# Towards Statistical and Computational Complexities of Polyak Step Size Gradient Descent

Tongzheng Ren[⋆,◇,‡] Fuheng Cui[⋆,♭] Alexia Atsidakou[⋆,†] Sujay Sanghavi[†] Nhat Ho[♭,‡]

Department of Computer Science, University of Texas at Austin[◇],
Department of Statistics and Data Sciences, University of Texas at Austin[♭]
Department of Electrical and Computer Engineering, University of Texas at Austin[†],
October 15, 2021

**Abstract**

We study the statistical and computational complexities of the Polyak step size gradient descent algorithm under generalized smoothness and Łojasiewicz conditions of the population loss function, namely, the limit of the empirical loss function when the sample size goes to infinity, and the stability between the gradients of the empirical and population loss functions, namely, the polynomial growth on the concentration bound between the gradients of sample and population loss functions. We demonstrate that the Polyak step size gradient descent iterates reach a final statistical radius of convergence around the true parameter after logarithmic number of iterations in terms of the sample size. It is computationally cheaper than the polynomial number of iterations on the sample size of the fixed-step size gradient descent algorithm to reach the same final statistical radius when the population loss function is not locally strongly convex. Finally, we illustrate our general theory under three statistical examples: generalized linear model, mixture model, and mixed linear regression model.

## 1 Introduction

From its origin in mathematics, gradient descent algorithm [32, 5, 30] has played a central role in large-scale machine learning and data science applications. In general unconstrained settings, this algorithm can be used for finding optimal solutions of optimization problems of the following form:

$$\min_{\theta \in \mathbb{R}^d} f_n(\theta). \tag{1}$$

Here, $n$ stands for the sample size of i.i.d. data $X_1, X_2, \ldots, X_n$ coming from an unknown distribution $P_{\theta^*}$ where $\theta^*$ is true but unknown parameter and $f_n$ is a given empirical loss function whose optimal solutions, denoted by $\widehat{\theta}_n$, can be used to approximate the true parameter $\theta^*$. While the difference between $\widehat{\theta}_n$ and $\theta^*$ had been studied extensively in the literature via several tools from the empirical process theory, the convergence rates of $\theta_n^t$, updates from the gradient descent algorithm, to optimal neighborhood around the true parameter $\theta^*$, has still remained a nascent topic.

A natural approach to analyze the difference between the updates $\theta_n^t$ and the true parameter $\theta^*$ is to study the convergence rate of $\theta_n^t$ to $\widehat{\theta}_n$, stationary points of optimization problem (1), and the gap between $\widehat{\theta}_n$ and $\theta^*$, namely, we use the following triangle inequality:

$$\|\theta_n^t - \theta^*\| \leq \|\theta_n^t - \widehat{\theta}_n\| + \|\widehat{\theta}_n - \theta^*\|. \tag{2}$$

⋆ Alexia Atsidakou, Fuheng Cui and Tongzheng Ren contributed equally to this work.

‡ Correspondence to: Tongzheng Ren (tongzheng@utexas.edu) and Nhat Ho (minhnhat@utexas.edu).

This approach is often referred to as *direct approach* and has been used in several earlier works (e.g., [1, 39, 26, 8]). However, to ensure that the radius of convergence for $\|\theta_n^t - \theta^*\|$ is at the order of final statistical rate, we need to obtain a tight optimization convergence rate of the term $\|\theta_n^t - \widehat{\theta}_n\|$ based on the sample size $n$ and the number of iterations $t$. It requires a precise understanding of the noise-structure in the gradient of the empirical loss function, which is generally non-trivial to study in practice.

To circumvent the challenges of the direct analysis (2), a popular approach to analyze the difference between the updates $\theta_n^t$ and the true parameter $\theta^*$ is the *population to sample analysis* [38, 14, 2, 23, 37, 7, 12, 11, 16, 24]. In particular, we define the corresponding population version of optimization problem (3) as follows:

$$\min_{\theta \in \mathbb{R}^d} f(\theta), \tag{3}$$

where $f(\cdot) := \mathbb{E}_{X^n}[f_n(\cdot)]$ is the population loss function and $X^n = (X_1, \ldots, X_n)$. When the step size $\eta$ of the gradient descent algorithm is fixed, which we refer to as *fixed-step size gradient descent algorithm*, the idea of the population to sample analysis is to analyze the radius of convergence of $\theta_n^t$ via the following triangle inequality:

$$\|\theta_n^{t+1} - \theta^*\| \leq \|F_{\mathrm{GD}}(\theta_n^t) - \theta^*\| + \eta \|\nabla f_n(\theta_n^t) - \nabla f(\theta_n^t)\| := A + B, \tag{4}$$

where $F_{\mathrm{GD}}(\theta) := \theta - \eta \nabla f(\theta)$ is the corresponding population operator of the fixed-step size gradient descent algorithm. The bound (4) suggests that we can relate the behaviors of the sample fixed-step size gradient descent iterate $\theta_n^{t+1}$ to two terms: (i) Term A: the convergence rate of gradient descent iterates for solving population loss function (3); (ii) Term B: the uniform concentration of $\nabla f_n(\theta)$ around $\nabla f(\theta)$ when $\theta$ lies in a certain neighborhood around $\theta^*$.

**Complexity of fixed-step size gradient descent:** When the population loss function is locally strongly convex and smooth around $\theta^*$, under the local initialization the convergence rate of gradient descent iterates for solving the population loss function is linear, i.e., the term A in equation (4) behaves like $\kappa \|\theta_n^t - \theta^*\|$ where $\kappa < 1$ is some constant. When the deviation bound between $\nabla f_n(\theta)$ and $\nabla f(\theta)$ is at the order $\varepsilon(n, \delta)$ with probability $1 - \delta$ as long as $\|\theta - \theta^*\| \leq r$ where $\varepsilon(n, \delta)$ is the noise function, the statistical radius of the sample fixed-step size gradient descent updates is at the order of $\mathcal{O}(\varepsilon(n, \delta))$ as long as the number of iterations is at least $\mathcal{O}(\log(1/\varepsilon(n, \delta)))$. For practical high dimensional statistical models, the noise function $\varepsilon(n, \delta)$ is at the order of $\sqrt{d/n}$ (here we skip $\delta$ for simplicity); therefore, we have parametric statistical radius of the sample gradient descent iterates after $\log(n/d)$ number of iterations.

When the population loss function is no longer locally strongly convex around the true parameter $\theta^*$, analyzing the convergence rate of $\theta_n^t$ is non-trivial as simply applying triangle inequality in equation (4) can get to sub-optimal rate. To get a sharp statistical radius of $\theta_n^t$, Ho et al. [16] recently utilize a localization argument from the empirical process theory to progressively balance the two terms A and B when the sample fixed-step size gradient descent updates $\theta_n^t$ move closer to the true parameter $\theta^*$. They show that when the convergence rate of the population fixed-step size gradient descent iterates is at the order of $\mathcal{O}(1/t^{1/\alpha})$ for some $\alpha > 0$ and the deviation bound between $\nabla f_n(\theta)$ and $\nabla f(\theta)$ is slow and at the order of $\mathcal{O}(r^\gamma \varepsilon(n, \delta))$ with probability $1 - \delta$ as long as $\|\theta - \theta^*\| \leq r$ where $\gamma \geq 0$, the final statistical radius of the fixed-step size gradient descent iterates $\|\theta_n^t - \theta^*\|$ is upper bounded by $\mathcal{O}(\varepsilon(n, \delta)^{\frac{1}{1+\alpha-\gamma}})$ as long as $t \geq \mathcal{O}(\varepsilon(n, \delta)^{-\frac{\alpha}{\alpha+1-\gamma}})$ and $\alpha \geq \gamma$. In practical high dimensional

statistical models, the noise function $\varepsilon(n, \delta)$ is proportional to $\sqrt{d/n}$; therefore, the required number of iterations for the fixed-step size gradient descent updates to reach the final radius is proportional to $(n/d)^{\frac{\alpha}{\alpha+1-\gamma}}$. Since each iteration of the gradient descent requires $\mathcal{O}(nd)$ arithmetic operations, the total computational complexity for the fixed-step size gradient descent algorithm to reach the final statistical radius is of the order of $\mathcal{O}(n^{\frac{\alpha}{\alpha+1-\gamma}+1})$ for fixed dimension $d$. It is much more computationally expensive than the optimal computational complexity $\mathcal{O}(n)$ when the sample size is sufficiently large in practice.

**Contribution.** In this paper, we show that by using Polyak step size gradient descent method [32], an adaptive gradient descent algorithm, we can overcome the high computational complexity of the fixed-step size gradient descent algorithm for reaching the final statistical radius when the population loss function is not locally strongly convex. Our contribution is two-fold and can be summarized as follows:

1. **Complexity of Polyak step size gradient descent algorithm**: We study the computational and statistical complexities of the Polyak step size gradient descent iterates under the generalized smoothness and Łojasiewicz properties of the population loss function, which are characterized by parameter $\alpha \geq 0$. Under these assumptions, we demonstrate that the population Polyak step size gradient descent iterates have a linear convergence rate to the true parameter $\theta^*$. When the deviation bound between the gradients of sample and population loss functions is growing at the order of $\mathcal{O}(r^\gamma \varepsilon(n, \delta))$ with probability $1-\delta$, we further prove that the sample Polyak step size gradient descent updates reach the final statistical radius $\mathcal{O}(\varepsilon(n, \delta)^{\frac{1}{1+\alpha-\gamma}})$ around the true parameter $\theta^*$ as long as $t \geq \mathcal{O}(\log(1/\varepsilon(n, \delta)))$. It indicates that the sample Polyak step size gradient descent iterates reach the same final statistical radius as that of the fixed-step size gradient descent iterates and they only require a logarithmic number of iterations, which is much smaller than those from the fixed-step size gradient descent updates. Since each iteration of the Polyak step size gradient descent algorithm only requires $\mathcal{O}(nd)$ arithmetic operations, the total computational complexity for the Polyak step size algorithm to reach the final statistical radius is at the order of $\mathcal{O}(n \log(1/\varepsilon(n, \delta)))$ for fixed dimension $d$, which is much cheaper than $\mathcal{O}(n \cdot \varepsilon(n, \delta)^{-\frac{\alpha}{\alpha+1-\gamma}})$ from the fixed-step size gradient descent algorithm. See Table 1 for a more detailed comparison between the Polyak step size and fixed-step size methods.

2. **Illustrative examples**: We illustrate the general theory under three statistical models: generalized linear model, symmetric two-component mixture model, and mixed linear regression model. For the generalized linear model with link function $g(x) = x^p$ where $p \in \mathbb{N}$ and $p \geq 2$, we demonstrate that when we have no signal, i.e., $\theta^* = 0$, the Polyak step size gradient descent iterates converge to a radius of convergence $\mathcal{O}((d/n)^{1/2p})$ around the true parameter after $\mathcal{O}(\log(n/d))$ number of iterations. It is much faster than the required number of iterations $\mathcal{O}((n/d)^{\frac{p-1}{p}})$ of the fixed-step size gradient descent algorithm. For both the symmetric two-component mixture model and mixed linear regression, under the low signal-to-noise regime, e.g., $\theta^* = 0$, we prove that the final optimal statistical radius of the Polyak step size iterates are at the order of $\mathcal{O}((d/n)^{1/4})$ as long as we run the algorithm for $\mathcal{O}(\log(n/d))$ iterations, which is faster than $\mathcal{O}(\sqrt{n/d})$ number of iterations required for the EM algorithm, which in these settings is equivalent to gradient descent with step size 1, in order to reach the same final statistical radii.

**Organization.** The paper is organized as follows. In Section 2, we first introduce our as-

sumptions on generalized smoothness and Łojasiewicz property of the population loss function and the growth condition on the concentration of the gradient of sample loss function around the gradient of the population loss function. Then, we establish convergence rates of the Polyak step size gradient descent iterates under these assumptions. In Section 3, we illustrate these convergence rates under specific settings of generalized linear model, mixture model, and mixed linear regression. We carry out experiments in Section 4 to verify the convergence rates studied in Section 3 while concluding the paper with a few discussions in Section 6. Proofs of main results are in Section 5 while proofs of the remaining results are deferred to the Appendices.

**Notation.** For any matrix $A \in \mathbb{R}^{d\times d}$, we denote by $\lambda_{\max}(A)$ the maximum eigenvalue of the matrix A. For any $x \in \mathbb{R}^d$, $\|x\|$ denotes the $\ell_2$ norm of $x$. For any two sequences $\{a_n\}_{n\geq 1}, \{b_n\}_{n\geq 1}$, we denote $a_n = \mathcal{O}(b_n)$ to mean that $a_n \leq Cb_n$ for all $n \geq 1$ where $C$ is some universal constant. Furthermore, we denote $a_n = \Theta(b_n)$ to indicate that $C_1 b_n \leq a_n \leq C_2 b_n$ for any $n \geq 1$ where $C_1, C_2$ are some universal constants.

## 2 Polyak Step Size Gradient Descent

In this section, we first provide a set of assumptions used in our analysis of the Polyak step size gradient descent algorithm in Section 2.1. We then study the convergence rate of that algorithm under these assumptions in Section 2.2.

### 2.1 Assumptions

We first start with the following assumption about the local generalized smoothness of the population loss function in equation (3).

(W.1) (Generalized Smoothness) There exists a constant $\alpha \geq 0$ such that for all $\theta \in \mathbb{B}(\theta^*, \rho)$ for some radius $\rho > 0$, we have

$$\lambda_{\max}(\nabla^2 f(\theta)) \leq c_1 \|\theta - \theta^*\|^{\alpha},$$

where $c_1 > 0$ is some universal constant.

When $\alpha = 0$, Assumption (W.1) corresponds to the standard local smoothness condition. When $\alpha > 0$, Assumption (W.1) provides a polynomial growth condition on the Lipschitz constant when the parameter lies in some neighborhood around the true parameter $\theta^*$. An example of the function $f$ that satisfies Assumption (W.1) is $f(\theta) = \sum_{i=1}^{d} \theta_i^{2\alpha_i}$ for all $\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \mathbb{R}^d$ where $\alpha_1, \alpha_2, \ldots, \alpha_d \geq 1$ are some given positive integers. In this simple example, the true parameter $\theta^* = 0$ and the constant $\alpha$ in Assumption (W.1) takes the value $\alpha = \min_{1\leq i\leq d}\{2\alpha_i - 2\}$.

Now, to obtain a convergence rate for the Polyak step size gradient descent algorithm for solving the minima of the population loss function, we need another assumption, which we refer to as *generalized Łojasiewicz property*, on the growth of the gradient of the population loss function $f$.

(W.2) (Generalized Łojasiewicz Property) For all $\theta \in \mathbb{B}(\theta^*, \rho)$ for some radius $\rho > 0$, there exists a constant $\alpha \geq 0$ such that we have

$$\|\nabla f(\theta)\| \geq c_2 (f(\theta) - f(\theta^*))^{1-\frac{1}{\alpha+2}}$$

where $c_2 > 0$ is some universal constant.

| Method | Smoothness (W.1), Łojasiewicz (W.2) | Concentration Bound (W.3) | Number of Iterations | Statistical Radius |
|---|---|---|---|---|
| Fixed-step size gradient descent (Proposition 1) | $\alpha > 0$ | $\gamma \geq 0$ | $\varepsilon(n,\delta)^{-\frac{\alpha}{1+\alpha-\gamma}}$ | $\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$ |
| | $\alpha = 0$ | $\gamma = 0$ | $\log(1/\varepsilon(n,\delta))$ | $\varepsilon(n,\delta)$ |
| Polyak step size gradient descent (Theorem 1) | $\alpha \geq 0$ | $\gamma \geq 0$ | $\log(1/\varepsilon(n,\delta))$ | $\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$ |

**Table 1.** An overview of the convergence rates of fixed-step size and Polyak step size gradient descent iterates under the assumptions on generalized smoothness of the population loss function (Assumptions (W.1)), generalized Łojasiewicz property of the population loss function (Assumption (W.2)), and uniform concentration bound between the gradients of the population and sample loss functions (Assumption (W.3)). The results in the table show that when $\alpha > 0$, the Polyak step size gradient descent iterates reach to the same statistical radius $\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$ as that of fixed-step size gradient descent iterates after much fewer number of iterations ($\log(1/\varepsilon(n,\delta))$ iterations of Polyak step size method versus $\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$ of fixed-step size method). As the complexity per iteration of the Polyak step size method and the fixed-step size method is similar, the Polyak method is more computationally efficient than the fixed-step size method for reaching the same final statistical radius. When $\alpha = 0$ and $\gamma = 0$, e.g., locally strongly convex setting, both the Polyak and fixed-step size methods reach the statistical radius $\varepsilon(n,\delta)$ after a logarithmic number of iterations.

When $\alpha = 0$, the generalized Łojasiewicz property is simply the well-known local Polyak-Łojasiewicz inequality [5]. This inequality has been used to guarantee the linear convergence of the fixed-step size gradient descent algorithm. When $\alpha > 0$, the inequality in Assumption (W.2) indicates that the gradient locally grows faster than a high order polynomial function as we move around the global minima $\theta^*$ where the maximum degree of the polynomial function is determine by the constant $\alpha$. Similar to Assumption (W.1), a simple example of the function $f$ that satisfies Assumption (W.2) is $f(\theta) = \sum_{i=1}^{d} \theta_i^{2\alpha_i}$ for all $\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \mathbb{R}^d$ where $\alpha_1, \alpha_2, \ldots, \alpha_d \geq 1$ are some given positive integers. The constant $\alpha$ in Assumption (W.2) takes the value $\alpha = \max_{1 \leq i \leq d}\{2\alpha_i - 2\}$. If we would like the function $f$ in this example to satisfy both Assumptions (W.1) and (W.2) with the same constant $\alpha$, we need to have $\alpha_1 = \alpha_2 = \ldots = \alpha_d = \alpha$, namely, homogeneous polynomial function. This behavior turns out to be popular in several statistical models, such as generalized linear model, mixture model, and mixed linear regression that we study in Section 3. In Appendix C, we also briefly discuss the behavior of the Polyak step size gradient descent algorithm when the simple polynomial function $f$ does not have homogeneous order, i.e., the constants in Assumptions (W.1) and (W.2) are different.

