satisfies that $q \geq 4$. We should note that for $q \geq 4$, the considered objective is not strictly convex because the Hessian matrix 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 (8) for $q = 2p$. Instead of solving (8), we focus on the convergence analysis of BFGS quasi-Newton method for solving the convex function in (9), as it is a more general function and our results are of general interest from an optimization point of view. To the best of our knowledge, there is no global convergence theory (without line-search) for BFGS when the function is not strictly convex, and our analysis provides the first result for such general setting.

Assumption 4.1 *We assume that there exists $\hat{\theta} \in \mathbb{R}^d$, such that $b = A\hat{\theta}$. This is equivalent to that vector b is in the range space of matrix A.*

Note that the above assumption is indeed satisfied in our considered setting low SNR case in (8) as we assume $\theta^* = 0$ which implies $b = 0$.

Assumption 4.2 *The matrix $A^\top A \in \mathbb{R}^{d \times d}$ is invertible. This is equivalent to that matrix $A^\top A$ is symmetric positive definite, i.e. $A^\top A \succ 0$.*

The above assumption is also easily satisfied for our considered setting as we assume that the covariance matrix for our input features is positive definite. Combining Assumption 4.1 and 4.2, we conclude that $\hat{\theta}$ is the unique optimal solution of the problem (9). Next, we formally state the convergence rate of BFGS for solving problem 9 under the disclosed assumptions.

Theorem 1 *Consider the update of BFGS in (3), (4) and (5). Suppose Assumptions 4.1 and 4.2 are satisfied, and the initial point θ_0 is an arbitrary vector in $\mathbb{R}^d$ and the initial Hessian inverse approximation matrix is selected as $H_0 = \nabla^2 f(\theta_0)^{-1}$. If the step size of BFGS is selected as $\eta_k = 1$ for all $k \geq 0$, then the iterates of BFGS converge to the optimal solution $\hat{\theta}$ at a linear rate of*

$$\|\theta_k - \hat{\theta}\| \leq r_{k-1}\|\theta_{k-1} - \hat{\theta}\|, \quad \forall k \geq 1, \tag{10}$$

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

$$r_0 = \frac{q-2}{q-1}, \qquad r_k = \frac{1 - r_{k-1}^{q-2}}{1 - r_{k-1}^{q-1}}, \quad \forall k \geq 1. \tag{11}$$

The above theorem shows that the iterates of BFGS converge globally at a linear rate to the optimal solution of (9). 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 $k \geq 0$. Moreover, the initial point θ_0 could be any vector and there is no restriction on the distance between θ_0 and optimal solution $\hat{\theta}$. Most analyses of quasi-Newton methods 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. Note that the result in Theorem 1 does not specify the exact complexity of BFGS for solving problem(9), as the contraction factors r_k are not explicitly given. In the following theorem, we show that for $q \geq 4$, 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 2 *Consider the linear convergence factors $\{r_k\}_{k=0}^{\infty}$ defined in (11) from Theorem 1. If $q \geq 4$, then the sequence $\{r_k\}_{k=0}^{\infty}$ converges to $r_* \in (0, 1)$ that is determined by the equation*

$$r_*^{q-1} + r_*^{q-2} = 1, \tag{12}$$

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

$$|r_k - r_*| \leq (1/2)^k \, |r_0 - r_*|, \qquad \forall k \geq 0. \tag{13}$$

Theorem 2 presented that eventually, the iterations generated by BFGS converge to the optimal solution at the linear rate r_* determined by (12). More specifically, the factors $\{r_k\}_{k=0}^{\infty}$ converge to the fixed point r_* with the linear rate $1/2$. Therefore, the linear convergence factors $\{r_k\}_{k=0}^{\infty}$

and the fixed point r_* are totally determined by the parameter q. They are all independent of the dimension d and the condition number κ_A of the matrix A. Hence, the performance of the BFGS method is not influenced by high-dimensional or ill-conditioned problems. This is different from the common convergence theories of most optimization algorithms, whose performance is deteriorated heavily under the circumstance of high dimension or ill conditioning. We should add that this result is independently important from the 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.

References

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

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

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

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

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

[6] Emmanuel J. Candes, Yonina Eldar, Thomas Strohmer, and Vlad Voroninski. Phase retrieval via matrix completion, 2011.

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

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

[9] WC Davidon. Variable metric method for minimization. Technical report, Argonne National Lab., Lemont, Ill., 1959.

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

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

[12] Xiaoming Chen Feiyan Tian, Lei Liu. Generalized memory approximate message passing. *https://arxiv.org/abs/2110.06069*, 2021.

[13] J. R. Fienup. 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.

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

[15] Roger Fletcher and Michael JD Powell. A rapidly convergent descent method for minimization. *The computer journal*, 6(2):163–168, 1963.

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

[17] Kazimierz Goebel and W. A. Kirk. *Topics in Metric Fixed Point Theory*. Cambridge University Press, 1990.

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

[19] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/hardt16.html.

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

[21] Qiujiang Jin and Aryan Mokhtari. Non-asymptotic superlinear convergence of standard quasi-newton methods. *arXiv preprint arXiv:2003.13607*, 2020.

[22] Qiujiang Jin, Alec Koppel, Ketan Rajawat, and Aryan Mokhtari. Sharpened quasi-newton methods: Faster superlinear rate and larger local convergence neighborhood. *The 39th International Conference on Machine Learning (ICML 2022)*, 2022.

[23] Dachao Lin, Haishan Ye, and Zhihua Zhang. Explicit superlinear convergence of broyden's method in nonlinear equations. *arXiv preprint arXiv:2109.01974*, 2021.

[24] Dachao Lin, Haishan Ye, and Zhihua Zhang. Greedy and random quasi-newton methods with faster explicit superlinear convergence. *Advances in Neural Information Processing Systems 34*, 2021.

[25] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[26] Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16: 559–616, 2015.

[27] Wenlong Mou, Nhat Ho, Martin J Wainwright, Peter Bartlett, and Michael I Jordan. A diffusion process perspective on posterior contraction rates for parameters. *arXiv preprint arXiv:1909.00966*, 2019.

[28] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015. doi: 10.1109/TSP.2015.2448516.

[29] Jorge Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[30] Tongzheng Ren, Fuheng Cui, Alexia Atsidakou, Sujay Sanghavi, and Nhat Ho. Towards statistical and computational complexities of Polyak step size gradient descent. *Artificial Intelligence and Statistics Conference*, 2022.

[31] Anton Rodomanov and Yurii Nesterov. Greedy quasi-newton methods with explicit superlinear convergence. *SIAM Journal on Optimization*, 31(1):785–811, 2021.

[32] Anton Rodomanov and Yurii Nesterov. Rates of superlinear convergence for classical quasi-newton methods. *Mathematical Programming*, pages 1–32, 2021.

[33] Anton Rodomanov and Yurii Nesterov. New results on superlinear convergence of classical quasi-newton methods. *Journal of Optimization Theory and Applications*, 188(3):744–769, 2021.