Finally, to analyze the iterates from the Polyak step size gradient descent algorithm for minimizing the sample loss function in equation (1), we need a growth condition on the uniform deviation bound between the gradients of the sample and population loss functions.

(W.3) (Stability Property) For a given parameter $\gamma \geq 0$, there exist a noise function $\varepsilon : \mathbb{N} \times (0,1] \to \mathbb{R}^+$, universal constant $c_3 > 0$, and some positive parameter $\rho > 0$ such that

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

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

A simple interpretation of the Assumption (W.3) is that we would like to control the growth of the noise function, resulting from the difference between the sample and population loss functions, when the radius of the ball around $\theta^*$ goes to 0. That assumption also suggests that $\theta^*$ is some stationary point of the sample loss function $f_n$ when $\gamma > 0$. A simple example for Assumption (W.3) is that $f_n(\theta) = \frac{\|\theta\|^{2p}}{2p} - \frac{\omega\|\theta\|^{2q}}{2q}\sqrt{\frac{d}{n}}$ where $\omega \sim \mathcal{N}(0,1)$ and $p, q$ are positive integers such that $p > q$. Under this simple case, $f(\theta) = \frac{\|\theta\|^{2p}}{2p}$ and the constant $\gamma$ in Assumption (W.3) takes the value $\gamma = 2q - 1$ while the noise function $\varepsilon(n,\delta) = \sqrt{\frac{d\log(1/\delta)}{n}}$. For more practical examples, we refer readers to Section 3.

### 2.2 Convergence rate of the Polyak step size gradient descent

The Polyak step size gradient descent iterates $\{\theta_n^t\}_{t\geq 0}$ for solving the sample loss function $f_n$ in equation (1) take the following form:

$$F_n(\theta_n^t) := \theta_n^{t+1} = \theta_n^t - \frac{f_n(\theta_n^t) - f_n(\widehat{\theta}_n)}{\|\nabla f_n(\theta_n^t)\|^2} \cdot \nabla f_n(\theta_n^t), \tag{5}$$

where $\widehat{\theta}_n$ is some optimal solution of the optimization problem (1) (See our discussion after Theorem 1 about an adaptive version of Polyak step size gradient descent algorithm to deal with the unknown value of $f_n(\widehat{\theta}_n)$). The operator $F_n$ in equation (5) is referred to as *sample Polyak operator*. To analyze the convergence rate of the sample iterates $\theta_n^t$, we will use the population to sample analysis discussed in equation (4). In particular, we define the following *population Polyak operator* for solving the population loss function $f$ in equation (3):

$$F(\theta) := \theta - \frac{f(\theta) - f(\theta^*)}{\|\nabla f(\theta)\|^2} \cdot \nabla f(\theta), \tag{6}$$

As being indicated in the population to sample analysis for analyzing the fixed-step size gradient descent algorithm, to analyze the sample iterates $\{\theta_n^t\}_{t\geq 0}$ of the Polyak step size gradient descent algorithm we use the following triangle inequality:

$$\|\theta_n^{t+1} - \theta^*\| \leq \|F_n(\theta_n^t) - F(\theta_n^t)\| + \|F(\theta_n^t) - \theta^*\|. \tag{7}$$

Therefore, to obtain an upper bound for the gap between $\theta_n^{t+1}$ and $\theta^*$, we need to understand the contraction of the population operator $F$ to $\theta^*$ as well as the deviation between the sample operator $F_n$ and population operator $F$. The following lemma shows the linear contraction of the population operator $F$ towards $\theta^*$.

**Lemma 1.** *Assume that Assumptions (W.1) and (W.2) hold. Then, given the definition of Polyak population operator in equation (6) we have*

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

*where* $\kappa := \left(1 - \frac{c_2^{\alpha+2}}{2c_1(\alpha+2)^{\alpha+2}}\right)^{1/2}$ *and* $c_1, c_2$ *are universal constants in Assumptions (W.1) and (W.2).*

The proof of Lemma 1 is in Section 5.1. The result of Lemma 1 indicates that if $\{\theta^t\}_{t\geq 0}$ is a sequence of population Polyak step size gradient descent iterates, i.e., $\theta^{t+1} = F(\theta^t)$, then we have

$$\|\theta^t - \theta^*\| \leq \kappa^t\|\theta^0 - \theta^*\|.$$

The linear convergence of population Polyak step size gradient descent iterates is in stark different from the sub-linear convergence $\Theta(t^{-1/\alpha})$ of the fixed-step size gradient descent iterates under Assumptions (W.1) and (W.2) (See Lemma 4 in Appendix B).

Our next result establishes an uniform concentration bound between the sample Polyak operator $F_n$ and the population Polyak operator $F$.

**Lemma 2.** *Assume that Assumptions (W.1), (W.2), and (W.3) hold with $\alpha \geq \gamma$. Assume that $\|\widehat{\theta}_n - \theta^*\| \leq r_n$ where $\widehat{\theta}_n$ is the optimal solution of the sample loss function $f_n$ and $r_n := \bar{C}\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$ where $\bar{C} = \left(\frac{C \cdot c_3(\alpha+2)^{\alpha+1}}{c_2^{\alpha+2}}\right)^{\frac{1}{1+\alpha-\gamma}}$, $c_2$, $c_3$ are the universal constant in Assumption (W.2) and (W.3) and $C$ is some universal constant. Then for any $r_n \leq r < \rho$ and for some universal constants $c_4 \geq 1$, we have*

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

The proof of Lemma 2 is in Section 5.2. A few comments with that lemma are in order. First, the condition $\alpha \geq \gamma$ is to guarantee that the signal is stronger than the noise in statistical model in which we can derive the meaningful statistical rate for our estimator. Second, the assumption that $\|\widehat{\theta}_n - \theta^*\| \leq r_n$ is natural as from Proposition 1, we demonstrate that that statistical radius is at the order of $\mathcal{O}(\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}})$. Third, as indicated in Lemma 2, the uniform concentration bound between the sample Polyak operator $F_n$ and the population Polyak operator $F$ only holds when $r_n \leq \|\theta - \theta^*\| \leq r$. The condition $\|\theta - \theta^*\| \geq r_n$ is important to ensure that the concentration bound is stable. When $\|\theta - \theta^*\| < r_n$, it happens that $\|F_n(\theta) - F(\theta)\|$ goes to infinity. This instability behavior of the concentration bound between $F_n$ and $F$ when the parameter approaches $\theta^*$ is different from the stable concentration bound of the sample fixed-step size gradient descent operator around the population fixed-step size gradient descent operator, which is proportional to $r^\gamma \cdot \varepsilon(n,\delta)$ according to Assumption (W.3) and holds for all $\theta \in \mathbb{B}(\theta^*, r)$.

Equipped with the linear convergence of the population Polyak operator in Lemma 1 and the uniform deviation bound between the sample Polyak operator $F_n$ and the population Polyak operator $F$, we are ready to state our main result about the statistical and computational complexity of the sample Polyak step size gradient descent iterates.

**Theorem 1.** *Assume that Assumptions (W.1), (W.2) and (W.3) and assumptions in Lemma 2 hold with $\alpha \geq \gamma$. Assume that the sample size $n$ is large enough such that $\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}} \leq \frac{(1-\kappa)\rho}{c_4 \bar{C}^{\gamma-\alpha}}$ where $\kappa$ is defined in Lemma 1, $c_4$ and $\bar{C}$ are the universal constants in Lemma 2, and $\rho$ is the local radius. Then, there exist universal constants $C_1$, $C_2$ such that for $t \geq C_1 \log(1/\varepsilon(n,\delta))$, the following holds:*

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

The proof of Theorem 1 is in Section 5.3. Below, we have the following discussions with the result of Theorem 1:

**Comparing to fixed-step size gradient descent:** Since the convergence rate of the fixed-step size gradient descent iterates is at the order of $t^{-1/\alpha}$ when $\alpha > 0$ under Assumptions (W.1) and (W.2) (See Lemma 4 in Appendix B) and the concentration bound between the sample gradient descent and population gradient descent operators are of the order $r^\gamma \cdot \varepsilon(n,\delta)$ under Assumption (W.3), the result of Theorem 1 in [16] indicates the following convergence rate of the fixed-step size gradient descent updates when $\alpha > 0$ and $\alpha \geq \gamma$.

**Proposition 1.** *Assume that Assumptions (W.1), (W.2) and (W.3) hold with $\alpha \geq \gamma$ and $\alpha > 0$. As long as the sample size $n$ is large enough such that $\varepsilon(n, \delta) \leq C$ for some universal constant $C$, there exist universal constants $C_1'$ and $C_2'$ such that for any fixed $\tau \in (0, \frac{1}{1+\alpha-\gamma})$ as long as $t \geq C_1' \varepsilon(n, \delta)^{-\frac{\alpha}{1+\alpha-\gamma}} \log(1/\tau)$, we have*

$$\|\theta_{n,GD}^t - \theta^*\| \leq C_2' \varepsilon(n, \delta)^{\frac{1}{\alpha+1-\gamma}-\tau},$$

*where $\{\theta_{n,GD}^t\}_{t\geq 0}$ is a sequence of sample fixed-step size gradient descent iterates.*

When $\alpha \geq \gamma$ and $\alpha > 0$, the result of Proposition 1 indicates that the fixed-step size gradient descent algorithm requires $\mathcal{O}(\varepsilon(n, \delta)^{-\frac{\alpha}{1+\alpha-\gamma}})$ number of iterations such that its updates can reach to the final statistical radius $\mathcal{O}(\varepsilon(n, \delta)^{\frac{1}{\alpha+1-\gamma}})$. Since each step of the gradient descent algorithm takes $\mathcal{O}(nd)$ arithmetic operations, it demonstrates that the total computational complexity for the fixed-step size gradient descent algorithm to reach the final statistical radius is $\mathcal{O}(n \cdot \varepsilon(n, \delta)^{-\frac{\alpha}{1+\alpha-\gamma}})$ for fixed dimension $d$. On the other hand, with a similar argument, Theorem 1 indicates that the total computational complexity for the Polyak step size gradient descent iterates to reach the final statistical radius is at the order of $\mathcal{O}(n \cdot \log(1/\varepsilon(n, \delta))$, which is much cheaper than that of the fixed-step size gradient descent algorithm when $\alpha \geq \gamma$.

**Cross-validation with the minimum number of iterates:** Note that, in Theorem 1 we only guarantee for the existence of some $k < t$ in the iterate that $\|\theta_n^k - \theta^*\| = \mathcal{O}(\varepsilon(n, \delta)^{\frac{1}{\alpha+1-\gamma}})$, instead of the generally desired last iterate $t$. As Ho et al. [16] pointed out, such minimum is unavoidable without further regularity conditions. Fortunately, we can still obtain the desired estimator in the iterate by cross-validation [35], which only accounts for an additional $\mathcal{O}(nd)$ computation and keeps the computational efficiency of the Polyak step-size gradient descent algorithm.

**Practical consideration of the Polyak step size gradient descent:** A practical issue of the original Polyak step size gradient descent algorithm is that it requires the knowledge of the optimal value of the sample loss function $f_n(\widehat{\theta}_n)$ (see equation (5)). Even though it may look restrictive at the first sight, it appears that we can utilize an adaptive version of that algorithm, named *adaptive Polyak step size gradient descent*, from [15] to deal with the unknown value of $f_n(\widehat{\theta}_n)$. The detailed description of that algorithm is in Algorithm 1.

As indicated in Algorithm 1, we first choose some lower bound $\tilde{f}_0$ of $f_n(\widehat{\theta}_n)$ and using it as a surrogate for $f_n(\widehat{\theta}_n)$. Then, we run the Polyak step size algorithm for $T$ times, which is the time horizon, with that surrogate choice. We then perform binary search to update that surrogate value to $\tilde{f}_1$ based on the current Polyak step size gradient descent iterates. We repeat that procedure $K$ times where $K$ is some given number of epochs to obtain a surrogate value $\tilde{f}_K$ of $f_n(\widehat{\theta}_n)$. As indicated in Theorem 2 of [15], to have $\tilde{f}_K - f_n(\widehat{\theta}_n) < \varepsilon$, we can choose $K = \mathcal{O}(\log(\frac{f_n(\widehat{\theta}_n)-\tilde{f}_0}{\varepsilon}))$ and $T = \mathcal{O}(\log(\frac{1}{\varepsilon}))$. Therefore, if we choose $\varepsilon = \mathcal{O}(\varepsilon(n, \delta)^{\frac{\alpha+2}{\alpha+1-\gamma}})$ (note that here $\varepsilon$ is the gap for value of the objective function), then based on the proof of Theorem 1, the adaptive Polyak step size gradient descent iterates converge to a final radius of convergence $\mathcal{O}(\varepsilon(n, \delta)^{\frac{1}{\alpha+1-\gamma}})$ after $\mathcal{O}(\log(1/\varepsilon(n, \delta))^2)$ number of iterations. It indicates that the adaptive Polyak step size gradient descent is still cheaper than the fixed-step gradient descent algorithm for reaching the same final statistical radius when $\alpha \geq \gamma$ and $\alpha > 0$, i.e., when the population loss function is not locally strongly convex.

**Algorithm 1:** Adaptive Polyak Step Size Gradient Descent

**Input:** Sample loss function $f_n$, initialization $\theta_n^0$, lower bound function $\tilde{f}_0$ such that $\tilde{f}_0 < f_n(\widehat{\theta}_n)$ where $\widehat{\theta}_n$ is some optimal solution of $f_n$, time horizon $T$, number of epochs $K$

$\bar{\theta} = \theta_n^0$
**for** $k = 0, 1, 2, \ldots, K-1$ **do**
  $\theta_n^{Tk} = \bar{\theta}$
  **for** $i = 0, 1, 2, \ldots, T-1$ **do**
    $\theta_n^{Tk+i+1} = \theta_n^{Tk+i} - \frac{f_n(\theta_n^{Tk+i}) - \tilde{f}_k}{\|\nabla f_n(\theta_n^{Tk+i})\|^2} \nabla f_n(\theta_n^{Tk+i})$
  **end**
  $\bar{\theta} = \arg\min_{0 \le i \le T} f_n(\theta_n^{Tk+i})$
  $\tilde{f}_{k+1} = \frac{f_n(\bar{\theta}) - \tilde{f}_k}{2}$
**end**

**Output:** $\bar{\theta}$

# 3 Examples

In this section, we consider an application of our theories in Section 2 to three specific examples: generalized linear model, over-specified Gaussian mixture model, and mixed linear regression model.

## 3.1 Generalized Linear Model

Generalized linear model is a generalization of linear regression model that allows the response variable to relate to the covariates via a link function. In particular, assume that $(Y_1, X_1), \ldots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$ satisfy

$$Y_i = g(X_i^\top \theta^*) + \varepsilon_i, \qquad \forall i \in [n] \tag{8}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a given link function, $\theta^*$ is a true but unknown parameter, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. noises from $\mathcal{N}(0, \sigma^2)$ where $\sigma > 0$ is a given variance parameter. Note that, the Gaussian assumption on the noise is just for the simplicity of the proof argument; the result in this section still holds for sub-Gaussian i.i.d. noise. Furthermore, we assume the random design setting of the generalized linear model, namely, $X_1, X_2, \ldots, X_n$ are i.i.d. from $\mathcal{N}(0, I_d)$.

For our study, we specifically consider $g(r) := r^p$ for any $p \in \mathbb{N}$ and $p \ge 2$. Note that, our choice of $g$ is motivated by phase retrieval problem [13, 34, 6, 31] where $g(r) = r^2$. To estimate $\theta^*$, we consider minimizing the least-square loss function, which is given by:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}_n(\theta) := \frac{1}{2n} \sum_{i=1}^{n} (Y_i - (X_i^\top \theta)^p)^2. \tag{9}$$

We then also have the corresponding population least-square loss function, which admits the following form:

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

where the outer expectation is taken with respect to $X \sim \mathcal{N}(0, I_d)$ and $Y = g(X^\top \theta^*) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Note that $\mathbb{E}[Y^2|X] = \mathbb{E}[g(X^\top\theta^*)^2] + \sigma^2$. Thus, by taking conditional expectation, the population loss function has the following form:

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{E}\left[\frac{1}{2}(Y - (X^\top\theta)^p)^2\right] \\ &= \frac{1}{2}\left(\mathbb{E}\left[\left((X^\top\theta^*)^p - (X^\top\theta)^p\right)^2\right] + \sigma^2\right). \end{aligned} \tag{10}$$

**Strong signal-to-noise regime:** When $\theta^*$ is bounded away from 0, i.e., $\|\theta^*\| \geq C$ for some universal constant $C$, the population least-square loss function $\mathcal{L}$ is locally strongly convex around $\theta^*$ and locally smooth, namely, the Assumptions (W.1) and (W.2) become

$$\lambda_{\max}(\nabla^2\mathcal{L}(\theta)) \leq c_1, \quad \|\nabla\mathcal{L}(\theta)\| \geq c_2(f(\theta) - f(\theta^*))^{1/2} \tag{11}$$

for all $\theta \in \mathbb{B}(\theta^*, \rho)$ where $\rho$ is some universal constant depending on $p$, as we demonstrate in Appendix A.1. Furthermore, for Assumption (W.3), for any $r > 0$ we can demonstrate that there exist universal constants $C_1$ and $C_2$ such that as long as $n \geq C_1(d\log(d/\delta))^{2p}$ with probability $1 - \delta$

$$\sup_{\theta\in\mathbb{B}(\theta^*,r)} \|\nabla\mathcal{L}_n(\theta) - \nabla\mathcal{L}(\theta)\| \leq C_2\sqrt{\frac{d + \log(1/\delta)}{n}}. \tag{12}$$

The proof for this uniform concentration bound is also in Appendix A.1. These results indicate that $\alpha = \gamma = 0$ in Assumptions (W.1)-(W.3). Therefore, a direct application of Theorem 1 shows that we have the iterates of Polyak step size gradient descent algorithm converge to a radius of convergence $\mathcal{O}(\sqrt{d/n})$ around $\theta^*$ within $\mathcal{O}(\log(n/d))$ number of iterations.