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

[35] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. 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.

[36] Haishan Ye, Dachao Lin, Zhihua Zhang, and Xiangyu Chang. Explicit superlinear convergence rates of the sr1 algorithm. *arXiv preprintarXiv:2105.07162*, 2021.

[37] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(Apr):899–925, 2013.

Appendix A. Comparison with Newton's method

Next, we compare the convergence results of BFGS for solving problem (9) with the one for Newton's method. The following theorem characterize the global linear convergence of Newton's method with unit step size applied to the objective function in (9).

Theorem 3 *Consider applying Newton's method to optimization problem* (9) *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*

$$\|\theta_k - \hat{\theta}\| = \frac{q-2}{q-1}\|\theta_{k-1} - \hat{\theta}\|, \quad \forall k \geq 1. \tag{14}$$

Moreover, this linear convergence rate $\frac{q-2}{q-1}$ is smaller than the fixed point r_* defined in (12) *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 D.3. The convergence results of Newton's method are also global without any backtracking 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 we illustrate in our numerical results. On the other hand, the computational cost per iteration of Newton's method is $\mathcal{O}(d^3)$ which is much worse than the $\mathcal{O}(d^2)$ of BFGS.

Moving back to our main problem, one important implication of the above convergence results is that in the low signal-to-noise ratio setting the iterates of BFGS converge linearly to the optimal solution of the population least-square loss function, while the contraction coefficient of BFGS is comparable to that of Newton's method which is $(2p-2)/(2p-1)$. For example, 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. We use this intuition to establish a similar result for the finite sample case in the following section.

Appendix B. Statistical rate of BFGS for solving the sample least-square loss

Thus far, we have demonstrated that the BFGS iterates converge linearly to the true parameter θ^* for solving the population least-square loss function $\mathcal{L}$ of the generalized linear models in equation (8). It provides important insight into the behaviors of the BFGS iterates when the sample size n is infinite. In this section, we would like to study the statistical behaviors of the BFGS iterates for solving the least-square loss function $\mathcal{L}_n$, which is given by:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - (X_i^\top \theta)^p)^2. \tag{15}$$

To simplify the proof argument and gain the insight into the statistical behaviors of the BFGS iterates, we focus on the univariate setting, namely, $d = 1$. Note that the statistical behaviors of BFGS iterates still hold in multivariate settings, namely, $d \geq 2$ (See our experiments in Figure 3). When $d = 1$, the BFGS iterates for solving the sample loss function take the following form:

$$\theta_{k+1}^n = \theta_k^n - \eta_k \frac{\theta_k^n - \theta_{k-1}^n}{\nabla \mathcal{L}_n(\theta_k^n) - \nabla \mathcal{L}_n(\theta_{k-1}^n)} \nabla \mathcal{L}_n(\theta_k^n). \tag{16}$$

Throughout this section, we consider the step size $\eta_k = 1$ in the updates (16) of BFGS. We show that the BFGS iterates (16) $\{\theta_k^n\}_{k\geq 0}$ converge to the final statistical radius after a logarithmic number of iterations under the low SNR regimes of the generalized linear models.

Theorem 4 *Consider the low SNR regime of the generalized linear model* (6) *namely,* $|\theta^*| \leq \bar{C}\sigma n^{-1/(2p)}$ *for some constant* $\bar{C}$*, for* $d = 1$*. For any failure probability* $\delta \in (0, 1)$*, if the number of samples is* $n \geq C_1 \log^{2p}(n/\delta)$*, where* C_1 *is a universal constant independent of* n *and* δ*, and the number of iterations satisfies* $k \geq \frac{\log\log(n/\delta) - \frac{\log n}{2p} - \log(|\theta_0^n - \theta_n^*|)}{\log\left(1 - \frac{1}{e(2p-1)}\right)}$*, then with probability* $1 - \delta$ *we have*

$$|\theta_k^n - \theta^*| \leq C_2 \left(\frac{\log^{2p}(n/\delta)}{n}\right)^{\frac{1}{2p}}, \tag{17}$$

where C_2 *is a universal constant independent of* n *and* δ*.*

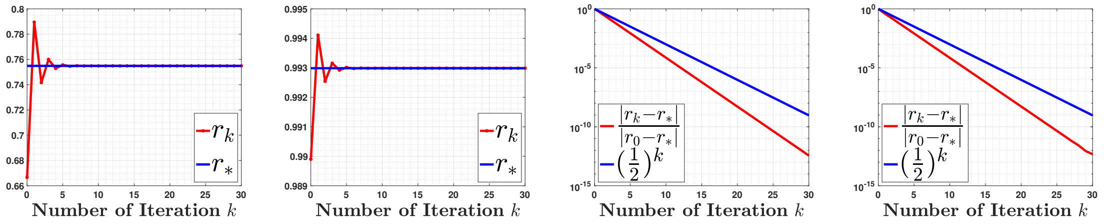
The proof is available in Appendix D.4. As we often start the optimization algorithm with an iterate that is not very close to the solution and hence we have $\theta_0^n = \Theta(1)$. Considering these bounds, the lower bound on the number of iterations can be simplified as $k = \Theta(\frac{2p-1}{2p}\log n)$. This shows that BFGS achieves the statistical accuracy in $O(\log n)$ iterations, which is faster than the sublinear convergence $O(n^{\frac{p-1}{p}})$ of GD shown in [30]. A few comments about Theorem 4 are in order.

Comparing to GD, GD with Polyak step size , and Newton's method: The result of Theorem 4 indicates that under the low SNR regime, the BFGS iterates reach the final statistical radius $\mathcal{O}(n^{-1/(2p)})$ within the true parameter θ^* after $\mathcal{O}(\log(n))$ number of iterations. This complexity indeed is better than the polynomial number of iterations of GD, which is at the order of $\mathcal{O}(n^{(p-1)/p})$ (Corollary 3 in [20]). It is also comparable to the logarithmic number of iterations of GD with Polyak step size which requires $\mathcal{O}(\kappa\log(n))$ iterations (Corollary 1 in [30]), where κ is the condition number of the covriance matrix Σ, and the $\mathcal{O}(\log(n))$ of Newton's method (Corollary 3 in [20]) to reach the similar statistical radius.

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, though the BFGS algorithm and the Newton's method have similar computational complexity in terms of n, BFGS has lower per iteration cost in comparison to Newton's method.