**Low signal-to-noise regime:** On the other hand, when $\|\theta^*\|$ is sufficiently small, the population loss function is no longer locally strongly convex and the precise understandings of the sample updates from the Polyak step size gradient descent algorithm for solving the sample loss function $\mathcal{L}_n$ have remained poorly understood. To illustrate the behaviors of the Polyak step size gradient descent algorithm, we only focus on the no signal-to-noise setting $\theta^* = 0$ in this section. Under this setting, the population least-square loss function can be written as

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta) = \frac{\sigma^2 + (2p-1)!!\|\theta - \theta^*\|^{2p}}{2}. \tag{13}$$

Different from the setting when $\theta^*$ is bounded away from 0, the population loss function in equation (13) is not locally strongly convex around $\theta^*$ when $\theta^* = 0$. Indeed, we demonstrate in Appendix A.1 that for all $\theta \in \mathbb{B}(\theta^*, \rho)$ for some radius $\rho$, we have

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

$$\|\nabla\mathcal{L}(\theta)\| \geq c_2(\mathcal{L}(\theta) - \mathcal{L}(\theta^*))^{1-\frac{1}{2p}}, \tag{15}$$

where $c_1, c_2$ are some universal constants depending on $r$. Furthermore, for Assumption (W.3), from Appendix A.2 in [29], there exist universal constants $C_1$ and $C_2$ such that for any $r > 0$ and $n \geq C_1(d\log(d/\delta))^{2p}$ we have

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

with probability at least $1-\delta$. These results suggest that as long as $r \in (0, \rho)$ for some given $\rho$, the values of constants $\alpha$ and $\gamma$ in Assumptions (W.1)-(W.3) are $\alpha = 2p-2$ and $\gamma = p-1$.

Given the above studies, a direct application of Theorem 1 leads to the following bounds on the statistical radius of the sample Polyak step size gradient descent iterates.

**Corollary 1.** *For the generalized linear model (8) with the link function $g(r) = r^p$ for some natural number $p \geq 2$, as long as $n \geq c(d\log(d/\delta))^{2p}$ for some positive universal constant $c$ and $\theta_n^0 \in \mathbb{B}(\theta^*, \rho)$ for some $\rho > 0$, with probability $1-\delta$ the sequence of sample Polyak step size gradient descent iterates $\{\theta_n^t\}_{t\geq 0}$ satisfies the following bounds*

*(i) Strong signal-to-noise regime: When $\|\theta^*\| \geq C$ for some constant $C$, we have*

$$\min_{1\leq k\leq t} \|\theta_n^k - \theta^*\| \leq c_1\sqrt{\frac{d+\log(1/\delta)}{n}}, \qquad \text{for } t \geq c_2 \log\left(\frac{n}{d+\log(1/\delta)}\right),$$

*(ii) Low signal-to-noise regime: When $\theta^* = 0$, we find that*

$$\min_{1\leq k\leq t} \|\theta_n^k - \theta^*\| \leq c_1'\left(\frac{d+\log(1/\delta)}{n}\right)^{1/(2p)}, \qquad \text{for } t \geq c_2' \log\left(\frac{n}{d+\log(1/\delta)}\right)$$

*Here, $c_1, c_2, c_1', c_2'$ are some universal constants.*

In light of Proposition 1, when $\theta^* = 0$ the iterates from the fixed-step size gradient descent algorithm have similar statistical radius $(d/n)^{1/(2p)}$ as that of the Polyak step size gradient descent updates. However, the fixed-step size gradient descent algorithm need at least $\mathcal{O}((n/d)^{\frac{p-1}{p}})$ number of iterations to reach that radius of convergence. It demonstrates that the Polyak step size gradient descent algorithm is much cheaper than the fixed-step size gradient descent algorithm in terms of the sample size $n$.

### 3.2 Mixture model

Gaussian mixture models are one of the most popular tools in machine learning and statistics for modeling heterogeneous data [25, 27]. In these models, learning location and scale parameters associated with each sub-population is important to understand the heterogeneity of the data. A popular approach to estimate these parameters is to maximize the log-likelihood function. Since the log-likelihood function of Gaussian mixture models is non-concave and complicated to study, a full picture about the convergence rates of optimization algorithms for solving the log-likelihood function of the over-specified Gaussian mixture models has still remained elusive.

In this section, we aim to shed light on the convergence rates of the Polyak step size gradient descent algorithm for solving the parameters of Gaussian mixture models. We specifically consider the symmetric two-component Gaussian mixture and provide comprehensive analysis of that algorithm. In particular, we assume that the data $X_1, X_2, \ldots, X_n$ are i.i.d. samples from $\frac{1}{2}\mathcal{N}(-\theta^*, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta^*, \sigma^2 I_d)$ where $\sigma > 0$ is given and $\theta^*$ is true but unknown parameter. To estimate $\theta^*$, we fit the data by the symmetric two-component Gaussian mixture

$$\frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d). \tag{17}$$

The maximum likelihood estimation is then given by:

$$\min_{\theta\in\mathbb{R}^d} \bar{\mathcal{L}}_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{2}\phi(X_i|\theta,\sigma^2 I_d) + \frac{1}{2}\phi(X_i|-\theta,\sigma^2 I_d)\right), \tag{18}$$

where $\phi(\cdot|\theta,\sigma^2 I_d)$ is the density function of multivariate Gaussian distribution with mean $\theta$ and covariance matrix $\sigma^2 I_d$. The corresponding population version of the maximum likelihood estimation (18) takes the following form:

$$\min_{\theta\in\mathbb{R}^d} \bar{\mathcal{L}}(\theta) := -\mathbb{E}_X\left[\log\left(\frac{1}{2}\phi(X|\theta,\sigma^2 I_d) + \frac{1}{2}\phi(X|-\theta,\sigma^2 I_d)\right)\right], \tag{19}$$

where the outer expectation is taken with respect to $X \sim \frac{1}{2}\mathcal{N}(-\theta^*,\sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta^*,\sigma^2 I_d)$.

**Strong signal-to-noise regime:** When $\|\theta^*\| \geq C\sigma$ for some universal constant $C$, the Corollary 1 in [2] demonstrates that the population loss function $\bar{\mathcal{L}}$ is locally strongly convex and locally smooth as long as $\theta \in \mathbb{B}(\theta^*, \frac{\|\theta^*\|}{4})$. It indicates that we have

$$\lambda_{\max}(\nabla^2\bar{\mathcal{L}}(\theta)) \leq c_1, \quad \|\nabla\bar{\mathcal{L}}(\theta)\| \geq c_2(f(\theta) - f(\theta^*))^{1/2}, \tag{20}$$

i.e., the Assumptions (W.1) and (W.2) are satisfied with the constant $\alpha = 0$. Furthermore, for any $r \leq \frac{\|\theta^*\|}{4}$ and $n \geq C_1 d\log(1/\delta)$ for some universal constant $C_1$ we have

$$\sup_{\theta\in\mathbb{B}(\theta^*,r)} \|\nabla\bar{\mathcal{L}}_n(\theta) - \nabla\bar{\mathcal{L}}(\theta)\| \leq C_2\sqrt{\frac{d\log(1/\delta)}{n}} \tag{21}$$

with probability at least $1-\delta$ where $C_2$ is some universal constant. See Corollary 4 in [2] for the proof of this concentration result.

**Low signal-to-noise regime:** We specifically consider the setting $\theta^* = 0$. This setting corresponds to the popular over-specified Gaussian mixture models [33, 17], namely, when we choose some given number of components that can be (much) larger than the true number of components and estimating the parameters from the mixture models with that chosen number of components. We prove in Appendix A.2 that for all $\theta \in \mathbb{B}(\theta^*, \frac{\sigma}{2})$:

$$\lambda_{\max}(\nabla^2\bar{\mathcal{L}}(\theta)) \leq c_1\|\theta-\theta^*\|^2, \tag{22}$$

$$\|\nabla\bar{\mathcal{L}}(\theta)\| \geq c_2(\bar{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta^*))^{3/4}. \tag{23}$$

Furthermore, from Lemma 1 in [12], there exist universal constants $C_1$ and $C_2$ such that for any $r > 0$ and $n \geq C_1 d\log(1/\delta)$ we have:

$$\sup_{\theta\in\mathbb{B}(\theta^*,r)} \|\nabla\bar{\mathcal{L}}_n(\theta) - \nabla\bar{\mathcal{L}}(\theta)\| \leq C_2 r\sqrt{\frac{d\log(1/\delta)}{n}} \tag{24}$$

with probability at least $1-\delta$.

Combining the above results to Theorem 1, we have the following results on the final statistical radius of the Polyak step size iterates under different regimes of the two-component Gaussian mixture model.

**Corollary 2.** *For the symmetric two-component mixture model (17), there exist positive universal constants $c_1, c_2, c_1', c_2'$ such that when $n \geq cd\log(1/\delta)$ for some universal constant $c$, with probability $1-\delta$ the sequence of sample Polyak step size gradient descent iterates $\{\theta_n^t\}_{t\geq 0}$ satisfies the following bounds:*

*(i) Strong signal-to-noise regime: When* $\|\theta^*\| \geq C$ *for some constant* $C$ *and* $\theta_n^0 \in \mathbb{B}(\theta^*, \frac{\|\theta\|^*}{4})$, *we have*

$$\min_{1\leq k\leq t} \|\theta_n^k - \theta^*\| \leq c_1 \sqrt{\frac{d\log(1/\delta)}{n}}, \qquad \text{for } t \geq c_2 \log\left(\frac{n}{d\log(1/\delta)}\right),$$

*(ii) Low signal-to-noise regime: When* $\theta^* = 0$ *and* $\theta_n^0 \in \mathbb{B}(\theta^*, \frac{\sigma}{2})$ *we find that*

$$\min_{1\leq k\leq t} \|\theta_n^k - \theta^*\| \leq c_1' \left(\frac{d\log(1/\delta)}{n}\right)^{1/4}, \qquad \text{for } t \geq c_2' \log\left(\frac{n}{d\log(1/\delta)}\right).$$

A few comments with the results of Corollary 2 are in order. First, the Expectation-Maximization (EM) algorithm [9] is a popular algorithm for solving the parameters of Gaussian mixture models. In the symmetric two-component Gaussian mixture (17), the EM algorithm is simply the gradient descent with step size being 1. In light of the results of Proposition 1 and the results in [12], the EM iterates reach to the final statistical radius $\mathcal{O}((d/n)^{1/4})$ after $\mathcal{O}(\sqrt{n})$ number of iterations. The results in Corollary 2 indicate that the Polyak step size gradient descent iterates reach to the final statistical radius with a much fewer number of iterations, namely, $\mathcal{O}(\log(n))$, while each iteration of the Polyak step size gradient descent has similar computational complexity as that of the EM algorithm. Therefore, the Polyak step size gradient descent algorithm is more efficient than the EM algorithm for the low-signal-to noise regime of symmetric two-component Gaussian mixture model. Second, the statistical radius $(d/n)^{1/4}$ that the Polyak iterates reach to in the low signal-to-noise regime is optimal according to the work [18].

### 3.3 Mixed linear regression

Mixed linear regression is a generalization of vanilla linear regression model when we have multiple mean parameters and each data can associate with one of these parameters. In statistics, mixed linear regression is often referred to as mixture of regression [21], which is also a special case of mixture of experts [19, 20] where the mixing weights are assumed to be independent of the covariates.

Similar to mixture model in Section 3.2, we also aim to shed light on the convergence rate of the Polyak step size gradient descent algorithm under the simple symmetric two-component mixed linear regression setting. In particular, we assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are i.i.d. samples from symmetric two components

$$\left(\frac{1}{2}\mathcal{N}(Y|-(\theta^*)^\top X, \sigma^2) + \frac{1}{2}\mathcal{N}(Y|(\theta^*)^\top X, \sigma^2)\right) \cdot \mathcal{N}(X|0, I_d), \tag{25}$$

where $\sigma > 0$ is known variance and $\theta^*$ is true but unknown parameter. To estimate $\theta^*$, we fit the data with the following symmetric two-component mixed linear regression:

$$\left(\frac{1}{2}\mathcal{N}(Y|-\theta^\top X, \sigma^2) + \frac{1}{2}\mathcal{N}(Y|\theta^\top X, \sigma^2)\right) \cdot \mathcal{N}(X|0, I_d). \tag{26}$$

A common approach to obtain an estimator of $\theta^*$ is maximum likelihood estimator, which is given by:

$$\min_{\theta\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n(\theta) := -\frac{1}{n}\sum_{i=1}^n \log\left(\frac{1}{2}\phi(Y_i|\theta^\top X_i, \sigma^2) + \frac{1}{2}\phi(Y_i|-\theta^\top X_i, \sigma^2)\right). \tag{27}$$

The corresponding population version of the optimization problem (27) is

$$\min_{\theta \in \mathbb{R}^d} \widetilde{\mathcal{L}}(\theta) := -\mathbb{E}_{X,Y}\left[\log\left(\frac{1}{2}\phi(Y|\theta^\top X, \sigma^2) + \frac{1}{2}\phi(Y|-\theta^\top X, \sigma^2)\right)\right], \tag{28}$$

where the outer expectation is taken with respect to $X \sim \mathcal{N}(0, I_d)$ and $Y|X \sim \frac{1}{2}\mathcal{N}(Y|-(\theta^*)^\top X, \sigma^2) + \frac{1}{2}\mathcal{N}(Y|(\theta^*)^\top X, \sigma^2)$.

**Strong signal-to-noise regime:** We first consider the setting when $\|\theta^*\| \geq C$ where $C$ is some universal constant. Corollary 2 in [2] proves that for that strong signal-to-noise regime, the population negative log-likelihood function $\tilde{\mathcal{L}}$ is locally strongly convex and smooth when $\theta \in \mathbb{B}(\theta^*, \frac{\|\theta^*\|}{32})$. Therefore, the Assumptions (W.1) and (W.2) are satisfied with the constant $\alpha = 0$. Furthermore, according to the result of Corollary 5 in [2], Assumption (W.3) is satisfied with $\gamma = 0$ and for any radius $r \leq \|\theta^*\|/32$.

**Low signal-to-noise regime:** We consider specifically the setting that $\theta^* = 0$. We prove in Appendix A.3 that for all $\theta \in \mathbb{B}(\theta^*, \frac{\sigma}{\sqrt{20}})$, there exist universal constants $c_1$ and $c_2$ such that:

$$\lambda_{\max}(\nabla^2 \widetilde{\mathcal{L}}(\theta)) \leq c_1 \|\theta - \theta^*\|^2, \tag{29}$$

$$\|\nabla \widetilde{\mathcal{L}}(\theta)\| \geq c_2 (\widetilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\theta^*))^{3/4}. \tag{30}$$

These results indicate that the Assumptions (W.1) and (W.2) are satisfied with the constant $\alpha = 2$. Furthermore, from the concentration result from Lemma 2 of [24], there exist universal constants $C_1$ and $C_2$ such that as long as $n \geq C_1 d \log(1/\delta)$, we have for any $r > 0$

$$\mathbb{P}\left(\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla \widetilde{\mathcal{L}}_n(\theta) - \nabla \widetilde{\mathcal{L}}(\theta)\| \leq C_2 r \sqrt{\frac{d\log(1/\delta)}{n}}\right) \geq 1 - \delta.$$

It indicates that Assumption (W.3) is satisfied when $\gamma = 1$. Collecting all of the above results under both the strong and low signal-to-noise regimes, we have the following bounds on the statistical radii of the Polyak step size gradient descent iterates.

**Corollary 3.** *For the symmetric two-component mixed linear regression (25), when $n \geq c \cdot d\log(1/\delta)$ for some universal constant $c$, there exist positive universal constants $c_1, c_2, c_1', c_2'$ such that with probability $1-\delta$ the sequence of sample Polyak step size gradient descent iterates $\{\theta_n^t\}_{t \geq 0}$ satisfies the following bounds:*

*(i) Strong signal-to-noise regime: When $\|\theta^*\| \geq C$ for some constant $C$ and $\theta_n^0 \in \mathbb{B}(\theta^*, \frac{\|\theta^*\|}{32})$, we have*

$$\min_{1 \leq k \leq t} \|\theta_n^k - \theta^*\| \leq c_1 \sqrt{\frac{d\log(1/\delta)}{n}}, \qquad \text{for } t \geq c_2 \log\left(\frac{n}{d\log(1/\delta)}\right),$$

*(ii) Low signal-to-noise regime: When $\theta^* = 0$ and $\theta_n^0 \in \mathbb{B}(\theta^*, \frac{\sigma}{\sqrt{20}})$ we find that*

$$\min_{1 \leq k \leq t} \|\theta_n^k - \theta^*\| \leq c_1' \left(\frac{d\log(1/\delta)}{n}\right)^{1/4}, \qquad \text{for } t \geq c_2' \log\left(\frac{n}{d\log(1/\delta)}\right).$$

Similar to the symmetric two-component Gaussian mixture, the EM algorithm for solving the symmetric two-component mixed linear regression is simply the gradient descent with step size one. The results of Corollary 3 and Proposition 1 indicate that the Polyak iterates take much fewer number of iterations, i.e., $\mathcal{O}(\log(n))$ than that of the EM algorithm, which is $\mathcal{O}(\sqrt{n})$. It indicates that the Polyak step size gradient descent algorithm is computationally more efficient than the EM algorithm for reaching to the optimal statistical radius $\mathcal{O}((d/n)^{1/4})$ in the low signal-to-noise regime.