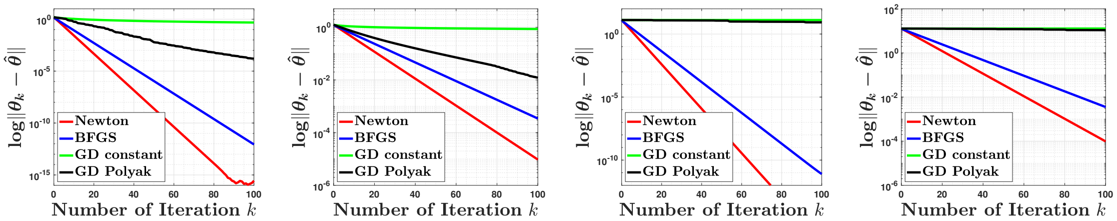
On the minimum number of iterations: The results of BFGS in Theorem 4 involve the minimum number of iterations, namely, these results only hold for some $1 \leq t \leq k$. It suggests that the BFGS iterates may diverge after they reach the final statistical radius under each regime of the generalized linear models. As highlighted in [20], such instability behavior of BFGS is inherent to fast and unstable methods. While it may sound limited, the minimum number of iterations can be overcome 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 3.

Generalization of the results to multivariate settings: While the results of Theorem 4 are only established for the univariate setting, we remark that it is mainly for the simplicity of the proof argument and of the BFGS iterates. In the experiments of Figures 3, we run BFGS for the case that dimension is $d = 4$ and observe that both the statistical radius and the iteration complexities of BFGS are still consistent with those in Theorem 4. We leave a theoretical verification of these results in multivariate settings for the future work.



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

Figure 1: Convergence of factors $\{r_k\}_{k=0}^{\infty}$ to r_*.



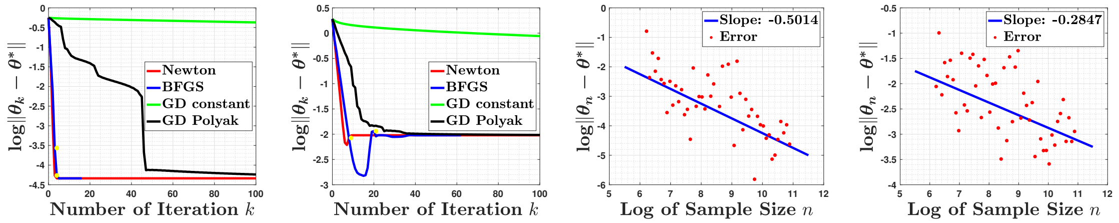
(a) $d = 10, q = 4$. (b) $d = 10, q = 10$. (c) $d = 10^3, q = 4$. (d) $d = 10^3, q = 10$.

Figure 2: Convergence rates of Newton's method, BFGS, GD with constant step size and GD with Polyak step size for different d and q. In plot (a), $m = 100$ and $\eta = 10^{-4}$. In plot (b), $m = 100$ and $\eta = 10^{-8}$. In plot (c), $m = 2000$ and $\eta = 10^{-12}$. In plot (d), $m = 2000$ and $\eta = 10^{-15}$.

Appendix C. Numerical experiments

Numerical experiment 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 (9) which corresponds to the population loss. We choose different values of parameter m, dimension d and the exponential parameter q in (9). 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 2.

We observe that GD with constant step converges very slowly since it can only reach a 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 the BFGS method generate iterations with linear convergence rates, and their linear convergence rates are only affected by the parameter q. 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 the BFGS method, their gap is not significantly large. We should add that these empirical results are consistent with the theoretical results we obtained in Section 4.



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

Figure 3: 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).

Numerical experiment for the empirical loss function. We now move to illustrate the statistical and computational complexities of BFGS for parameter estimation of the generalized linear model. In our experiments, we specifically consider the dimension to be $d = 4$ and the power of the link function to be $p = 2$, namely, we consider the multivariate setting of the phase retrieval problem. The data is generated by first sampling the inputs according to $\{X_i\}_{i=1}^n \sim \mathcal{N}(0, \text{diag}(\sigma_1^2, \cdots, \sigma_4^2))$ where $\sigma_k = (0.5)^{k-1}$, and then generating their labels based on $Y_i = (X_i^\top \theta^*)^2 + \zeta_i$ where $\{\zeta_i\}_{i=1}^n$ 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. Furthermore, for the GD, we choose the step size to be $\eta = 0.1$, while for Newton's method and BFGS, we select the unit stepsize $\eta = 1$.

In (a) and (b) of Figure 3, we consider the sample size $n = 10^4$ and run GD, GD with Polyak step size, BFGS, and Newton's method to find the optimal solution of the sample least-square loss $\mathcal{L}_n$. 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 (a) and (b) of Figure 3 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 (c) and (d) of Figure 3, we run the BFGS when the sample size is 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 $\mathcal{O}(n^{-1/2})$, while under the low SNR regime, its statistical radius becomes $\mathcal{O}(n^{-1/4})$. These empirical results are consistent with the theoretical results of the BFGS in Theorem 4.

Appendix D. Proof of Lemmas and Theorems

Lemma 5 *Consider the objective function in (9) 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 (9) 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

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

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 6 ***Banach's Fixed-Point Theorem.*** *Consider the differentiable function $f : D \subset \mathbb{R} \to D \subset \mathbb{R}$. 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*

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

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

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

Proof Check [17]. ■

D.1. Proof of Theorem 1

We use induction to prove the convergence results in (10) and (11). Note that $b = A\hat{\theta}$ by Assumption 4.1 and the gradient and Hessian of the objective function in (9) 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 5, 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

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

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

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

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

Therefore, we obtain that

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

Condition (10) holds for $k = 1$ with $r_0 = \frac{q-2}{q-1}$. Now we assume that condition (10) 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 $I - \frac{u_{t-1}s_{t-1}^\top}{s_{t-1}^\top u_{t-1}}$ 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 in (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 (10) holds for $k = t + 1$. By induction, we prove the linear convergence results in (10) and (11).

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 (10). 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 (11) and the direction remains all the same.

D.2. Proof of Theorem 2

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

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

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 4. 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

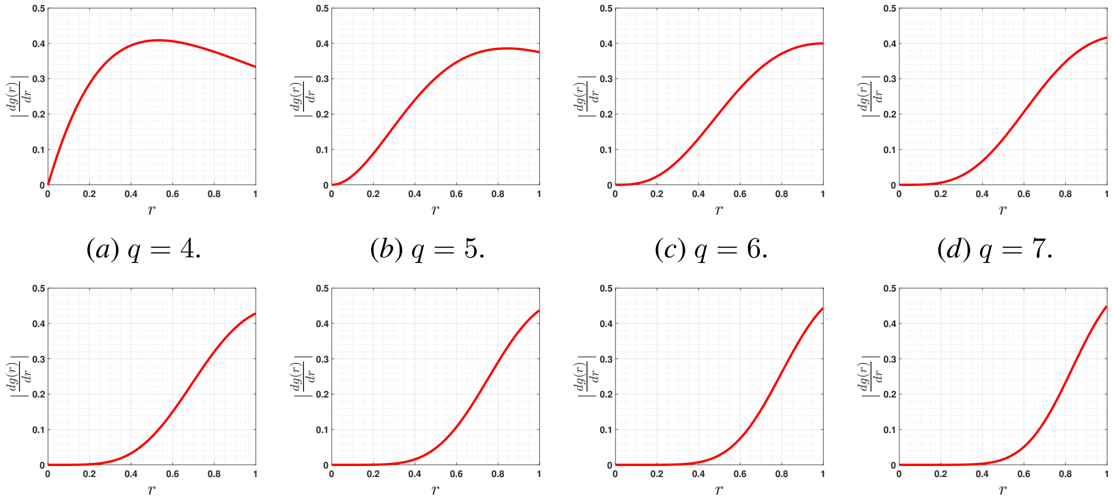
$$|g'(r)| = \frac{(q-1)r^{q-2} - r^{2q-4} - (q-2)r^{q-3}}{(1-r^{q-1})^2} \leq \frac{1}{2}, \tag{54}$$

which is equivalent to

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

Define the function $h(r)$ as

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



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

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 $r \in [0, 1]$,

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

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

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 4, we obtain that $|g'(r)| \leq 1/2$ holds for all $q \geq 4$. Applying Banach's Fixed-Point Theorem from Lemma 6, we prove the final conclusion (13).

D.3. Proof of Theorem 3

Notice that the gradient and the Hessian of the objective function (9) 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 5, 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 $k \geq 1$,

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

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

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

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

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

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

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 $\frac{q-2}{q-1} < r_* < \frac{2q-3}{2q-2}$.

D.4. Proof of Theorem 4

Recall that, we utilize the BFGS for solving the least-square loss function $\mathcal{L}_n$ in equation (15), which is given by:

$$\mathcal{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - (X_i^\top \theta)^p\Big)^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{2}{n}\sum_{i=1}^{n} Y_i (X_i^\top \theta)^p + \frac{1}{n}\sum_{i=1}^{n} (X_i^\top \theta)^{2p}, \tag{74}$$

where $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ that are i.i.d. samples from the following generalized linear model with polynomial link function of power p:

$$Y_i = (X_i^\top \theta^*)^p + \zeta_i,$$

where $\zeta_i \sim \mathcal{N}(0, \sigma^2)$. In this proof, we focus on the low SNR regime of the generalized linear model, namely, $\theta^* = 0$.

Optimal solution θ_n^*: For the case of $d = 1$, we study the optimal solution θ_n^* of the least-square loss function $\mathcal{L}_n$. First of all, directly solving the gradient of the least-square loss function leads to either $\theta_n^* = 0$ or the following form of θ_n^*:

$$(\theta_n^*)^p = \frac{\sum_{i=1}^n Y_i X_i^p}{\sum_{i=1}^n X_i^{2p}}. \tag{75}$$

To bound θ_n^*, we only focus on bounding the later value of θ_n^* in equation (75). Given the generative model of the data, we have

$$\sum_{i=1}^{n} Y_i X_i^p = \left(\sum_{i=1}^{n} X_i^{2p}\right)(\theta^*)^p + \sum_{i=1}^{n} \zeta_i X_i^p.$$

Therefore, we obtain that

$$(\theta_n^*)^p = (\theta^*)^p + \frac{\sum_{i=1}^n \zeta_i X_i^p}{\sum_{i=1}^n X_i^{2p}}.$$

We then consider $\frac{1}{n}\sum_{i=1}^n \zeta_i X_i^p$ and $\frac{1}{n}\sum_{i=1}^n X_i^{2p}$ separately. For the term $\frac{1}{n}\sum_{i=1}^n \zeta_i X_i^p$, note that, for any even integer q, we have

$$\mathbb{E}\left[(\zeta X^p)^q\right] = \mathbb{E}\left[\zeta^q X^{pq}\right] \leq (\sigma^2 q)^{q/2} \cdot (pq)^{pq/2} \leq \big(\max\{\sigma^2, p\} q\big)^{(p+1)q/2}.$$

Invoking Lemma 2 in [27], we have that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\zeta_i X_i^p\right| \geq (4\max\{\sigma^2, p\})^{(p+1)/2}\sqrt{\frac{\log 4/\delta}{n}} + \left(\max\{\sigma^2, p\}\log\frac{n}{\delta}\right)^{(p+1)/2}\frac{\log 4/\delta}{n}\right] \leq \delta.$$

For the term $\frac{1}{n}\sum_{i=1}^{n} X_i^{2p}$, applying Lemma 5 in [10], we have that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i^{2p} - (2p-1)!!\right| \geq \frac{C_p \log^p(n/\delta)}{\sqrt{n}}\right] \leq \delta,$$

where $C_p = \Theta(p\log p)$ is a constant depends only on p.

For the case $\theta^* = 0$ which corresponds to the low signal-to-noise regime, with simple algebra, we know with probability at least $1 - 2\delta$,

$$|\theta_n^*| \leq \left(\frac{(4\max\{\sigma^2, p\})^{(p+1)/2}\sqrt{\frac{\log 4/\delta}{n}} + \left(\max\{\sigma^2, p\}\log\frac{n}{\delta}\right)^{(p+1)/2}\frac{\log 4/\delta}{n}}{(2p-1)!! - \frac{C_p\log^p(n/\delta)}{\sqrt{n}}}\right)^{1/p}.$$

With Minkowski's inequality, it indicates that as long as $n \geq \Theta(\max\{\sigma^2, p^2\log^2 p\}\log^{2p}(n/\delta))$ for some failure probability $\delta \in (0,1)$ to make sure $(4\max\{\sigma^2, p\})^{(p+1)/2}\sqrt{\frac{\log 4/\delta}{n}} \geq \left(\max\{\sigma^2, p\}\log\frac{n}{\delta}\right)^{(p+1)/2}\frac{\log 4/\delta}{n}$ and $\frac{C_p\log^p(n/\delta)}{\sqrt{n}} \leq 1$, if $|\theta^*| \leq C_0\sigma\left(\frac{\log^2 p(n/\delta)}{n}\right)^{1/2p}$, there exists constant $C = \Theta\left(\frac{(\max\{\sigma^2,p\})^{(p+1)/2p}}{p} + C_0\sigma\right)$ such that with probability $1-\delta$, we have $|\theta_n^*| \leq C\left(\frac{\log^{2p}(n/\delta)}{n}\right)^{\frac{1}{2p}}$.