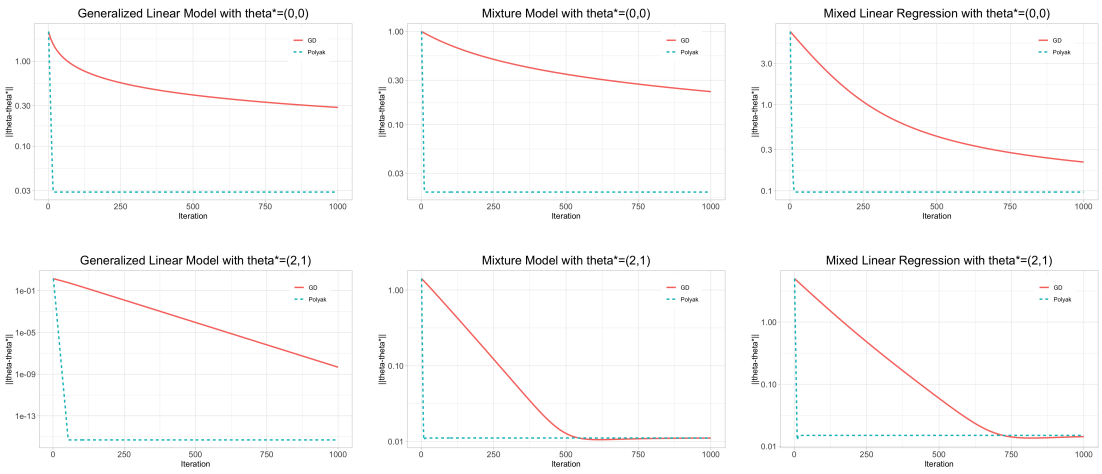


**Figure 1.** The convergence rates of Polyak step size and fixed-step size gradient descent iterates for solving the population losses of generalized linear model, Gaussian mixture model, and mixed linear regression model in Section 3. The first row corresponds to the low signal-to-noise regime $\theta^* = (0, 0)$ while the second row is for the strong signal-to-noise regime $\theta^* = (2, 1)$.

# 4 Experiments

In this section, we illustrate the behaviors of Polyak step size gradient descent iterates for three statistical examples in Section 3. In Section 4.1, we compare the behaviors of population Polyak step size gradient descent iterates and population fixed-step size gradient descent iterates for solving the population loss functions of the given statistical models. In Section 4.2, we compare the sample iterates from both (adaptive) Polyak step size and fixed-step size gradient descent methods.

## 4.1 Population loss function

We first use Polyak step size and fixed-step size gradient descent algorithms to find the minima of the population losses of three examples in Section 3. We consider these examples in $d = 2$ dimensions. For the strong signal-to-noise regime, we choose $\theta^* = (2, 1)$. We compare the convergence rates of Polyak step size and fixed-step size iterates to the optimal solution $\theta^*$ of the population losses in Figure 1. In this figure, GLM, GMM, MLR respectively stand for generalized linear model, Gaussian mixture model, and mixed linear regression. All the plots in this figure are log-log scale plots. From this figure, the Polyak step size GD iterates converge linearly to $\theta^*$ while the fixed-step size gradient descent iterates converge sub-linearly to $\theta^*$. These experiment results are consistent with our theories in Section 3.

## 4.2 Sample loss function

Now, we carry out the experiments to compare the behaviors of Polyak step size and fixed-step size gradient descent iterates for solving the sample loss functions in three examples in Section 3. In these examples, since we only observe the data, we do not have access to the optimal value of the sample loss functions. Therefore, we instead use the adaptive Polyak step size gradient descent in Algorithm 1 for these examples. The strategy for choosing the lower bound of the optimal value of the sample loss functions in that algorithm will be described

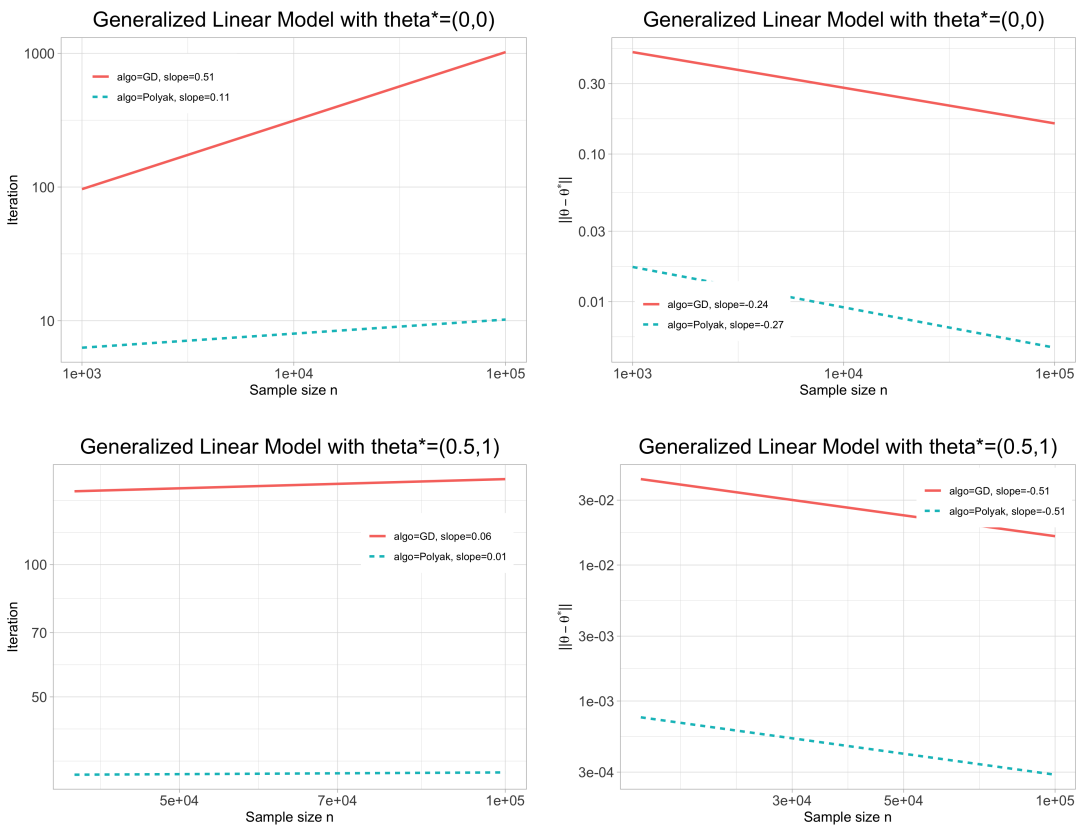


**Figure 2.** The convergence rates of adaptive Polyak step size gradient descent and fixed-step size gradient descent iterates for solving the sample loss function of the generalized linear model when the link function $g(r) = r^2$. The first row corresponds to the low signal-to-noise regime $\theta^* = (0, 0)$ while the second row is for the strong signal-to-noise regime $\theta^* = (0.5, 1)$. For the left images, we use log-log plots to illustrate the iteration complexities of these algorithms to reach the final estimate. For the right images, log-log plots for the final statistical radius versus the sample size are presented. For the low signal-to-noise regime, both the adaptive Polyak step size and fixed-step size gradient descent iterates reach the statistical radius $n^{-1/4}$. The adaptive Polyak step size method takes much fewer number of iterations to reach the final statistical radius than the fixed-step size method, namely, from $\log(n)$ number of iterations of adaptive Polyak step size method to to $\sqrt{n}$ number of iterations of fixed-step size method. For the strong signal-to-noise regime, both adaptive Polyak and fixed-step size methods only take logarithmic number of iterations to reach the statistical radius $n^{-1/2}$.

in details in each example. In our experiments, the sample size $n$ is chosen to be in the set $\{1000, 2000, \cdots, 100000\}$.

**Generalized linear model:** We first consider the generalized linear model in Section 3.1. We specifically choose the link functions $g(r) = r^2$ , i.e., $p = 2$. The data $(Y_1, X_1), \ldots, (Y_n, X_n)$ satisfy

$$Y_i = (X_i^\top \theta^*)^2 + \varepsilon_i,$$

where $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_2)$ and $\varepsilon_1, \ldots, \varepsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$. We choose $\theta^* = (0, 0)$ for the low signal-to-noise regime and $\theta^* = (0.5, 1)$ for the strong signal-to-noise regime in our experiments.

Since we do not have access to $\mathcal{L}_n(\widehat{\theta}_n)$ where $\widehat{\theta}_n$ is the optimal solution of the sample loss

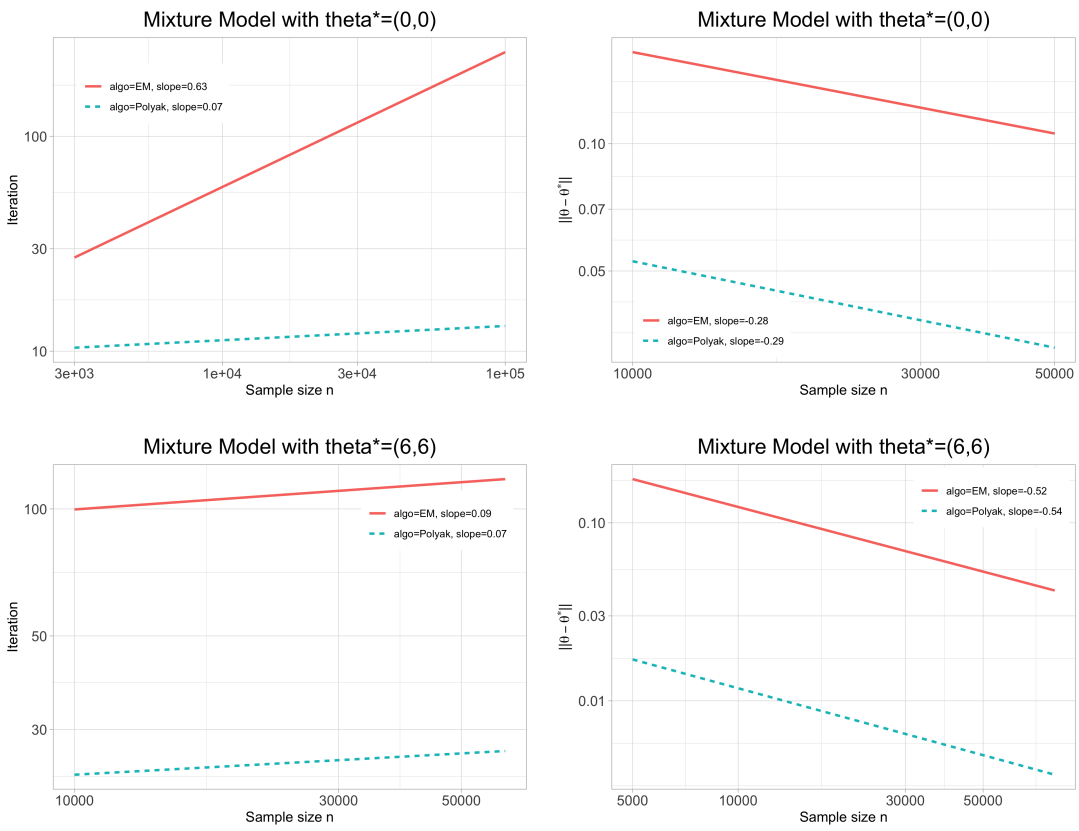


**Figure 3.** Illustrations for the convergence rates of adaptive Polyak step size and the EM algorithm (equivalently gradient descent algorithm with step size 1) for solving the sample log-likelihood function of the symmetric two-component Gaussian mixtures. The first row corresponds to the low signal-to-noise regime $\theta^* = (0, 0)$ while the second row is for the strong signal-to-noise regime $\theta^* = (6, 6)$. The structures of the images are similar to those in Figure 2. The images in the first row for low signal-to-noise regime show that the adaptive Polyak step size iterates only need roughly $\log(n)$ number of iterations in comparison to $\sqrt{n}$ number of iterations of the EM algorithm to reach the final statistical radius $n^{-1/4}$. The images in the second row for strong signal-to-noise regime show that these optimization methods have similar sample and iteration complexities.

function $\mathcal{L}_n$ in equation (9), we will consider its approximated value according to the adaptive Polyak step size gradient descent algorithm in Algorithm 1. By concentration inequality with the chi-squared random variables, the concentration of $\mathcal{L}_n(\hat{\theta}_n)$ is at the order of $\mathcal{O}(\frac{1}{\sqrt{n}})$ with high probability. Therefore, we use $\frac{c}{\sqrt{n}}$ to approximate $\mathcal{L}_n(\hat{\theta}_n)$, where $c$ here is a parameter to choose in the our experiment.

The updates from the adaptive Polyak step size gradient descent based on that approximation are given by:

$$\theta_n^{t+1} = \theta_n^t - \frac{\mathcal{L}_n(\theta_n^t) - \frac{c}{\sqrt{n}}}{\|\nabla \mathcal{L}_n(\theta_n^t)\|^2} \nabla \mathcal{L}_n(\theta_n^t).$$

When implementing the adaptive Polyak step size gradient descent algorithm, we use binary search to update the value of $c$ periodically. In particular, when the algorithm is stuck at some point, we decrease $c$; when it become very unstable, we increase $c$. For the fixed-step

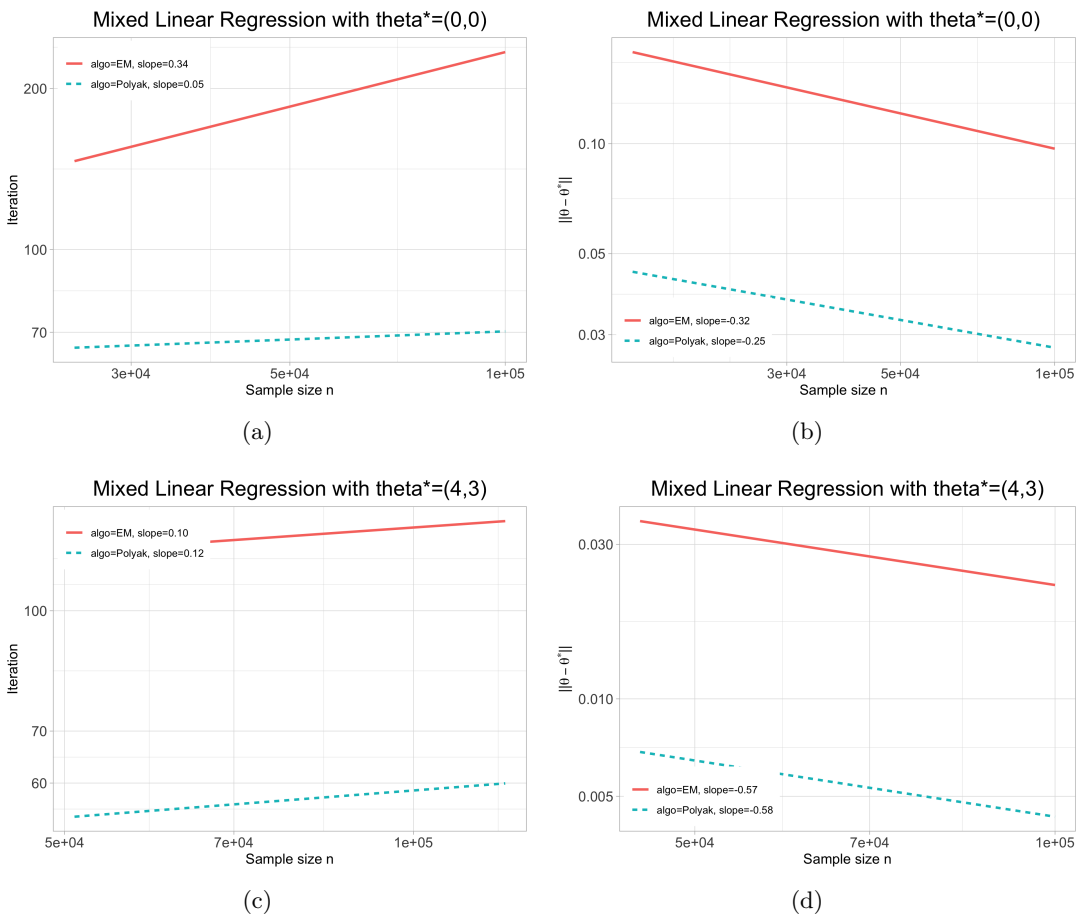


**Figure 4.** Plots characterizing the convergence rates of adaptive Polyak step size and EM algorithm (equivalently gradient descent algorithm with step size 1) for solving the sample log-likelihood function of the symmetric two-component mixed linear regression model. The first row corresponds to the low signal-to-noise regime $\theta^* = (0, 0)$ while the second row is for the strong signal-to-noise regime $\theta^* = (4, 3)$. From the images in the first row for low signal-to-noise regime, the iteration complexity of the adaptive Polyak step size method is roughly $\log(n)$ while that of the EM algorithm scales like $\sqrt{n}$ to reach the final statistical radius $n^{-1/4}$. From the images in the second row for strong signal-to-noise regime, both these optimization algorithms have sample complexity $n^{-1/2}$ and iteration complexity $\log(n)$.

size gradient descent algorithm, we choose the step size to be 0.01.

The experiment results are shown in Figure 2. For the left images in that figure, we use log-log plot to illustrate the iteration complexity of the adaptive Polyak step size and fixed-step size gradient descent algorithms versus the sample size under the low signal-to-noise setting $\theta^* = (0, 0)$ (first row) and the strong signal-to-noise setting $\theta^* = (0.5, 1)$ (second row). When $\theta^* = (0, 0)$, we observe that the number of iterations for the fixed-step size gradient descent algorithm to reach the final statistical radius is at the order close to $\sqrt{n}$ while the iteration complexity for the adaptive Polyak step size gradient descent algorithm is roughly $\log(n)$. On the other hand, when $\theta^* = (0.5, 1)$, both the iteration complexities of these algorithms scale like $\log(n)$. For the right images in Figure 2, we plot the final statistical radii of the adaptive Polyak and fixed-step size gradient descent iterates versus the sample size under different settings of $\theta^*$. As being indicated in these images, the radius scales like $n^{-1/4}$

when $\theta^* = (0, 0)$ while it is roughly $n^{-1/2}$ when $\theta^* = (0.5, 1)$. These results, along with our comments about the adaptive Polyak step size gradient descent algorithm after Theorem 1, confirm our theories in Section 3.1.