Statistical analysis of the BFGS: Without loss of generality, we assume that θ_n^* takes the value in equation (75) (the proof when $\theta_n^* = 0$ can be argued in the similar fashion). When we run the BFGS for solving the least-square loss function $\mathcal{L}_n$, its updates take the following form

$$\begin{aligned}\theta_{k+1}^n &= \theta_k^n - \frac{\theta_k^n - \theta_{k-1}^n}{\nabla\mathcal{L}(\theta_k^n) - \nabla\mathcal{L}(\theta_{k-1}^n)}\nabla\mathcal{L}(\theta_k^n)\\ &= \theta_k^n - \frac{(\frac{1}{n}\sum_{i=1}^n X_i^{2p})(\theta_k^n)^{2p-1} - (\frac{1}{n}\sum_{i=1}^n Y_i X_i^p)(\theta_k^n)^{p-1}}{(\frac{1}{n}\sum_{i=1}^n X_i^{2p})(\sum_{j=0}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j) - (\frac{1}{n}\sum_{i=1}^n Y_i X_i^p)(\sum_{j=0}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j)}.\end{aligned}$$

Given the formulation of θ_n^* in equation (75), we can rewrite the update of θ_{k+1}^n as follows:

$$\theta_{k+1}^n = \theta_k^n - \frac{(\theta_{k-1}^n - \theta_k^n)(\theta_k^n)^{p-1}((\theta_k^n)^p - (\theta_n^*)^p)}{(\theta_{k-1}^n)^{p-1}((\theta_{k-1}^n)^p - (\theta_n^*)^p) - (\theta_k^n)^{p-1}((\theta_k^n)^p - (\theta_n^*)^p)}. \tag{76}$$

Assume that $|\theta_k^n| \geq 2|\theta_n^*|$ for all $1 \leq k \leq T$ where T indicates the first iteration that the BFGS iterates reach the statistical radius. Without loss of generality, we assume that the initializations θ_0^n and θ_1^n of the BFGS satisfy $\theta_0^n > \theta_1^n \geq 2|\theta_n^*|$, namely, these initializations are positive and lie above the statistical radius (otherwise, the conclusion of the theorem trivially holds). Then, we will first demonstrate by induction that $2|\theta_n^*| < \theta_k^n < \theta_{k-1}^n$ for all $1 \leq k \leq T$. Indeed, the induction hypothesis holds for any $k \leq 1$ based on the conditions of the initializations; we now prove the hypothesis for $k+1$. From equation (76), we have

$$\theta_{k+1}^n = \theta_k^n \left(1 - \frac{A_n}{B_n}\right),$$

where we define

$$\begin{aligned} A_n =& (\theta_{k-1}^n - \theta_k^n)(\theta_k^n)^{p-2}((\theta_k^n)^p - (\theta_n^*)^p), \\ B_n =& (\theta_{k-1}^n)^{p-1}((\theta_{k-1}^n)^p - (\theta_n^*)^p) - (\theta_k^n)^{p-1}((\theta_k^n)^p - (\theta_n^*)^p). \end{aligned}$$

With the induction hypothesis, we can check that $A_n > 0$. Furthermore, direct computation shows that

$$B_n \geq ((\theta_{k-1}^n)^{p-1} - (\theta_k^n)^{p-1})((\theta_k^n)^p - (\theta_n^*)^p) \geq 0.$$

Similarly, by taking the difference between A_n and B_n, we have

$$\begin{aligned} B_n - A_n = & (\theta_{k-1}^n)^{p-1}((\theta_{k-1}^n)^p - (\theta_n^*)^p) - (\theta_k^n)^{p-1}((\theta_k^n)^p - (\theta_n^*)^p) - (\theta_{k-1}^n - \theta_k^n)(\theta_k^n)^{p-2}((\theta_k^n)^p - (\theta_n^*)^p) \\ \geq & ((\theta_{k-1}^n)^{p-1} - (\theta_k^n)^{p-1} - (\theta_{k-1}^n - \theta_k^n)(\theta_k^n)^{p-2})((\theta_k^n)^p - (\theta_n^*)^p) \\ = & \theta_{k-1}^n((\theta_{k-1}^n)^{p-2} - (\theta_k^n)^{p-2})((\theta_k^n)^p - (\theta_n^*)^p) \geq 0. \end{aligned}$$

Putting these results together, it indicates that $0 < A_n/B_n < 1$; therefore, we have $0 < \theta_{k+1}^n < \theta_k^n$, which implies that the induction hypothesis holds for $k+1$. As a consequence, we have $0 < \theta_k^n < \theta_{k-1}^n$ for all $0 \leq k \leq T$.

Now, we divide our proof into two settings: $\theta_n^* > 0$ and $\theta_n^* < 0$.

Setting 1 — When $\theta_n^* > 0$: We will prove the following bound:

$$\theta_{k+1}^n \geq \frac{2p-2}{2p-1}\theta_k^n. \tag{77}$$

Indeed, the inequality (77) is equivalent to show that

$$A_n \leq \frac{1}{2p-1} B_n$$

Note that,

$$\begin{aligned} \frac{B_n}{A_n} =& \frac{(\theta_{k-1}^n)^{p-1}}{(\theta_{k-1}^n - \theta_k^n)(\theta_k^n)^{p-2}} \frac{(\theta_{k-1}^n)^p - (\theta_n^*)^p}{(\theta_k^n)^p - (\theta_n^*)^p} - \frac{\theta_k^n}{\theta_{k-1}^n - \theta_k^n} \\ \geq & \frac{\theta_{k-1}^n}{\theta_{k-1}^n - \theta_k^n} \left(\frac{(\theta_{k-1}^n)^{2p-2}}{(\theta_k^n)^{2p-2}} - 1 \right) \\ \geq & (2p-1)\frac{\theta_{k-1}^n}{\theta_k^n} \\ \geq & (2p-1), \end{aligned}$$

where the first inequality is due to the fact that $\frac{z^p - x^p}{y^p - x^p}$ monotonically increases with x for $0 < x < y < z$, the second inequality is due to the convexity of the function t^{2p-1} for $t > 0$ and the last one is due to the fact that $\theta_{k-1}^n > \theta_k^n$.

Now, we would like to demonstrate the linear convergence rate of the BFGS iterates $\{\theta_k^n\}$ to the optimal solution θ_n^*:

$$|\theta_{k+1}^n - \theta_n^*| \leq \left(1 - \frac{1}{e(2p-2)}\right) |\theta_k^n - \theta_n^*|, \tag{78}$$

for all $2 \leq k \leq T-1$. As $\theta_k^n > |\theta_n^*|$ for all $2 \leq k \leq T$, the result in equation (78) is equivalent to

$$(\theta_{k+1}^n - \theta_n^*) \leq \left(1 - \frac{1}{e(2p-2)}\right) (\theta_k^n - \theta_n^*), \tag{79}$$

for all $2 \leq k \leq T-1$. From direct computation, we have

$$\begin{aligned} \theta_{k+1}^n - \theta_n^* = (\theta_k^n - \theta_n^*) & \\ \cdot \left(1 - \frac{(\theta_{k-1}^n - \theta_k^n)(\theta_k^n)^{p-1}((\theta_k^n)^p - (\theta_n^*)^p)}{(\theta_k^n - \theta^*)((\theta_{k-1}^n)^{p-1}((\theta_{k-1}^n)^p - (\theta_n^*)^p) - (\theta_k^n)^{p-1}((\theta_k^n)^p - (\theta_n^*)^p))}\right) & . \end{aligned}$$