**Mixture model:** We now move to the symmetric two-component Gaussian mixture model considered in Section 3.2. We set dimension $d = 2$, the variance $\sigma = 1, \theta^* = (0, 0)$ for the low signal-to-noise regime and $\theta^* = (6, 6)$ for the strong signal-to-noise regime. To obtain an estimation of $\theta^*$, we maximize the log-likelihood in equation (18). We use $\frac{c}{n}$ to approximate the optimal value of sample log-likelihood function $\bar{\mathcal{L}}_n$ where $c$ is some universal constant. We also use binary search to adaptively update the value of the constant $c$ when we run the adaptive Polyak step size algorithm. We compare the performance of that algorithm to the EM algorithm (equivalently gradient descent algorithm with step size 1) in Figure 3. When $\theta^* = (0, 0)$, the images in the first row of Figure 3 show that the adaptive Polyak step size iterates only need roughly $\log(n)$ number of iterations in comparison to $\sqrt{n}$ number of iterations of the EM algorithm to reach the final statistical radius $n^{-1/4}$. When $\theta^* = (6, 6)$, the images in the second row for strong signal-to-noise regime show that these optimization methods have similar sample complexities $n^{-1/2}$ and iteration complexities $\log(n)$. These experiment results prove that the adaptive Polyak step size gradient descent algorithm is computationally more efficient than the EM algorithm to reach the final estimate for the low signal-to-noise regime, which confirms out theories in Section 3.2.

**Mixed linear regression:** Finally, we consider the two-component mixed linear regression example in Section 3.3. We consider $\theta^* = (0, 0)$ for the low signal-to-noise regime and $\theta^* = (4, 3)$ for the strong signal-to-noise regime. We choose the variance $\sigma = 1$ in model (25). Our goal is to maximize the log-likelihood in equation (27). Similar to the two-component Gaussian mixture model, we use use $\frac{c}{n}$ to approximate the optimal value of $\widetilde{\mathcal{L}}_n$ in the adaptive Polyak step size gradient descent method where $c$ is adaptively updated via the binary search. We compare the adaptive Polyak step size algorithm to the EM algorithm (equivalently gradient descent algorithm with step size 1) in Figure 4. When $\theta^* = (4, 3)$, both optimization algorithms reach the final statistical radius $n^{-1/2}$ around $\theta^*$ after $\log(n)$ number of iterations. When $\theta^* = (0, 0)$, the adaptive Polyak step size iterates reach the statistical radius $n^{-1/4}$ after $\log(n)$ number of iterations while the EM algorithm needs roughly $\sqrt{n}$ number of iterations to reach the same radius. These observations are consistent with our theories in Section 3.3.

# 5 Proofs

In this section, we provide the proofs for main results in Section 2.2.

## 5.1 Proof of Lemma 1

First, we notice that

$$\begin{aligned}
\|\theta^{t+1} - \theta^*\|^2 - \|\theta^t - \theta^*\|^2 &= \frac{(f(\theta^t) - f(\theta^*))^2}{\|\nabla f(\theta^t)\|^2} - \frac{2(f(\theta^t) - f(\theta^*))}{\|\nabla f(\theta^t)\|^2}\langle \nabla f(\theta^t), \theta^t - \theta^* \rangle \\
&= \frac{f(\theta^t) - f(\theta^*)}{\|\nabla f(\theta^t)\|^2}\left(f(\theta^t) - f(\theta^*) - 2\langle \nabla f(\theta^t), \theta^t - \theta^* \rangle\right) \\
&\leq -\frac{(f(\theta^t) - f(\theta^*))^2}{\|\nabla f(\theta^t)\|^2} \leq 0
\end{aligned}$$

where the inequality is due to the convexity of the population loss function $f$. This result indicates that the sequence $\{\|\theta^t - \theta^*\|\}_{t\geq 0}$ is monotonically decreasing and thus $\theta^t \in \mathbb{B}(\theta^*, \rho)$ for all $t \geq 0$ as long as $\theta^1 \in \mathbb{B}(\theta^*, \rho)$. Furthermore, we find that

$$\begin{aligned}\|\theta^{t+1} - \theta^*\|^2 - \|\theta^t - \theta^*\|^2 \leq -\frac{(f(\theta^t) - f(\theta^*))^2}{\|\nabla f(\theta^t)\|^2} &\leq -\frac{f(\theta^t) - f(\theta^*)}{2c_1\|\theta^t - \theta^*\|^\alpha} \\ &\leq -\frac{c_2^{\alpha+2}}{2c_1(\alpha+2)^{\alpha+2}}\|\theta^t - \theta^*\|^2,\end{aligned}$$

where the second inequality is based on the fact that $f(\theta^t) - f(\theta^*) \geq \frac{\|\nabla f(\theta^t)\|^2}{2c_1\|\theta^t - \theta^*\|^\alpha}$ which can be recovered from Lemma 3.5 in [5] (also stated in Lemma 5) and Assumption (W.1), and the third inequality is from Lemma 3. The above inequality is equivalent to

$$\|\theta^{t+1} - \theta^*\|^2 \leq \left(1 - \frac{c_2^{\alpha+2}}{2c_1(\alpha+2)^{\alpha+2}}\right)\|\theta^t - \theta^*\|^2. \tag{31}$$

We can further see that for any $\theta \in \mathbb{B}(\theta^*, \rho)$

$$\|\theta - \theta^*\| \leq \frac{\alpha+2}{c_2}(f(\theta) - f(\theta^*))^{\frac{1}{\alpha+2}} \leq \frac{\alpha+2}{c_2} \cdot \left(\frac{c_1}{2}\right)^{\frac{1}{\alpha+2}}\|\theta - \theta^*\|,$$

which means $\left(\frac{c_2}{\alpha+2}\right)^{\alpha+2} \leq \frac{c_1}{2}$ and $\frac{c_2^{\alpha+2}}{2c_1(\alpha+2)^{\alpha+2}} \leq \frac{1}{4}$. Thus, the contraction coefficient $\frac{3}{4} \leq 1 - \frac{c_2^{\alpha+2}}{2c_1(\alpha+2)^{\alpha+2}} < 1$, which means that it is positive and strictly less than 1. By repeating the inequality (31), we eventually have the following inequality:

$$\|\theta^{t+1} - \theta^*\|^2 \leq \left(1 - \frac{c_2^{\alpha+2}}{2c_1(\alpha+2)^{\alpha+2}}\right)^t \|\theta^0 - \theta^*\|^2.$$

As a consequence, we reach the conclusion of Lemma 1.

### 5.2 Proof of Lemma 2

Recall that, from Assumptions (W.1), (W.2) and Lemma 3, as long as $\theta \in \mathbb{B}(\theta^*, \rho)$ we have the following relations:

$$f(\theta) - f(\theta)^* \leq \frac{c_1}{2}\|\theta - \theta^*\|^{\alpha+2}, \tag{32}$$

$$c_2\left(\frac{c_2}{\alpha+2}\|\theta - \theta^*\|\right)^{\alpha+1} \leq \|\nabla f(\theta)\| \leq c_1\|\theta - \theta^*\|^{\alpha+1}. \tag{33}$$

From the definitions of the population and sample Polyak operators $F_n$ and $F$ in equations (5) and (6), we make the following decomposition on $\|F_n(\theta) - F(\theta)\|$:

$$\begin{aligned}\|F_n(\theta) - F(\theta)\| &= \left\|\frac{f_n(\theta) - f_n(\widehat{\theta}_n)}{\|\nabla f_n(\theta)\|^2}\nabla f_n(\theta) - \frac{f(\theta) - f(\theta^*)}{\|\nabla f(\theta)\|^2}\nabla f(\theta)\right\| \\ &\leq \left\|\left(\frac{f_n(\theta) - f_n(\widehat{\theta}_n)}{\|\nabla f_n(\theta)\|^2} - \frac{f(\theta) - f(\theta^*)}{\|\nabla f(\theta)\|^2}\right)\nabla f_n(\theta)\right\| \\ &\quad + \left\|\frac{f(\theta) - f(\theta^*)}{\|\nabla f(\theta)\|^2}(\nabla f(\theta) - \nabla f_n(\theta))\right\| \\ &:= T_1 + T_2.\end{aligned} \tag{34}$$

**Upper bound on $T_2$:** We first deal with the second term $T_2$. With Assumption (W.3), with probability $1-\delta$ we have that

$$\begin{aligned}\left\|\frac{f(\theta)-f(\theta^*)}{\|\nabla f(\theta)\|^2}(\nabla f(\theta)-\nabla f_n(\theta))\right\| =& \frac{f(\theta)-f(\theta^*)}{\|\nabla f(\theta)\|^2}\|\nabla f(\theta)-\nabla f_n(\theta)\| \\ \leq& \frac{c_3 r^\gamma \varepsilon(n,\delta)}{c_2^2}(f(\theta)-f(\theta^*))^{\frac{2}{\alpha+2}-1}\end{aligned}$$

for all $\theta \in \mathbb{B}(\theta^*, r)\backslash\mathbb{B}(\theta^*, r_n)$ where $r < \rho$. Combining the above inequality with the inequality (32), we obtain

$$\left\|\frac{f(\theta)-f(\theta^*)}{\|\nabla f(\theta)\|^2}(\nabla f(\theta)-\nabla f_n(\theta))\right\| \leq \frac{c_3}{c_2^2}\cdot\left(\frac{c_1}{2}\right)^{\frac{2}{\alpha+2}-1} r^{\gamma-\alpha}\varepsilon(n,\delta) \tag{35}$$

for all $\theta \in \mathbb{B}(\theta^*, r)\backslash\mathbb{B}(\theta^*, r_n)$ where $r < \rho$.

**Upper bound on $T_1$:** For the first term $T_1$, we have that

$$\begin{aligned}&\left\|\left(\frac{f_n(\theta)-f_n(\widehat{\theta}_n)}{\|\nabla f_n(\theta)\|^2}-\frac{f(\theta)-f(\theta^*)}{\|\nabla f(\theta)\|^2}\right)\nabla f_n(\theta)\right\| \\ &\quad\leq \frac{\left|(f_n(\theta)-f_n(\widehat{\theta}_n))\|\nabla f(\theta)\|^2-(f(\theta)-f(\theta^*))\|\nabla f_n(\theta)\|^2\right|}{\|\nabla f_n(\theta)\|\|\nabla f(\theta)\|^2} \\ &\quad\leq \frac{\left|(f_n(\theta)-f_n(\widehat{\theta}_n)-f(\theta)+f(\theta^*))\|\nabla f(\theta)\|^2\right|+(f(\theta)-f(\theta^*))\left|\|\nabla f_n(\theta)\|^2-\|\nabla f(\theta)\|^2\right|}{(\|\nabla f(\theta)\|-c_3 r^\gamma\varepsilon(n,\delta))\|\nabla f(\theta)\|^2} \\ &\quad\leq \frac{|f_n(\theta)-f_n(\widehat{\theta}_n)-f(\theta)+f(\theta^*)|\|\nabla f(\theta)\|^2+2(f(\theta)-f(\theta^*))\|\nabla f(\theta)\|c_3 r^\gamma\varepsilon(n,\delta)}{(\|\nabla f(\theta)\|-c_3 r^\gamma\varepsilon(n,\delta))\|\nabla f(\theta)\|^2} \\ &\qquad\qquad+\frac{(f(\theta)-f(\theta^*))c_3^2 r^{2\gamma}\varepsilon^2(n,\delta)}{(\|\nabla f(\theta)\|-c_3 r^\gamma\varepsilon(n,\delta))\|\nabla f(\theta)\|^2}.\end{aligned} \tag{36}$$

To bound the RHS of equation (36), we need to upper bound

$$\begin{aligned}&f_n(\theta)-f(\theta)-(f_n(\widehat{\theta}_n)-f(\theta^*)) \\ &\qquad= f_n(\theta)-f_n(\theta^*)-(f(\theta)-f(\theta^*))-(f_n(\widehat{\theta}_n)-f_n(\theta^*))\end{aligned}$$

Indeed, from Assumption (W.3), with probability $1-\delta$ we have that

$$\begin{aligned}&f_n(\theta)-f_n(\theta^*)-(f(\theta)-f(\theta^*)) \\ &\qquad\leq \int_0^1 \|\nabla f_n(\theta^*+t(\theta-\theta^*))-\nabla f(\theta^*+t(\theta-\theta^*))\|dt \\ &\qquad\leq \frac{c_3 r^{\gamma+1}\varepsilon(n,\delta)}{\gamma+1}\end{aligned} \tag{37}$$

for any $\theta \in \mathbb{B}(\theta^*, r)$. Furthermore, we find that

$$|f_n(\widehat{\theta}_n)-f_n(\theta^*)| \leq |f_n(\widehat{\theta}_n)-f(\widehat{\theta}_n)-f_n(\theta^*)+f(\theta^*)|+|f(\widehat{\theta}_n)-f(\theta^*)|$$

$$\leq \frac{c_3 r^{\gamma+1}\varepsilon(n,\delta)}{\gamma+1} + \frac{c_1\|\widehat{\theta}_n - \theta^*\|^{\alpha+2}}{2}, \tag{38}$$

where the final inequality is due to inequalities (37) and (32). Plugging the bounds (37) and (38) into (36), we find that

$$\begin{aligned}
&\left\|\left(\frac{f_n(\theta)-f_n(\theta_n^*)}{\|\nabla f_n(\theta)\|^2} - \frac{f(\theta)-f(\theta^*)}{\|\nabla f(\theta)\|^2}\right)\nabla f_n(\theta)\right\| \\
&\qquad \leq \frac{\left(\frac{2c_3 r^{\gamma+1}\varepsilon(n,\delta)}{\gamma+1} + \frac{c_1\|\widehat{\theta}_n-\theta^*\|^{\alpha+2}}{2}\right) c_1^2 r^{2\alpha+2} + c_1^2 c_3 r^{2\alpha+3+\gamma}\varepsilon(n,\delta) + \frac{c_1 c_3^2}{2} r^{2\gamma+\alpha+2}\varepsilon^2(n,\delta)}{c_2^2\left(\frac{c_2 r}{(\alpha+2)}\right)^{2\alpha+2}\left(c_2\left(\frac{c_2 r}{(\alpha+2)}\right)^{\alpha+1} - c_3 r^{\gamma}\varepsilon(n,\delta)\right)}.
\end{aligned}$$

As $r \geq \bar{C}\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$ where $\bar{C} = \left(\frac{C\cdot c_3(\alpha+2)^{\alpha+1}}{c_2^{\alpha+2}}\right)^{\frac{1}{\alpha+1-\gamma}}$, we know $c_3 r^{\gamma}\varepsilon(n,\delta) \leq \frac{c_2}{C}\left(\frac{c_2 r}{\alpha+2}\right)^{\alpha+1}$, and we can simplify this term to

$$\begin{aligned}
&\left\|\left(\frac{f_n(\theta)-f_n(\theta_n^*)}{\|\nabla f_n(\theta)\|^2} - \frac{f(\theta)-f(\theta^*)}{\|\nabla f(\theta)\|^2}\right)\nabla f_n(\theta)\right\| \\
\leq &\frac{C}{C-1}\left(\frac{2c_1^2 c_3(\alpha+2)^{3\alpha+3}}{(\gamma+1)c_2^{3\alpha+6}} r^{\gamma-\alpha}\varepsilon(n,\delta) + \frac{c_1^3(\alpha+2)^{3\alpha+3}}{2(\gamma+1)c_2^{3\alpha+6}}\left(\frac{Cc_3(\alpha+2)^{3\alpha+3}}{c_2^{\alpha+2}}\varepsilon(n,\delta)\right)^{\frac{\alpha+2}{\alpha+1-\gamma}} r^{-\alpha-1}\right. \\
&\left. + \frac{c_1^2 c_3(\alpha+2)^{3\alpha+3}}{c_2^{3\alpha+6}} r^{\gamma-\alpha}\varepsilon(n,\delta) + \frac{c_1 c_3^2(\alpha+2)^{3\alpha+3}}{2c_2^{3\alpha+6}} r^{2\gamma-2\alpha-1}\varepsilon^2(n,\delta)\right) \\
\leq &\frac{C}{C-1}\left(\frac{2c_1^2 c_3(\alpha+2)^{3\alpha+3}}{(\gamma+1)c_2^{3\alpha+6}} r^{\gamma-\alpha}\varepsilon(n,\delta) + \frac{Cc_1^3 c_3(\alpha+2)^{4\alpha+4}}{2(\gamma+1)c_2^{4\alpha+8}} r^{\gamma-\alpha}\varepsilon(n,\delta)\right. \\
&\left. + \frac{c_1^2 c_3(\alpha+2)^{3\alpha+3}}{c_2^{3\alpha+6}} r^{\gamma-\alpha}\varepsilon(n,\delta) + + \frac{c_1 c_3(\alpha+2)^{2\alpha+2}}{2c_2^{2\alpha+4}C} r^{\gamma-\alpha}\varepsilon(n,\delta)\right),
\end{aligned} \tag{39}$$

for any $\theta \in \mathbb{B}(\theta^*, r)\backslash\mathbb{B}(\theta^*, r_n)$. Combining inequalities (35) and (39) and taking the constant $c_4$ accordingly, we can obtain the desired result.

### 5.3 Proof of Theorem 1

Recall that for the radius of $r_n$ in Lemma 2, we denote $r_n = \bar{C}\cdot\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$. Without loss of generality, we assume $\|\theta_n^k - \theta^*\| > \left(\frac{c_4\bar{C}^{\gamma-\alpha}}{1-\kappa}+1\right) r_n$ holds for all $k < T$ where $c_4$ is the universal constant in Lemma 2, $T := C\log(1/\varepsilon(n,\delta)$ and $C$ is some constant that will be chosen later; otherwise the conclusion of the theorem already holds.

We first show that, $\theta_n^k \in \mathbb{B}(\theta^*,\rho)\backslash\mathbb{B}(\theta^*, r_n)$ for all $k < T$. The inequality $\|\theta_n^k - \theta^*\| > r_n$ is direct from the hypothesis. Therefore, we only need to prove that $\|\theta_n^k - \theta^*\| \leq \rho$. Indeed, we have

$$\begin{aligned}
\|\theta_n^{k+1} - \theta^*\| =& \|F_n(\theta_n^k) - \theta^*\| \\
\leq& \|F_n(\theta_n^k) - F(\theta_n^k)\| + \|F(\theta_n^k) - \theta^*\| \\
\leq& \sup_{\theta\in\mathbb{B}(\theta^*,\rho)\backslash\mathbb{B}(\theta^*,r_n)} \|F_n(\theta) - F(\theta)\| + \|F(\theta_n^k) - \theta^*\| \\
\overset{(i)}{\leq}& c_4\rho^{\gamma-\alpha}\varepsilon(n,\delta) + \kappa\|\theta_n^k - \theta^*\|
\end{aligned}$$

$$\overset{(ii)}{\leq} c_4 \bar{C}^{\gamma-\alpha} \varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}} + \kappa\rho$$
$$\overset{(iii)}{\leq} \rho$$

with probability $1-\delta$ where the inequality (i) is due to Lemma 2 and $c_4$ is the universal constant in that lemma; the inequality (ii) is due to $\rho > r_n = \bar{C}\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$ and $\gamma \leq \alpha$; the inequality (iii) is due to the assumption that $n$ is sufficiently large such that $c_4\bar{C}^{\gamma-\alpha}\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}} \leq (1-\kappa)\rho$. As a consequence, we can guarantee that $\theta_n^k \in \mathbb{B}(\theta^*,\rho)\backslash\mathbb{B}(\theta^*,r_n)$ for all $k < T$.

Now, we would like to show that $\|\theta_n^T - \theta^*\| \leq \frac{2-\kappa}{1-\kappa} r_n$. Indeed, following the earlier argument, we find that

$$\begin{aligned}
\|\theta_n^T - \theta^*\| \leq & \|F_n(\theta_n^{T-1}) - F(\theta_n^{T-1})\| + \|F(\theta_n^{T-1}) - \theta^*\| \\
\leq & \sup_{\theta \in \mathbb{B}(\theta^*,\rho)\backslash\mathbb{B}(\theta^*,r_n)} \|F_n(\theta) - F(\theta)\| + \kappa\|\theta_n^{T-1} - \theta^*\| \\
\leq & c_4 \cdot r_n^{\gamma-\alpha}\varepsilon(n,\delta) + \kappa\|\theta_n^{T-1} - \theta^*\| \\
= & c_4\bar{C}^{\gamma-\alpha} \cdot \varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}} + \kappa\|\theta_n^{T-1} - \theta^*\|.
\end{aligned}$$

By repeating the above argument $T$ times, we finally obtain that

$$\begin{aligned}
\|\theta_n^T - \theta^*\| \leq & c_4\bar{C}^{\gamma-\alpha} \cdot \varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}} \left(\sum_{t=0}^{T-1} \kappa^t\right) + \kappa^T\|\theta_n^0 - \theta^*\| \\
\leq & \frac{c_4\bar{C}^{\gamma-\alpha}}{1-\kappa}\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}} + \kappa^T\rho.
\end{aligned}$$

By choosing $T$ such that $\kappa^T\rho \leq \varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$, which is equivalent to $T \geq \frac{\log(\rho) + \frac{1}{\alpha+1-\gamma}\log(1/\varepsilon(n,\delta))}{\log(1/\kappa)}$, we can guarantee that

$$\|\theta_n^T - \theta^*\| \leq \left(\frac{c_4\bar{C}^{\gamma-\alpha}}{1-\kappa} + 1\right)\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}.$$

As a consequence, we obtain the conclusion of the theorem.

# 6 Discussion

In this paper, we have provided statistical and computational complexities of the Polyak step size gradient descent iterates under the generalized smoothness and Łojasiewicz property of the population loss function as well as the uniform concentration bound between the gradients of the population and sample loss functions. Our results indicate that the Polyak step size iterates only take a logarithmic number of iterations to reach a final statistical radius, which is much fewer than the polynomial number of iterations of the fixed-step size gradient descent iterates to reach the same final statistical radius, when the population loss function is not locally strongly convex. Given that the complexity per iteration of the Polyak step size and fixed-step size gradient descent methods are similar, these results indicate that the Polyak step size gradient descent method is computationally more efficient than the fixed-step size gradient descent method in terms of the number of sample size when the dimension is fixed. Finally, we illustrate our findings under three statistical models: generalized linear model,

mixture model, and mixed linear regression model. A few natural future questions arising from our work.

First, our general theory for the convergence rate of the Polyak step size gradient descent iterates relies on the assumptions that the constants of the generalized smoothness and the generalized Łojasiewicz condition are similar. While this assumption is natural in several statistical models, there are also certain instances of statistical models that this requirement does not hold, such as general over-specified low rank matrix factorization problem, and factor analysis. Therefore, extending our theory of the Polyak step size gradient descent algorithm to the settings when the constants in these assumptions are not similar is of interest.

Second, our results are restricted to the settings of i.i.d. data in which we can define the corresponding population loss function of the sample loss function. In dependent settings, such as time series data, since the notion of population loss function is not well-defined, it is of interest to develop a new framework beyond the population to sample framework in the current paper to analyze the behavior of Polyak step size gradient descent method for solving the optimal solution of the sample loss function.

Finally, our results shed light on the favorable performance of adaptive gradient methods for dealing with the singular settings of the statistical models, namely, those settings when the Fisher information matrix around the true parameter is degenerate or close to be degenerate, which leads to the slow convergence rates of estimating the true parameters. For the future work, it is of practical interest to extend our general theoretical studies under these settings to other popular adaptive gradient descent methods, such as Adagrad [10] and Adam [22], that have been observed to have favorable performance in several machine learning and deep learning models.

# 7 Acknowledgements

This work was partially supported by the NSF IFML 2019844 award and research gifts by UT Austin ML grant to NH, and by NSF awards 1564000 and 1934932 to SS.

# A Proofs of remaining key results

In this appendix, we provide proofs for the generalized smoothness and PL conditions of the generalized linear model, over-specified mixture model, and over-specified mixed linear regression model in the main text.

## A.1 Generalized linear model

We first prove the local strong convexity (11) and uniform concentration bound (12) under the strong signal-to-noise regime in Section A.1.1. Then, we prove the generalized Łojasiewicz property (16) of the population loss $\mathcal{L}$ for the low signal-to-noise regime in Section A.1.2.

### A.1.1 Strong signal-to-noise regime

**Local strong convexity:** We first prove the local strong convexity in equation (11). Recall that, we have

$$\mathcal{L}(\theta) = \frac{1}{2}\left(\mathbb{E}\left[\left((X^{\top}\theta^{*})^{p} - (X^{\top}\theta)^{p}\right)^{2}\right] + \sigma^{2}\right),$$

where the outer expectation is taken with respect to $X \sim \mathcal{N}(0, I_d)$. Hence, $\mathcal{L}$ is a polynomial function with degree at most $2p$ and coefficients bounded (as for Gaussian we have any finite order moment bounded). So $\mathcal{L}$ should be smooth around the optima. Furthermore, when $\|\theta - \theta^*\|$ is small enough we have that

$$\left((X^\top\theta)^p - (X^\top\theta^*)^p\right)^2 = p(X^\top\theta^*)^{p-1}X^\top(\theta - \theta^*) + o(\|\theta - \theta^*\|).$$

Thus, we have that

$$\begin{aligned}\mathcal{L}(\theta) =& \frac{1}{2}\left(\mathbb{E}\left[\left((X^\top\theta^*)^p - (X^\top\theta)^p\right)^2\right] + \sigma^2\right)\\ =& \frac{p^2}{2}(\theta - \theta^*)^\top\mathbb{E}\left[X^\top(X^\top\theta^*)^{2p-2}X\right](\theta - \theta^*) + \frac{\sigma^2}{2} + o(\|\theta - \theta^*\|^2).\end{aligned}$$

As $2p - 2$ is even, it is clear that we have $\mathbb{E}\left[X^\top(X^\top\theta^*)^{2p-2}X\right]$ is positive definite matrix, which shows $\mathcal{L}$ is locally strongly convex function (by manipulating $\|\theta - \theta^*\|$ and the constant).

**Uniform concentration bound:** For the uniform concentration of the gradient in equation (12), direct calculations show that

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

Hence, with triangle inequality, we have that

$$\begin{aligned}\|\nabla\mathcal{L}_n(\theta) - \nabla\mathcal{L}(\theta)\| \le& \left\|\left(\frac{1}{n}\sum_{i=1}^n(Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-1}X_i\right)\right\|\\ &+ \left\|\left(\frac{1}{n}\sum_{i=1}^n(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-1}X_i - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-1}X]\right)\right\|\\ &+ \left\|\left(\frac{1}{n}\sum_{i=1}^n(X_i^\top\theta)^{2p-1}X_i - \mathbb{E}[(X^\top\theta)^{2p-1}X]\right)\right\|\\ :=& T_1 + T_2 + T_3.\end{aligned}$$

The first and the third terms $T_1$ and $T_3$ can be upper bounded via the identical method introduced in Section A.2 in [29] and we only need to change the radius from $r$ to $r + \|\theta^*\|$ when $\theta \in \mathbb{B}(\theta^*, r)$, namely, we have the following bounds:

$$T_1 \le c_1(r + \|\theta^*\|)^{p-1}\left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{1}{n}\left(d + \log\left(\frac{n}{\delta}\right)\right)^{p+1}\right), \tag{40}$$

$$T_3 \le c_2(r + \|\theta^*\|)^{2p-1}\left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{1}{n}\left(d + \log\left(\frac{n}{\delta}\right)\right)^{2p+1}\right) \tag{41}$$

with probability $1 - \delta$ where $c_1$ and $c_2$ are some universal constants. Therefore, it is sufficient to focus on the second term $T_2$. Without the loss of generality, we assume $\|\theta\| = 1$, and the results can be generalized to other norm of $\theta$ by rescaling. First, we know that

$$\left\|\left(\frac{1}{n}\sum_{i=1}^n(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-1}X_i - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-1}X]\right)\right\|$$

$$= \sup_{u\in\mathbb{S}^{d-1}} \left| \left( \frac{1}{n}\sum_{i=1}^{n} (X_i^\top \theta^*)^p (X_i^\top \theta)^{p-1} X_i^\top u - \mathbb{E}[(X^\top \theta^*)^p (X^\top \theta)^{p-1} X^\top u] \right) \right|,$$

where $\mathbb{S}^{d-1}$ is the unit norm Euclidean sphere in $\mathbb{R}^d$. With standard discretization arguments (e.g., Chapter 6 in [36]), let $U$ be a 1/8-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$ whose cardinality can be upper bounded by $17^d$, we know

$$\begin{aligned}
&\sup_{u\in\mathbb{S}^{d-1}} \left| \left( \frac{1}{n}\sum_{i=1}^{n} (X_i^\top \theta^*)^p (X_i^\top \theta)^{p-1} X_i^\top u - \mathbb{E}[(X^\top \theta^*)^p (X^\top \theta)^{p-1} X^\top u] \right) \right| \\
&\qquad \le 2 \sup_{u\in U} \left| \left( \frac{1}{n}\sum_{i=1}^{n} (X_i^\top \theta^*)^p (X_i^\top \theta)^{p-1} X_i^\top u - \mathbb{E}[(X^\top \theta^*)^p (X^\top \theta)^{p-1} X^\top u] \right) \right|.
\end{aligned}$$

Hence we can focus on the upper bound with a fixed $u$ where $\|u\| = 1$. We then apply a symmetrization argument (e.g., Theorem 4.10 in [36]), we have that, for any even integer $q$,

$$\begin{aligned}
&\mathbb{E}\left[ \left| \left( \frac{1}{n}\sum_{i=1}^{n} (X_i^\top \theta^*)^p (X_i^\top \theta)^{p-1} X_i^\top u - \mathbb{E}[(X^\top \theta^*)^p (X^\top \theta)^{p-1} X^\top u] \right) \right|^q \right] \\
&\qquad \le \mathbb{E}\left[ \left| \left( \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i (X_i^\top \theta^*)^p (X_i^\top \theta)^{p-1} X_i^\top u \right) \right|^q \right],
\end{aligned}$$

where $\{\varepsilon_i\}_{i\in[n]}$ is a set of i.i.d. Rademacher random variables. We then follow the proof strategy used in Section A.2 in [29]. For a compact set $\Omega$, define

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

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

$$\begin{aligned}
\mathcal{R}(\mathbb{S}^{d-1}) &= \sup_{\theta\in\mathbb{S}^{d-1}, p'\in[1,p]} \left| \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i (X_i^\top \theta^*)^p (X_i^\top \theta)^{p'-1} X_i^\top u \right| \\
&\le \sup_{\theta_t\in\mathcal{N}(t), \|\eta\|\le t, p'\in[1,p]} \left| \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i (X_i^\top \theta^*)^p (X_i^\top (\theta_t+\eta))^{p'-1} X_i^\top u \right| \\
&\le \sup_{\theta_t\in\mathcal{N}(t), p'\in[1,p]} \left| \frac{4}{n}\sum_{i=1}^{n} \varepsilon_i (X_i^\top \theta^*)^p (X_i^\top \theta_t)^{p'-1} X_i^\top u \right| \\
&\quad + \max_{p'\in[1,p]} 3^{p'-1} \left| \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i (X_i^\top \theta^*)^p (X_i^\top \eta)^{p'-1} X_i^\top u \right| \\
&\le 2\mathcal{R}(\mathcal{N}(t)) + 3^{p'-1} t \mathcal{R}(\mathbb{S}^{d-1}).
\end{aligned}$$

Take $t = 3^{-p}$, we have that $\mathcal{R}(\mathbb{S}^{d-1}) \le 3\mathcal{R}(\mathcal{N}(3^{-p}))$. We then move to the upper bound of $\mathcal{R}(\mathcal{N}(3^{-p}))$. With the union bound, for any $q \ge 1$ we have that

$$\sup_{\theta\in\mathbb{S}^{d-1}, p'\in[1,p]} \mathbb{E}\left[ \left| \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i (X_i^\top \theta^*)^p (X_i^\top \theta)^{p'-1} X_i^\top u \right|^q \right]$$

$$
\begin{aligned}
&= \sup_{\theta\in\mathbb{S}^{d-1},p'\in[1,p]} \int_0^\infty \mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-1} X_i^\top u\right|^q \geq \varepsilon\right) d\varepsilon \\
&\geq \sup_{\theta\in\mathcal{N}(3^{-p}),p'\in[1,p]} \int_0^\infty \mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-1} X_i^\top u\right|^q \geq \varepsilon\right) d\varepsilon \\
&\geq \frac{\sup_{p'\in[1,p]} \sum_{\theta\in\mathcal{N}(3^{-p})} \int_0^\infty \mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-1} X_i^\top u\right|^q \geq \varepsilon\right) d\varepsilon}{|\mathcal{N}(3^{-p})|} \\
&\geq \frac{\sup_{p'\in[1,p]} \int_0^\infty \mathbb{P}\left(\sup_{\theta\in\mathcal{N}(3^{-p})}\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-1} X_i^\top u\right|^q \geq \varepsilon\right) d\varepsilon}{|\mathcal{N}(3^{-p})|} \\
&\geq \frac{\int_0^\infty \mathbb{P}\left(\sup_{\theta\in\mathcal{N}(3^{-p}),p'\in[1,p]}\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-1} X_i^\top u\right|^q \geq \varepsilon\right) d\varepsilon}{p|\mathcal{N}(3^{-p})|} \\
&= \frac{\mathbb{E}[\mathcal{R}^q(\mathcal{N}(3^{-p}))]}{p|\mathcal{N}(3^{-p})|}.
\end{aligned}
$$

Hence, it's sufficient to consider $\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-1} X_i^\top u\right|^q\right]$. We apply Khintchine's inequality [4], which guarantees that there is an universal constant $C$, such that for all $p' \in [1,p]$, we have

$$
\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-1} X_i^\top u\right|^q\right] \leq \mathbb{E}\left[\left(\frac{Cq}{n^2}\sum_{i=1}^n (X_i^\top\theta^*)^{2p} (X_i^\top\theta)^{2(p'-1)} (X_i^\top u)^2\right)^{q/2}\right]
$$

To further upper bound the right hand side of the above equation, we consider the large deviation property of random variable $(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2(p'-1)}(X_i^\top u)^2$. It's straightforward to show that

$$
\begin{aligned}
\mathbb{E}\left[(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2(p'-1)}(X_i^\top u)^2\right] \leq & (2(p+p'))^{(p+p')}, \\
\mathbb{E}\left[\left((X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2(p'-1)}(X_i^\top u)^2\right)^{q/2}\right] \leq & (2(p+p')q)^{(p+p')q}.
\end{aligned}
$$

With Lemma 2 in [29], with probability at least $1-\delta$, we have

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

Hence, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2(p'-1)}(X_i^\top u)^2\right)^{q/2}\right] \\
\leq & 2^{q/2}\left(\mathbb{E}\left[(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2(p'-1)}(X_i^\top u)^2\right]\right)^{q/2} \\
&+ 2^{q/2}\mathbb{E}\left[\left|\sum_{i=1}^n \left((X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2(p'-1)}(X_i^\top u)^2\right)^{q/2} - \mathbb{E}\left[(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2(p'-1)}(X_i^\top u)^2\right]\right|^{q/2}\right]
\end{aligned}
$$

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

where $C'$ is a universal constant and $\Gamma(\cdot)$ is the Gamma function. Notice that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^*)^p(X_i^\top\theta)^p X_i^\top u - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)X^\top u]\right)\right|^q\right] \\
&\qquad \leq \mathbb{E}[\mathcal{R}^q(\mathbb{S}^{d-1})] \\
&\qquad \leq 3^q\mathbb{E}[\mathcal{R}(\mathcal{N}(3^{-p}))] \\
&\qquad \leq 3^q p|\mathcal{N}(3^{-p})| \sup_{\theta\in\mathbb{S}^{d-1}p'\in[1,p]} \mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p'-1}X_i^\top u\right|^q\right] \\
&\qquad \leq 3^q p(3^{p+1})^d\left(\frac{Cq}{n}\right)^{q/2}\left((16p)^{2pq} + 2C'pq\,(64p)^{pq}\,n^{-q/4}\Gamma(q/4)\right. \\
&\qquad \left.+(16p)^{(2p+1)q/2}n^{-q/2}\left((\log n)^{(2p+1)q/2} + \Gamma((2p+1)q/2)\right)\right),
\end{aligned}
$$

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

$$
\begin{aligned}
&\left(\mathbb{E}\left[\left\|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^*)^p(X_i^\top\theta)^p X_i - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)X]\right)\right\|^q\right]\right)^{1/q} \\
&\qquad \leq 2\left(\mathbb{E}\left[\sup_{u\in U}\left|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^*)^p(X_i^\top\theta)^p X_i^\top u - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)X^\top u]\right)\right|^q\right]\right)^{1/q} \\
&\qquad \leq 2\left(\mathbb{E}\left[\sum_{u\in[U]}\left|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^*)^p(X_i^\top\theta)^p X_i^\top u - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)X^\top u]\right)\right|^q\right]\right)^{1/q} \\
&\qquad \leq 2\cdot 17^{d/q}\sup_{u\in[U]}\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^*)^p(X_i^\top\theta)^p X_i^\top u - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)X^\top u]\right)\right|^q\right]^{1/q} \\
&\qquad \leq 6\cdot(17\cdot 3^{p+1})^{d/q}\left[\sqrt{\frac{C_p q}{n}} + \left(\frac{C_p q}{n}\right)^{3/4} + \frac{C_p}{n}(\log n + q)^{(2p+1)/2}\right],
\end{aligned}
$$

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

$$
T_2 \leq c_3(r + \|\theta^*\|)^{p-1}\left(\sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{1}{n}\left(d + \log\left(\frac{n}{\delta}\right)\right)^{\frac{2p+1}{2}}\right). \tag{42}
$$

Combining the bounds from equations (40), (42), and (41), as long as $n \geq C_1(d\log(d/\delta))^{2p}$ we have

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

where $C_1, C_2$ are some universal constants. Since $\|\theta^*\|$ is bounded away from 0, the above bound concludes our claim in equation (12).