As $0 < 2|\theta_n^*| < \theta_k^n < \theta_{k-1}^n$, we have

$$\begin{aligned} & \frac{\theta_k^n - \theta_n^*}{\theta_{k-1}^n - \theta_k^n} \left(\frac{(\theta_{k-1}^n)^{p-1}((\theta_{k-1}^n)^p - (\theta_n^*)^p)}{(\theta_k^n)^{p-1}((\theta_k^n)^p - (\theta_n^*)^p)} - 1 \right) \\ = & \frac{\theta_k^n - \theta_n^*}{\theta_{k-1}^n - \theta_k^n} \cdot \frac{(\theta_{k-1}^n)^{2p-1} - (\theta_n^k)^{2p-1} - ((\theta_{k-1}^n)^{p-1} - (\theta_k^n)^{p-1})(\theta_n^*)^p}{(\theta_k^n)^{p-1}((\theta_k^n)^p - (\theta_n^*)^p)} \\ \leq & \frac{(\theta_{k-1}^n)^{2p-1} - (\theta_k^n)^{2p-1}}{(\theta_{k-1}^n - \theta_k^n)(\theta_k^n)^{2p-2}} \\ \leq & (2p-1) \left(\frac{\theta_{k-1}^n}{\theta_k^n} \right)^{2p-2} \\ \leq & (2p-1) \left(1 + \frac{1}{2p-2} \right)^{2p-2} \\ \leq & e(2p-1), \end{aligned}$$

where we obtain the first inequality as the objective is monotonically decreasing with respect to θ_n^*.

Hence, as long as $2 \leq k \leq T-1$ we find that

$$(\theta_{k+1}^n - \theta_n^*) \leq \left(1 - \frac{1}{e(2p-1)}\right) (\theta_k^n - \theta_n^*),$$

which shows the linear convergence of the BFGS iterates to the optimal solution.

By repeating the inequalities (79), we obtain that

$$|\theta_T^n - \theta_n^*| \leq \left(1 - \frac{1}{e(2p-1)}\right)^{T-1} |\theta_0^n - \theta_n^*|. \tag{80}$$

As long as we choose $\left(1 - \frac{1}{e(2p-1)}\right)^T |\theta_0^n - \theta_n^*| \leq \left(\frac{\log^{2p}(n/\delta)}{n}\right)^{1/2p}$, which is equivalently to

$$T \geq \frac{\log\log(n/\delta) - \frac{\log n}{2p} - \log(|\theta_0^n - \theta_n^*|)}{\log\left(1 - \frac{1}{e(2p-1)}\right)},$$

which is satisfied according to the hypothesis. If this condition holds, then we obtain $|\theta_T^n - \theta_n^*| \leq |\theta_n^*|$. Hence, for any k larger than the above threshold our result holds. A direct application of the triangle inequality, we have

$$\begin{aligned} |\theta_T^n - \theta^*| \leq |\theta_T^n - \theta_n^*| + |\theta_n^*| &\leq \left(\frac{\log^{2p}(n/\delta)}{n}\right)^{1/2p} + |\theta_n^* - \theta^*| \\ &\leq \left(\frac{\log^{2p}(n/\delta)}{n}\right) + |\theta_n^*| + |\theta^*| \\ &\leq (\bar{C}\sigma + C + 1)\left(\frac{\log^{2p}(n/\delta)}{n}\right)^{\frac{1}{2p}} \end{aligned} \tag{81}$$

with probability $1-\delta$, where the final inequality is due to $|\theta_n^*| \leq C\left(\frac{\log^{2p}(n/\delta)}{n}\right)^{\frac{1}{2p}}$ and $|\theta^*| \leq \bar{C}\sigma\left(\frac{\log^{2p}(n/\delta)}{n}\right)^{\frac{1}{2p}}$ where C is the constant from the bound of θ_n^* and $\bar{C}$ is a constant from the hypothesis of the low SNR regime. As a consequence, by choosing $C_1 = \Theta(\max\{\sigma^2, p^2\log^2 p\})$ and $C_2 = \Theta\left(\frac{(\max\{\sigma^2, p\})^{(p+1)/2p}}{p} + C_0 + \bar{C}\sigma\right)$ from the bound on θ_n^*, we obtain the conclusion of the theorem when $\theta_n^* > 0$.

Setting 2 — When $\theta_n^* < 0$: We would like to prove the following bound:

$$\theta_{k+1}^n \geq \frac{p}{p+2}\theta_k^n. \tag{82}$$

Indeed, the inequality (82) is equivalent to show that

$$B_n - A_n \geq \frac{p}{p+2}B_n.$$

In light of the previous calculations, this inequality is equivalent to

$$\begin{aligned} &\sum_{j=1}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j[(\theta_k^n)^p - (\theta_n^*)^p] + \sum_{j=p-1}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j \\ &\qquad\geq \frac{p}{p+2}\left(\sum_{j=0}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j[(\theta_k^n)^p - (\theta_n^*)^p] + \sum_{j=p-1}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j\right), \end{aligned}$$

which is equivalent to

$$2\sum_{j=1}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j[(\theta_k^n)^p - (\theta_n^*)^p] + 2\sum_{j=p-1}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j \geq p(\theta_k^n)^{p-2}[(\theta_k^n)^p - (\theta_n^*)^p]. \tag{83}$$

Since $\sum_{j=1}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j \geq (p-2)(\theta_k^n)^{p-2}$ and $\theta_k^n > 2|\theta_n^*| > 0$, we have that

$$\sum_{j=1}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j[(\theta_k^n)^p - (\theta_n^*)^p] \geq (p-2)(\theta_k^n)^{p-2}[(\theta_k^n)^p - (\theta_n^*)^p].$$

Furthermore, as $-(\theta_n^*)^p < (\theta_k^n)^p/2^p$, we have $(\theta_k^n)^{p-2}[(\theta_k^n)^p - (\theta_n^*)^p] < (1+2^{-p})(\theta_k^n)^{2p-2}$. On the other hand, $\sum_{j=p-1}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j > p(\theta_k^n)^{2p-2} \geq (1+2^{-p})(\theta_k^n)^{2p-2} > (\theta_k^n)^{p-2}[(\theta_k^n)^p - (\theta_n^*)^p]$ as $p \geq 2$. Putting these results together indicates that the bound (83) holds. Therefore, we obtain the conclusion of the bound (83) and equivalently the bound (82).

Now, we would like to demonstrate the linear convergence rate of the BFGS iterates $\{\theta_k^n\}$ to the optimal solution θ_n^*:

$$|\theta_{k+1}^n - \theta_n^*| \leq \left(1 - \frac{\left(1-\frac{2}{p+2}\right)^{2p-2}}{2p^2-p}\right)|\theta_k^n - \theta_n^*|, \tag{84}$$

for all $2 \leq k \leq T-1$. As $\theta_k^n > 2|\theta_n^*|$ for all $2 \leq k \leq T$, the result in equation (84) is equivalent to

$$(\theta_{k+1}^n - \theta_n^*) \leq \left(1 - \frac{\left(1-\frac{2}{p+2}\right)^{2p-2}}{2p^2-p}\right)(\theta_k^n - \theta_n^*), \tag{85}$$

for all $2 \leq k \leq T-1$. From direct computation, we have

$$\begin{aligned}\theta_{k+1}^n - \theta_n^* &= (\theta_k^n - \theta_n^*)\\ &\cdot\left(1 - \frac{(\theta_k^n)^{p-1}(\sum_{j=0}^{p-1}(\theta_k^n)^{p-1-j}(\theta_n^*)^j)}{(\sum_{j=0}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j) - (\sum_{j=0}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j)(\theta_n^*)^p}\right).\end{aligned}$$

As $0 < 2|\theta_n^*| < \theta_k^n < \theta_{k-1}^n$, we have

$$\begin{aligned}&\left(\sum_{j=0}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j\right) - \left(\sum_{j=0}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j\right)(\theta_n^*)^p\\ &=\sum_{j=0}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j[(\theta_k^n)^p - (\theta_n^*)^p] + \sum_{j=p-1}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j\\ &=\sum_{j=0}^{p-2}(\theta_k^n)^{p-2-j}(\theta_{k-1}^n)^j\left[(\theta_k^n - \theta_n^*)\left(\sum_{l=0}^{p-1}(\theta_k^n)^l(\theta_n^*)^{p-1-l}\right)\right] + \sum_{j=p-1}^{2p-2}(\theta_k^n)^{2p-2-j}(\theta_{k-1}^n)^j\\ &\leq p(p-1)(\theta_{k-1}^n)^{2p-3}(\theta_k^n - \theta_n^*) + p(\theta_{k-1}^n)^{2p-2}\\ &\leq (2p^2-p)(\theta_{k-1}^n)^{2p-2}.\end{aligned}$$

Meanwhile, when $2 \leq k \leq T$ we have that

$$(\theta_k^n)^{p-1}\left(\sum_{j=0}^{p-1}(\theta_k^n)^{p-1-j}(\theta_n^*)^j\right) \geq (\theta_k^n)^{2p-2} \geq \left(1 - \frac{2}{p+2}\right)^{2p-2}(\theta_{k-1}^n)^{2p-2}.$$

Putting the above results together, as long as $2 \leq k \leq T-1$ we find that

$$(\theta^n_{k+1} - \theta^*_n) \leq \left(1 - \frac{\left(1 - \frac{2}{p+2}\right)^{2p-2}}{2p^2 - p}\right)(\theta^n_k - \theta^*_n),$$

which shows the linear convergence of the BFGS iterates to the optimal solution.

By repeating the inequalities (85), we obtain that

$$|\theta^n_T - \theta^*_n| \leq \left(1 - \frac{\left(1 - \frac{2}{p+2}\right)^{2p-2}}{2p^2 - p}\right)^T |\theta^n - \theta^*_n|.$$

As long as we choose $\left(1 - \frac{\left(1 - \frac{2}{p+2}\right)^{2p-2}}{2p^2 - p}\right)^T (\theta^n_0 - \theta^*_n) \leq \left(\frac{\log^{2p}(n/\delta)}{n}\right)^{1/2p}$, which is equivalent to $T \geq \frac{\log\log(n/\delta) - \frac{\log n}{2p} - \log(|\theta^n_0 - \theta^*_n|)}{\log\left(1 - \frac{\left(1 - \frac{2}{p+2}\right)^{2p-2}}{2p^2-p}\right)}$, then we obtain $|\theta^n_T - \theta^*_n| \leq |\theta^*_n|$. It is satisfied as $T \geq \frac{\log\log(n/\delta) - \frac{\log n}{2p} - \log(|\theta^n_0 - \theta^*_n|)}{\log\left(1 - \frac{1}{e(2p-1)}\right)} > \frac{\log\log(n/\delta) - \frac{\log n}{2p} - \log(|\theta^n_0 - \theta^*_n|)}{\log\left(1 - \frac{\left(1 - \frac{2}{p+2}\right)^{2p-2}}{2p^2-p}\right)}$ from the hypothesis. A direct application of the triangle inequality, we have

$$|\theta^n_T - \theta^*| \leq |\theta^n_T - \theta^*_n| + |\theta^*_n - \theta^*| \leq \left(\frac{\log^{2p}(n/\delta)}{n}\right)^{1/2p} + |\theta^*_n - \theta^*| \leq (\bar{C}\sigma + C + 1)\left(\frac{\log^{2p}(n/\delta)}{n}\right)^{\frac{1}{2p}}$$

with probability $1-\delta$. Here, C is the constant from the bound of θ^*_n and $\bar{C}$ is a constant from the hypothesis of the low SNR regime. As a consequence, by choosing $C_1 = \Theta(\max\{\sigma^2, p^2\log^2 p\})$ and $C_2 = \Theta\left(\frac{(\max\{\sigma^2, p\})^{(p+1)/2p}}{p} + C_0 + \bar{C}\sigma\right)$ from the bound on θ^*_n, we obtain the conclusion of the theorem when $\theta^*_n < 0$.

Appendix E. Additional Experiment Results

E.1. Statistical Radius with Error Bar

We additionally report the statistical radius with error bar, by sampling 40 different dataset with the same generative model and report the median, the 25% percentile and 75% percentile. For low SNR regime, BFGS still reach the statistical radius within $\mathcal{O}(n^{-1/4})$, while for high SNR regime, BFGS reach the statistical radius within $\mathcal{O}(n^{-1/2})$.

E.1.1. Experiments in High Dimension

To show that BFGS can also be applied to high dimension scenarios, we conduct additional experiments on the generalized linear model with input $d = 50$ and the power of link function $p = 2$. The inputs are generated by $\{X_i\}_{i=1}^n \sim \mathcal{N}(0, \text{diag}(\sigma_1^2, \cdot, \sigma_{50}^2))$ where $\sigma_k = (0.96)^{k-1}$, and the remaining setting and hyper-parameters are set identical to the low dimension scenarios.

The results are shown in Figure 6. As ther results show, the performance of BFGS in high dimensional scenarios are nearly identical to the low dimensional scenarios.