### A.1.2 Low signal-to-noise regime

Now, we prove the generalized Łojasiewicz property (16) of the population loss $\mathcal{L}$ for the low signal-to-noise regime. Recall that, we assume $\theta^* = 0$. Now, we will demonstrate that for all $\theta \in \mathbb{B}(\theta^*, \rho)$ for some $\rho > 0$, we have

$$\|\nabla\mathcal{L}(\theta)\| \geq c_2(\mathcal{L}(\theta) - \mathcal{L}(\theta^*))^{1-\frac{1}{2p}}.$$

For the form of $\mathcal{L}(\theta)$, we have that

$$\begin{aligned}\nabla\mathcal{L}(\theta) &= 2p(2p-1)!!(\theta-\theta^*)\|\theta-\theta^*\|^{2p-2},\\ \|\nabla\mathcal{L}(\theta)\| &= 2p(2p-1)!!\|\theta-\theta^*\|^{2p-1}.\end{aligned}$$

Also, due to equation (13) we obtain that

$$\begin{aligned}(\mathcal{L}(\theta) - \mathcal{L}(\theta^*))^{1-\frac{1}{2p}} &= \left(\frac{(2p-1)!!\|\theta-\theta^*\|^{2p}}{2}\right)^{1-\frac{1}{2p}}\\ &= \left(\frac{(2p-1)!!}{2}\right)^{1-\frac{1}{2p}}\|\theta-\theta^*\|^{2p-1}.\end{aligned}$$

Thus, the Assumption (W.2) follows by selecting the constant $c_2 \leq \frac{2p(2p-1)!!}{\left(\frac{(2p-1)!!}{2}\right)^{1-\frac{1}{2p}}}$.

Next, with direct computation, we have

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

Notice that, $(\theta-\theta^*)(\theta-\theta^*)^\top$ is a rank-1 matrix, so the maximum eigenvalue of $\|\theta-\theta^*\|^2 I + (2p-4)(\theta-\theta^*)(\theta-\theta^*)^\top$ is $(2p-3)\|\theta-\theta^*\|^2$, hence $\lambda_{\max}(\nabla^2\mathcal{L}(\theta)) = 2p(2p-3)(2p-1)!!\|\theta-\theta^*\|^{2p-2}$, which confirms our claim of Assumption (W.1).

## A.2 Over-specified mixture model

We first present a proof of claim (23) about the generalized PL property of the population log-likelihood function $\bar{\mathcal{L}}$ of low-signal-to-noise symmetric two-component Gaussian mixture model in Appendix A.2.1. Then, in Appendix A.2.2, we present a proof of claim (22) about the local smoothness of $\bar{\mathcal{L}}$.

### A.2.1 Proof of claim (23)

Recall that $\theta^* = 0$ and the population log-likelihood function is given by:

$$\bar{\mathcal{L}}(\theta) = -\mathbb{E}_X \left[ \log \left( \frac{1}{2}\phi(X|\theta, \sigma^2 I_d) + \frac{1}{2}\phi(X| - \theta, \sigma^2 I_d) \right) \right],$$

where the outer expectation is taken with respect to $X \sim \mathcal{N}(\theta^*, I_d)$. Using $Z$ to absorb the constant that is independent of $\theta$, we have

$$\bar{\mathcal{L}}(\theta) = \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_X \left[ \log \left( \exp\left( -\frac{X^\top \theta}{\sigma^2} \right) + \exp\left( \frac{X^\top \theta}{\sigma^2} \right) \right) \right] + Z.$$

It indicates that

$$\bar{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta^*) = \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_X \left[ \log \left( \exp\left( -\frac{X^\top \theta}{\sigma^2} \right) + \exp\left( \frac{X^\top \theta}{\sigma^2} \right) \right) \right] + Z.$$

To simplify the calculation, we perform a change of coordinates via an orthogonal matrix $R$ such that $R\theta = \|\theta\| e_1$ where $e_1$ denotes the first canonical basis in dimension $d$. By denoting $V = RX/\sigma$, we have $V = (V_1, \ldots, V_d) \sim \mathcal{N}(0, I_d)$. Therefore, we can rewrite the above equation as follows:

$$\bar{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta^*) = \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_{V_1} \left[ \log \left( \exp\left( -\frac{V_1 \|\theta\|}{\sigma} \right) + \exp\left( \frac{V_1 \|\theta\|}{\sigma} \right) \right) \right] + Z,$$

where the outer expectation is taken with respect to $V_1 \sim \mathcal{N}(0, 1)$. By using the basic inequality $\exp(-x) + \exp(x) \geq 2 + x^2$ for all $x \in \mathbb{R}$, we find that

$$\bar{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta^*) \leq \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_{V_1} \left[ \log \left( 1 + \frac{V_1^2 \|\theta\|^2}{2\sigma^2} \right) \right],$$

Applying further the inequality $\log(1 + x) \geq x - \frac{x^2}{2}$ for all $x \geq 0$, we have

$$\bar{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta^*) \leq \frac{3\|\theta\|^4}{8\sigma^4}. \tag{43}$$

Now, we proceed to lower bound $\|\nabla \bar{\mathcal{L}}(\theta)\|$. Direct calculation leads to

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

Direct application of the triangle inequality with $\|.\|$ norm indicates that

$$\|\nabla \bar{\mathcal{L}}(\theta)\| \geq \frac{1}{\sigma^2} \left( \|\theta\| - \left\| \mathbb{E}_X \left( X \tanh(\frac{X^\top \theta}{\sigma^2}) \right) \right\| \right).$$

Using the similar change of coordinates as we did earlier, we obtain that

$$\left\| \mathbb{E}_X \left( X \tanh(\frac{X^\top \theta}{\sigma^2}) \right) \right\| = \sigma \mathbb{E}_{V_1} \left( V_1 \tanh(\frac{V_1 \|\theta\|}{\sigma}) \right),$$

where the outer expectation is taken with respect to $V_1 \sim \mathcal{N}(0,1)$. An application of the inequality $x\tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15}$ for all $x \in \mathbb{R}$ leads to

$$\begin{aligned}\sigma\mathbb{E}_{V_1}\left(V_1\tanh(\frac{V_1\|\theta\|}{\sigma})\right) &\leq \frac{\sigma^2}{\|\theta\|}\mathbb{E}_{V_1}\left(\frac{V_1^2\|\theta\|^2}{\sigma^2} - \frac{V_1^4\|\theta\|^4}{3\sigma^4} + \frac{2V_1^6\|\theta\|^6}{15\sigma^6}\right) \\ &= \|\theta\| - \frac{\|\theta\|^3}{\sigma^2} + \frac{2\|\theta\|^5}{\sigma^4}.\end{aligned}$$

As long as $\|\theta\| \leq \frac{\sigma}{2}$, we have $2\|\theta\|^5/\sigma^4 \leq \|\theta\|^3/(2\sigma^2)$. Putting the above inequalities together, we find that

$$\|\nabla\bar{\mathcal{L}}(\theta)\| \geq \frac{\|\theta\|^3}{2\sigma^4} \tag{44}$$

when $\|\theta\| \leq \frac{\sigma}{2}$. Combining the results of equations (43) and (44), we obtain

$$\|\nabla\bar{\mathcal{L}}(\theta)\| \geq c_2\left(\bar{\mathcal{L}}(\theta) - \bar{\mathcal{L}}(\theta^*)\right)^{\frac{3}{4}}$$

when $\|\theta\| \leq \frac{\sigma}{2}$ where $c_2$ is some universal constant. Therefore, we obtain the conclusion of claim (23).

### A.2.2 Proof of claim (22)

Direct calculation shows that

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

where $\text{sech}^2(x) = \frac{4}{(\exp(-x)+\exp(x))^2}$ for all $x \in \mathbb{R}$. Via an application of the change of coordinates that we used earlier, we can write the above equation as:

$$\nabla^2\bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^2}\left(I_d - \mathbb{E}_V\left(VV^\top\text{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right)\right),$$

where the outer expectation is taken with respect to $V = (V_1, V_2, \ldots, V_d) \sim \mathcal{N}(0, I_d)$. The matrix $A = \mathbb{E}_V\left(VV^\top\text{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right)$ is a diagonal matrix that $A_{11} = \mathbb{E}_{V_1}\left[V_1^2\text{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right]$ and $V_{ii} = \mathbb{E}_{V_1}\left[\text{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right]$ for all $2 \leq i \leq d$.

An application of the inequality $\text{sech}^2(x) \geq 1 - x^2$ for all $x \in \mathbb{R}$ leads to

$$\begin{aligned}A_{11} &\geq \mathbb{E}_{V_1}\left[V_1^2\left(1 - \frac{V_1^2\|\theta\|^2}{\sigma^2}\right)\right] = 1 - \frac{3\|\theta\|^2}{\sigma^2}, \\ A_{ii} &\geq \mathbb{E}_{V_1}\left[1 - \frac{V_1^2\|\theta\|^2}{\sigma^2}\right] = 1 - \frac{\|\theta\|^2}{\sigma^2},\end{aligned}$$

for all $i \neq 1$. These results indicate that

$$\lambda_{\max}(\nabla^2\bar{\mathcal{L}}(\theta)) \leq \frac{3\|\theta\|^2}{\sigma^4}.$$

As a consequence, we obtain the conclusion of claim (22).

### A.3 Mixed linear regression model

We first present a proof of claim (30) about the generalized PL property of the population log-likelihood function $\tilde{\mathcal{L}}$ of low-signal-to-noise symmetric two-component Gaussian mixed linear regression in Appendix A.3.1. Then, in Appendix A.3.2, we present a proof of claim (29) about the local smoothness of $\tilde{\mathcal{L}}$. The proof ideas of these claims are similar to those in the mixture model case. Here, we provide the proofs for the completeness.

#### A.3.1 Proof of claim (30)

When $\theta^* = 0$, we have that $Y \sim \mathcal{N}(0, \sigma^2)$. Furthermore, the population log-likelihood function $\tilde{\mathcal{L}}$ admits the following form:

$$\tilde{\mathcal{L}}(\theta) = -\mathbb{E}_{X,Y}\left[\log\left(\frac{1}{2}\phi(Y|X^\top\theta, \sigma^2) + \frac{1}{2}\phi(Y|-X^\top\theta, \sigma^2)\right)\right].$$

Using $Z$ to absorb the constant that is independent of $\theta$, when $X \sim \mathcal{N}(0, I_d)$ and $Y|X \sim \mathcal{N}(Y|0, \sigma^2)$ we have

$$\tilde{\mathcal{L}}(\theta) = \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_{X,Y}\left[\log\left(\exp\left(\frac{Y\theta^\top X}{\sigma^2}\right) + \exp\left(\frac{-Y\theta^\top X}{\sigma^2}\right)\right)\right] + Z.$$

Similar to the proof of claim (23), to bound the expectation in the above equation we can perform a change of coordinates using an orthonormal matrix $R$ such that $R\theta = \|\theta\|e_1$. Let $V = RX$, then $V = (V_1, V_2, \ldots, V_d) \sim \mathcal{N}(0, I_d)$. Moreover, since $\tilde{\mathcal{L}}(\theta^*)$ does not depend on $\theta$, we can write:

$$\begin{aligned}
\tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\theta^*) &= \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_{X,Y}\left[\log\left(\exp\left(\frac{Y\theta^\top X}{\sigma^2}\right) + \exp\left(\frac{-Y\theta^\top X}{\sigma^2}\right)\right)\right] + Z \\
&= \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_{V_1,Y}\left[\log\left(\exp\left(\frac{Y\|\theta\|V_1}{\sigma^2}\right) + \exp\left(\frac{-Y\|\theta\|V_1}{\sigma^2}\right)\right)\right] + Z.
\end{aligned}$$

Using the standard inequality $\exp(-x) + \exp(x) \geq 2 + x^2$ for all $x \in \mathbb{R}$ we find that

$$\tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\theta^*) \leq \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_{V_1,Y}\left[\log\left(1 + \frac{Y^2\|\theta\|^2V_1^2}{2\sigma^4}\right)\right].$$

From here, the inequality $\log(1+x) \geq x - \frac{x^2}{2}$ for all $x \geq 0$ leads to

$$\begin{aligned}
\tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\theta^*) &\leq \frac{\|\theta\|^2}{2\sigma^2} - \mathbb{E}_{V_1,Y}\left[\frac{Y^2\|\theta\|^2V_1^2}{2\sigma^4} - \frac{Y^4\|\theta\|^4V_1^4}{8\sigma^8}\right] \\
&= \frac{\|\theta\|^2}{2\sigma^2} - \frac{\mathbb{E}_Y[Y^2]\|\theta\|^2\mathbb{E}_{V_1}[V_1^2]}{2\sigma^4} + \frac{\mathbb{E}_Y[Y^4]\|\theta\|^4\mathbb{E}_{V_1}[V_1^4]}{8\sigma^8} \\
&= \frac{9}{8\sigma^4}\|\theta\|^4. \qquad (45)
\end{aligned}$$

Now, we establish an lower bound for $\|\nabla\tilde{\mathcal{L}}(\theta)\|$. Indeed, direct calculation shows that

$$\nabla\tilde{\mathcal{L}}(\theta) = \frac{1}{\sigma^2}\left(\theta - \mathbb{E}_{X,Y}\left[YX\tanh(\frac{Y\theta^\top X}{\sigma^2})\right]\right)$$

Therefore, we find that $\|\nabla\tilde{\mathcal{L}}(\theta)\| \geq \frac{1}{\sigma^2}\left(\|\theta\| - \left\|\mathbb{E}_{X,Y}\left[YX\tanh(\frac{Y\theta^\top X}{\sigma^2})\right]\right\|\right)$. Using the earlier change of coordinates, we have

$$\left\|\mathbb{E}_{X,Y}\left[YX\tanh(\frac{Y\theta^\top X}{\sigma^2})\right]\right\| = \mathbb{E}_{V_1,Y}\left[YV_1\tanh(\frac{YV_1\|\theta\|}{\sigma^2})\right].$$

As we have the inequality $x\tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15}$ for all $x \in \mathbb{R}$, we obtain

$$\begin{aligned}\left\|\mathbb{E}_{X,Y}\left[YX\tanh(\frac{Y\theta^\top X}{\sigma^2})\right]\right\| &\leq \frac{\sigma^2}{\|\theta\|}\mathbb{E}_{V_1,Y}\left[\frac{Y^2V_1^2\|\theta\|^2}{\sigma^4} - \frac{Y^4V_1^4\|\theta\|^4}{3\sigma^8} + \frac{2Y^6V_1^6\|\theta\|^6}{15\sigma^{12}}\right] \\ &\leq \|\theta\| - \frac{3}{\sigma^2}\|\theta\|^3 + \frac{30}{\sigma^4}\|\theta\|^5 \leq \|\theta\| - \frac{3}{2\sigma^2}\|\theta\|^3,\end{aligned}$$

as long as $\|\theta\| \leq \frac{\sigma}{\sqrt{20}}$. Putting the above results together, we find that

$$\|\nabla\tilde{\mathcal{L}}(\theta)\| \geq \frac{3}{2\sigma^4}\|\theta\|^3. \tag{46}$$

A combination of the results from equation (45) and (46) indicate that

$$\|\nabla\tilde{\mathcal{L}}(\theta)\| \geq c_2\left(\tilde{\mathcal{L}}(\theta) - \tilde{\mathcal{L}}(\theta^*)\right)^{3/4},$$

for all $\|\theta\| \leq \frac{\sigma}{\sqrt{20}}$ where $c_2$ is some universal constant. As a consequence, we obtain the conclusion of claim (30).

### A.3.2 Proof of claim (29)

Similar to the proof of claim (22), we have

$$\begin{aligned}\nabla^2\tilde{\mathcal{L}}(\theta) &= \frac{1}{\sigma^2}\left(I_d - \frac{1}{\sigma^2}\mathbb{E}_{X,Y}\left[Y^2XX^\top\mathrm{sech}^2(\frac{Y\theta^\top X}{\sigma^2})\right]\right) \\ &= \frac{1}{\sigma^2}\left(I_d - \frac{1}{\sigma^2}\mathbb{E}_{Y,V}\left[Y^2VV^\top\mathrm{sech}^2(\frac{YV_1\|\theta\|}{\sigma^2})\right]\right),\end{aligned}$$

where the second equality is from the change of coordinates $R = VX$ and $R$ is an orthogonal matrix such that $R\theta = \|\theta\|e_1$. Here, the outer expectation is taken with respect to $Y \sim \mathcal{N}(0,\sigma^2)$ and $V = (V_1,\ldots,V_d) \sim \mathcal{N}(0, I_d)$.

The matrix $B = \frac{1}{\sigma^2}\mathbb{E}_{Y,V}\left[Y^2VV^\top\mathrm{sech}^2(\frac{YV_1\|\theta\|}{\sigma^2})\right]$ is a diagonal matrix such that $B_{11} = \mathbb{E}_{Y,V_1}\left[Y^2V_1^2\mathrm{sech}^2(\frac{YV_1\|\theta\|}{\sigma^2})\right]$ and $B_{ii} = \mathbb{E}_{Y,V_1}\left[Y^2\mathrm{sech}^2(\frac{YV_1\|\theta\|}{\sigma^2})\right]$ for all $i \neq 1$. Using the standard inequality $\mathrm{sech}^2(x) \geq 1 - x^2$ for all $x \in \mathbb{R}$ yields

$$\begin{aligned}B_{11} &\geq \mathbb{E}_{Y,V_1}\left[Y^2V_1^2 - \frac{Y^4V_1^4\|\theta\|^2}{\sigma^4}\right] = \sigma^2 - 9\|\theta\|^2, \\ B_{ii} &\geq \mathbb{E}_{Y,V_1}\left[Y^2 - \frac{Y^4V_1^2\|\theta\|^2}{\sigma^4}\right] = \sigma^2 - 3\|\theta\|^2,\end{aligned}$$

for all $i \neq 1$. Collecting the above results, we obtain

$$\lambda_{\max}(\nabla^2\tilde{\mathcal{L}}(\theta)) \leq \frac{9}{\sigma^2}\|\theta\|^2.$$

Hence, we obtain the conclusion of claim (29).

## B Auxiliary results

**Lemma 3.** *If Assumption (W.2) holds, then for all $\theta \in \mathbb{B}(\theta^*, \rho)$, we have that*

$$\|\theta - \theta^*\| \leq \frac{\alpha+2}{c_2}(f(\theta) - f(\theta^*))^{\frac{1}{\alpha+2}}.$$

*Furthermore, we have*

$$\|\nabla f(\theta)\| \geq c_2 \left(\frac{c_2}{\alpha+2}\|\theta - \theta^*\|\right)^{\alpha+1}.$$

*Proof.* The proof idea originates from the proof of Theorem 27 in [3]. We start from the gradient flow:

$$\frac{d\theta(t)}{dt} = -\nabla f(\theta(t)).$$

By the convexity, we have that

$$\frac{d\|\theta(t) - \theta^*\|_2^2}{dt} = 2\left\langle \theta(t) - \theta^*, \frac{d\theta(t)}{dt}\right\rangle = -2\langle \theta(t) - \theta^*, \nabla f(\theta(t))\rangle \leq 0,$$

which means if $\theta(0) \in \mathbb{B}(\theta^*, \rho)$, $\theta(t) \in \mathbb{B}(\theta^*, \rho), \forall t \geq 0$. Meanwhile, $\theta(t) \to \theta^*$ when $t \to \infty$. We then conclude the proof by

$$\begin{aligned}
(f(\theta(0)) - f(\theta^*))^{\frac{1}{\alpha+2}} &= \int_{\infty}^{0} d(f(\theta(t)) - f(\theta^*))^{\frac{1}{\alpha+2}} \\
&= \int_0^{\infty} \frac{f(\theta(t)) - f(\theta^*))^{\frac{1}{\alpha+2}-1}}{\alpha+2}\|\nabla f(\theta(t))\|^2 dt \\
&\geq \int_0^{\infty} \frac{c_2}{\alpha+2}\|\nabla f(\theta(t))\| dt \\
&= \int_0^{\infty} \frac{c_2}{\alpha+2}\left\|\frac{d\theta(t)}{dt}\right\| dt \\
&= \frac{c_2}{\alpha+2}\|\theta(0) - \theta^*\|.
\end{aligned}$$

The second argument can be directly obtained via Assumption (W.2), which concludes our proof. □

**Lemma 4.** *Under Assumptions (W.1) and (W.2), there exists a universal constant $c_0 > 0$ depending on the constants of these assumptions such that*

$$\|\theta_{GD}^t - \theta^*\| \leq \frac{c_0}{(\eta t)^{1/\alpha}},$$

*where $\theta_{GD}^{t+1} = \theta_{GD}^t - \eta\nabla f(\theta_{GD}^t)$ are the fixed-step size gradient descent iterates for minimizing the population loss function $f$. Furthermore, this bound is tight, means there are population loss functions $f$ satisfying Assumptions (W.1) and (W.2) and*

$$\|\theta_{GD}^t - \theta^*\| \geq \frac{c_0}{(\eta t)^{1/\alpha}}.$$

*Proof.* Our proof idea originates from [28] and we include it for completeness. We start from the following lemma.

**Lemma 5** (Lemma 3.5 in [5]). *If $f$ is $\beta$-smooth, then $\forall \theta_1, \theta_2 \in \mathbb{R}^d$, we have that*

$$f(\theta_1) - f(\theta_2) \leq \langle \nabla f(\theta_1), \theta_1 - \theta_2 \rangle - \frac{1}{2\beta} \|\nabla f(\theta_1) - \nabla f(\theta_2)\|^2.$$

**Corollary 4.** *If $f$ is $\beta$-smooth, then $\forall \theta_1, \theta_2 \in \mathbb{R}^d$, we have that*

$$\frac{1}{\beta} \|\nabla f(\theta_1) - \nabla f(\theta_2)\|^2 \leq \langle \nabla f(\theta_1) - \nabla f(\theta_2), x - y \rangle.$$

Notice that, if $\theta_1, \theta_2 \in \mathbb{B}(\theta^*, r)$, then $\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq c_1 r^\alpha \|\theta_1 - \theta_2\|$, which means $f$ is $c_1 r^\alpha$-smooth in $\mathbb{B}(\theta^*, r)$. We assume the step-size satisfies $0 < \eta < \frac{2}{c_1 r^\alpha}$, and define the "effective step-size" $\frac{1}{\beta} := \eta(2 - c_1 r^\alpha \eta) > 0$ where $\beta > c_1 r^\alpha$. If $\theta^t_{\mathrm{GD}} \in \mathbb{B}(\theta^*, r)$, we have that

$$\begin{aligned}
\|\theta^{t+1} - \theta^*\|^2 - \|\theta^t - \theta^*\|^2 =& \eta^2 \|\nabla f(\theta^t)\|^2 - 2\eta \langle \nabla f(\theta^t), \theta^t - \theta^* \rangle \\
\leq& -\frac{1}{\beta} \langle \nabla f(\theta^t), \theta^t - \theta^* \rangle \leq 0,
\end{aligned}$$

where the last inequality is due to Corollary 4. Hence, $\theta^{t+1}_{\mathrm{GD}} \in \mathbb{B}(\theta^*, r)$. Furthermore, from the generalized smoothness property of the function $f$ in Assumption (W.1) we have

$$\begin{aligned}
f(\theta^{t+1}) - f(\theta^t) \leq& \nabla f(\theta^t)^\top (\theta^{t+1} - \theta^t) + \frac{c_1 r^\alpha}{2} \|\theta^{t+1} - \theta^t\|^2 \\
=& -\frac{1}{2\beta} \|\nabla f(\theta^t)\|^2 \\
\leq& -\frac{c_2^2}{2\beta} (f(\theta^t) - f(\theta^*))^{2 - \frac{2}{\alpha+2}} \leq 0.
\end{aligned}$$

**Lemma 6.** *Given $\alpha > 0$, $\forall x \in [0, 1]$,*

$$\frac{1}{\alpha}(1 - x^\alpha) \geq x^\alpha (1 - x).$$

*Proof.* Consider the mapping $g : x \mapsto \frac{1}{\alpha}(x^\alpha - 1) - x^\alpha(1 - x)$.. We can see $g(0) = \frac{1}{\alpha}$ and $g(1) = 0$. Moreover,

$$\nabla g(x) = -(\alpha + 1)(x^{\alpha - 1} - x^\alpha) \leq 0,$$

which concludes the proof. □

Define $\delta(\theta^t) := f(\theta^t) - f(\theta^*)$, we have that

$$\begin{aligned}
\frac{1}{\delta(\theta^t)^{\frac{\alpha}{\alpha+2}}} =& \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \left( \frac{1}{\delta(\theta^s)^{\frac{\alpha}{\alpha+2}}} - \frac{1}{\delta(\theta^{s+1})^{\frac{\alpha}{\alpha+2}}} \right) \\
=& \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \frac{\frac{\alpha}{\alpha+2}}{\delta(\theta^{s+1})^{\frac{\alpha}{\alpha+2}}} \cdot \frac{\alpha + 2}{\alpha} \cdot \left( 1 - \left( \frac{\delta(\theta^{s+1})}{\delta(\theta^s)} \right)^{\frac{\alpha}{\alpha+2}} \right) \\
\geq& \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \frac{\frac{\alpha}{\alpha+2}}{\delta(\theta^{s+1})^{\frac{\alpha}{\alpha+2}}} \cdot \left( \frac{\delta(\theta^{s+1})}{\delta(\theta^s)} \right)^{\frac{\alpha}{\alpha+2}} \left( 1 - \frac{\delta(\theta^{s+1})}{\delta(\theta^s)} \right)
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}}+\sum_{s=1}^{t-1}\frac{\frac{\alpha}{\alpha+2}}{\delta(\theta^s)^{2-\frac{2}{\alpha+2}}}\cdot(\delta(\theta^s)-\delta(\theta^{s+1}))\\
\geq&\frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}}+\sum_{s=1}^{t-1}\frac{\frac{\alpha}{\alpha+2}}{\delta(\theta^s)^{2-\frac{2}{\alpha+2}}}\cdot\frac{c_2^2}{2\beta}(\delta(\theta^t))^{2-\frac{2}{\alpha+2}}\\
=&\frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}}+\sum_{s=1}^{t-1}\frac{c_2^2\left(\frac{\alpha}{\alpha+2}\right)}{2\beta}\\
=&\frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}}+\frac{c_2^2\left(\frac{\alpha}{\alpha+2}\right)}{2\beta}\cdot(t-1).
\end{aligned}
$$

We can conclude that

$$
f(\theta^t)-f(\theta^*)\leq\left[\frac{1}{(f(\theta^1)-f(\theta^*))^{\frac{\alpha}{\alpha+2}}}+\frac{c_2^2\left(\frac{\alpha}{\alpha+2}\right)}{2\beta}\cdot(t-1)\right]^{-\frac{\alpha+2}{\alpha}}\leq C(\eta\cdot t)^{-\frac{\alpha+2}{\alpha}},
$$

where $C$ is some universal constant. Combined Lemma 3 with the upper bound of $f(\theta^t)-f(\theta^*)$, we obtain that $\|\theta^t-\theta^*\|\leq c_0(\eta\cdot t)^{-1/\alpha}$ where $c_0$ is some universal constant. As a consequence, we reach the upper bound stated in Lemma 4.

For the tightness, consider the function $f:\mathbb{R}\to\mathbb{R}$, $f(\theta)=\frac{|\theta|^{\alpha+2}}{\alpha+2}$, which satisfies Assumptions (W.1) and (W.2). Consider the continuous limit of the fixed-step size gradient descent (i.e., the limit $\eta\to 0$) starting from $\theta^0=1$, which corresponds to the following ODE:

$$
\frac{d\theta}{dt}=-|\theta|^{\alpha+1},\quad\theta(0)=1.
$$

The solution of the ODE can be written as:

$$
\theta(t)=(t+1)^{-1/\alpha}.
$$

Notice that the $t$ in the solution of ODE is equivalent to $\eta t$ in the gradient descent dynamics, which concludes the proof. □

# C Beyond homogeneous assumptions

In this Appendix, we provide a brief discussion on the behaviors of the Polyak step size gradient descent iterates when the constants in Assumptions (W.1) and (W.2) are different. In particular, we consider the following two-dimensional population loss function $f(\theta)=\theta_1^2+\theta_2^4$ for all $\theta=(\theta_1,\theta_2)\in\mathbb{R}^2$. Under this case, the optima is $(0,0)$, and the updates of the Polyak step size gradient descent algorithm are given by:

$$
\begin{aligned}
\theta_1^{t+1}&=\theta_1^t-\frac{(\theta_1^t)^3+\theta_1^t(\theta_2^t)^4}{2(\theta_1^t)^2+8(\theta_2^t)^6}=\frac{(\theta_1^t)^3+\theta_1^t(\theta_2^t)^4(8(\theta_2^t)^2-1)}{2(\theta_1^t)^2+8(\theta_2^t)^6},\\
\theta_2^{t+1}&=\theta_2^t-\frac{(\theta_1^t)^2(\theta_2^t)^3+(\theta_2^t)^7}{(\theta_1^t)^2+4(\theta_2^t)^6}=\frac{3(\theta_2^t)^7+(\theta_1^t)^2\theta_2^t(1-(\theta_2^t)^2)}{(\theta_1^t)^2+4(\theta_2^t)^6}.
\end{aligned}
$$

Consider the local convergence in $\mathbb{B}(0, \rho)$ for some sufficiently small radius $\rho$, such that $(\theta_2^t)^2 \ll 1/8$, which corresponds to the approximate update:

$$\theta_1^{t+1} \approx \theta_1^t \cdot \frac{(\theta_1^t)^2 - (\theta_2^t)^4}{2(\theta_1^t)^2 + 8(\theta_2^t)^6},$$
$$\theta_2^{t+1} \approx \theta_2^t \cdot \frac{(\theta_1^t)^2 + 3(\theta_2^t)^6}{(\theta_1^t)^2 + 4(\theta_2^t)^6}.$$

For $\theta_1$, the update is only stable when $\theta_1^t \geq C(\theta_2^t)^2$ where $C$ is some universal constant. However, in this regime, $\theta_2$ can converge slowly, as

$$\frac{(\theta_1^t)^2 + 3(\theta_2^t)^6}{(\theta_1^t)^2 + 4(\theta_2^t)^6} = 1 - \mathcal{O}((\theta_2^t)^2) \to 1 \quad (\text{as} \quad \theta_2^t \to 0).$$

On the other hand, if we want $\theta_2$ to converge linearly, we need $\theta_1^t = \mathcal{O}((\theta_2^t)^3)$. In this regime, the update of $\theta_1$ can be unstable, as

$$\frac{(\theta_1^t)^2 - (\theta_2^t)^4}{2(\theta_1^t)^2 + 8(\theta_2^t)^6} \geq C_1(\theta_2^t)^{-2}$$

where $C_1$ is some constant. Hence, it's pretty hard to characterize the behaviour of Polyak step-size gradient descent iterates when the constants in Assumption (W.1) and (W.2) are different. We leave the understanding of this case as an interesting future direction.

## 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. (Cited on page 2.)

[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. (Cited on pages 2, 12, and 14.)

[3] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017. (Cited on page 34.)

[4] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. (Cited on page 27.)

[5] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. (Cited on pages 1, 5, 20, and 35.)

[6] E. J. Candes, Y. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion, 2011. (Cited on page 9.)

[7] Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754, 2018. (Cited on page 2.)

[8] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, pages 1–33, 2018. (Cited on page 2.)

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–38, 1997. (Cited on page 13.)

[10] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011. (Cited on page 24.)

[11] 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. (Cited on page 2.)

[12] 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. (Cited on pages 2, 12, and 13.)

[13] J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21(15):2758–2769, Aug 1982. (Cited on page 9.)

[14] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In M. F. Balcan and K. 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. (Cited on page 2.)

[15] E. Hazan and S. M. Kakade. Revisiting the Polyak step size. *Arxiv Preprint Arxiv: 1905.00313*, 2019. (Cited on page 8.)

[16] 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. (Cited on pages 2, 7, and 8.)

[17] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016. (Cited on page 12.)

[18] N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10:271–307, 2016. (Cited on page 13.)

[19] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3, 1991. (Cited on page 13.)

[20] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994. (Cited on page 13.)

[21] A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102:1025–1038, 2007. (Cited on page 13.)

[22] D. P. Kingma and J. L. Ba. Adam: a method for stochastic optimization. In *ICLR*, 2015. (Cited on page 24.)

[23] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824, 2018. (Cited on page 2.)

[24] J. Y. Kwon, N. Ho, and C. Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *AISTATS*, 2021. (Cited on pages 2 and 14.)

[25] B. Lindsay. *Mixture Models: Theory, Geometry and Applications*. In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995. (Cited on page 11.)

[26] P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015. (Cited on page 2.)

[27] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs.* New York, 1988. (Cited on page 11.)

[28] J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. *arXiv preprint arXiv:2105.06072*, 2021. (Cited on page 35.)

[29] W. Mou, N. Ho, M. J. Wainwright, P. Bartlett, and M. I. Jordan. A diffusion process perspective on posterior contraction rates for parameters. *arXiv preprint arXiv:1909.00966*, 2019. (Cited on pages 10, 25, 26, and 27.)

[30] Y. Nesterov. *Lectures on Convex Optimization*. Springer, 2018. (Cited on page 1.)

[31] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015. (Cited on page 9.)

[32] B. T. Polyak. *Introduction to Optimization.* Optimization Software, Inc., New York, 1987. (Cited on pages 1 and 3.)

[33] J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:689–710, 2011. (Cited on page 12.)

[34] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. (Cited on page 9.)

[35] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974. (Cited on page 8.)

[36] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. (Cited on page 26.)

[37] F. Yang, S. Balakrishnan, and M. Wainwright. Statistical and computational guarantees for the Baum-Welch algorithm. *Journal of Machine Learning Research*, 18:1–53, 2017. (Cited on page 2.)

[38] X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015. (Cited on page 2.)

[39] X.-T. Yuan and T. Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(Apr):899–925, 2013. (Cited on page 2.)