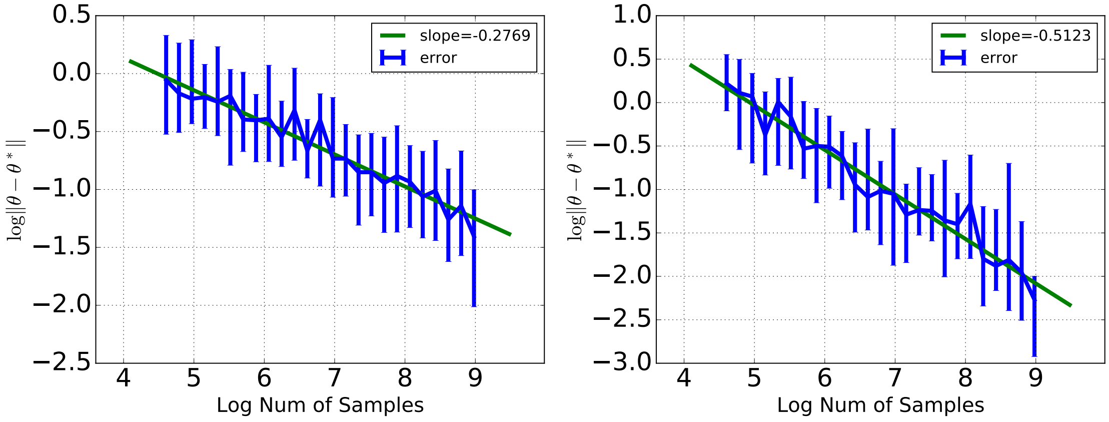
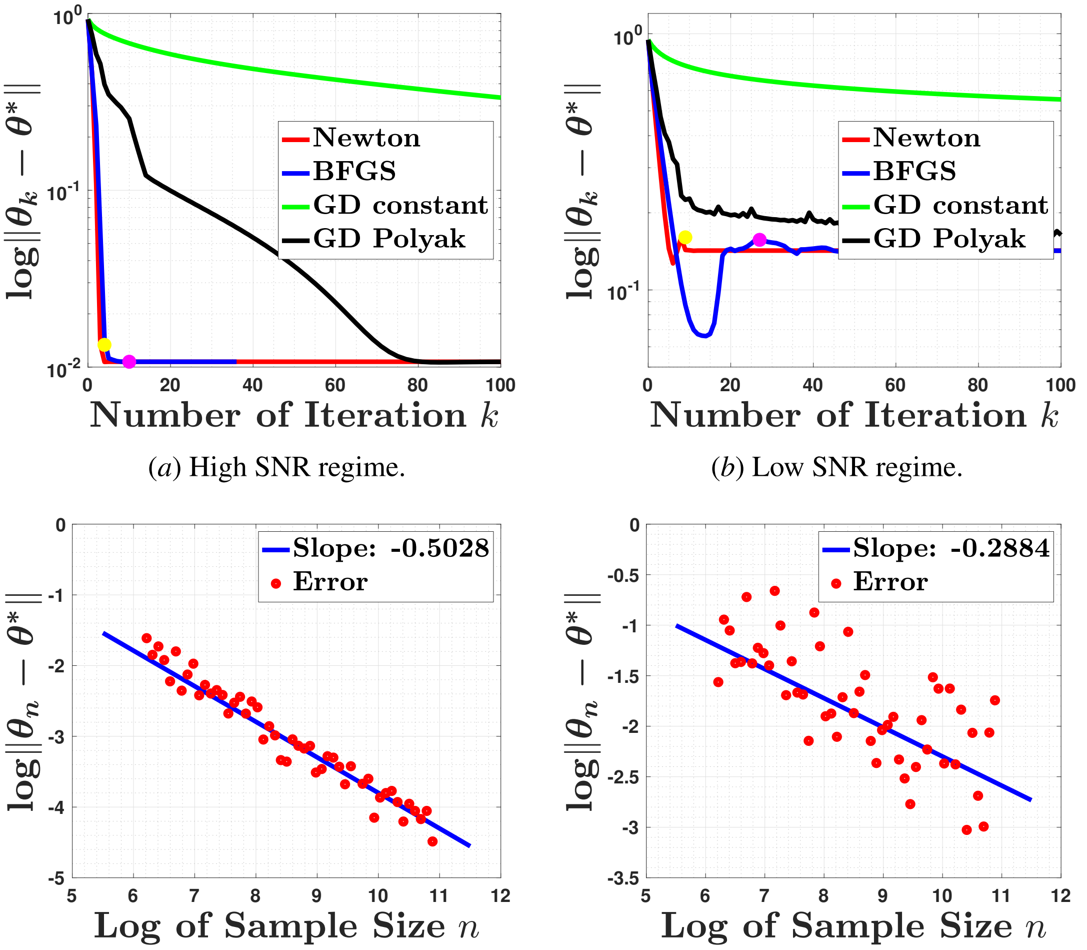


Figure 5: Illustration of the statistical radius of BFGS with error bar. **Left:** Low SNR regime. **Right:** High SNR regime.

E.1.2. Experiments in Middle SNR Regime

Here we briefly illustrate the behavior of BFGS in Middle SNR regime. We consider the generalized linear model with $d = 50$ and $p = 2$. The inputs are still generated by $\{X_i\}_{i=1}^n$, but θ^* now is uniformly sampled from the sphere with radius $n^{-1/6}$.

The results are shown in Figure 7. We can see BFGS still converges linearly, and the statistical radius of middle SNR regime lies between the Hign SNR and Low SNR. A rigorous characterization of the statistical radius of middle SNR regime will be left as future work.



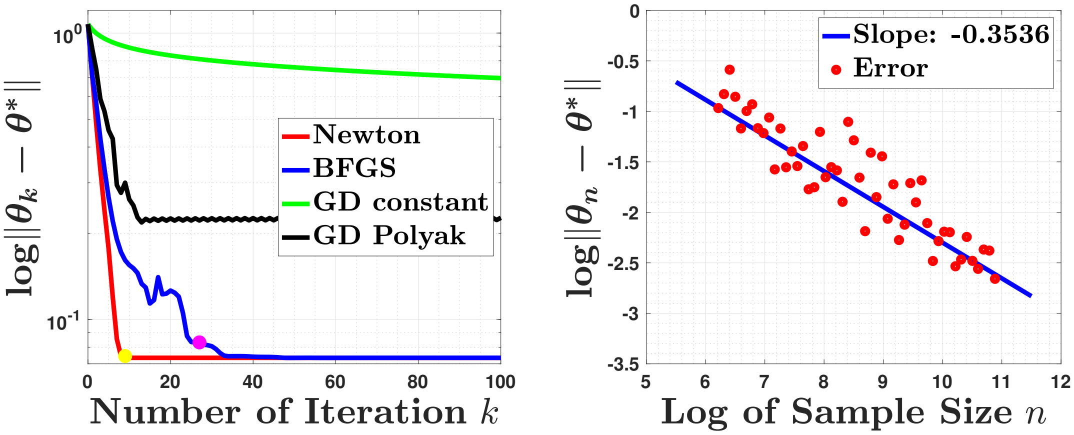
(a) High SNR regime.

(b) Low SNR regime.

(c) High SNR regime.

(d) Low SNR regime.

Figure 6: Convergence and Statistical Results in $d = 50$. 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) correspondingly.



(a) Middle SNR regime.

(b) Middle SNR regime.

Figure 7: Convergence and Statistical Results for Medium SNR in $d = 50$.