Towards Statistical and Computational Complexities of Polyak Step Size Gradient Descent

Tongzheng Ren*, Fuheng Cui*, Alexia Atsidakou*, Sujay Sanghavi, Nhat Ho

UT Austin, AISTATS 2022

Table of contents

Introduction

Problem Studied

Consider optimization problems:

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

where n stands for the sample size of i.i.d. data $X_1, X_2, \ldots, X_n$ coming from an unknown distribution P_{θ^*} where θ^* is true but unknown parameter, and f_n is a given empirical loss function whose optimal solution $\widehat{\theta}_n$ can be used to approximate the true parameter.

We also needs to use *population to sample analysis*. Thus define the population loss $f(\cdot) := \mathbb{E}_{X^n}[f_n(\cdot)]$ where $X^n = (X_1, \ldots, X_n)$.

Problems and Our Contribution

Fixed-step size GD Problems:

- Statistical radius ($\|\hat{\theta}_n - \theta^*\|$), which is $\mathcal{O}\left((n/d)^{\frac{\alpha}{\alpha+1-\gamma}}\right)$
- Iteration needed: $\mathcal{O}\left(n^{\frac{\alpha}{\alpha+1-\gamma}+1}\right)$.
- Optimal computational complexity: $\mathcal{O}(n)$.

Our Work: We show that by using Polyak step size gradient descent method, 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. 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.

Main results

Assumptions

Assumption 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.

Assumption 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.

Assumption 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 $1 - \delta$.

Lemma

Define Polyak operators for population and sample loss:
$F(\theta) := \theta - \frac{f(\theta)-f(\theta^*)}{\|\nabla f(\theta)\|^2} \cdot \nabla f(\theta)$ and $F_n(\theta_n^t) = \theta_n^t - \frac{f_n(\theta_n^t)-f_n(\widehat{\theta}_n)}{\|\nabla f_n(\theta_n^t)\|^2} \cdot \nabla f_n(\theta_n^t)$.

Lemma 1

Assume that Assumptions 1 and 2 hold. Then, given the definition of Polyak population operator, 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 Assumptions1 and 2.

Lemma 2

Assume that Assumptions 1, 2 and 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 2 and 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
$\sup_{\theta \in \mathbb{B}(\theta^*, r) \setminus \mathbb{B}(\theta^*, r_n)} \|F_n(\theta) - F(\theta)\| \leq c_4 r^{\gamma-\alpha} \varepsilon(n,\delta)$.

Theorem

Theorem 1

Assume that Assumptions 1, 2 and 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 κ is defined in Lemma 1, c_4 and $\bar{C}$ are the universal constants in Lemma 2, and ρ 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:

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

Fixed-step v.s. Polyak-step

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)	$\alpha \geq 0$	$\gamma \geq 0$	$\log(1/\varepsilon(n,\delta))$	$\varepsilon(n,\delta)^{\frac{1}{\alpha+1-\gamma}}$

Adaptive Polyak Step Size GD (Hazan and Kakade (2019))

Algorithm 1: Adaptive Polyak Step Size Gradient Descent

Input: Sample loss function f_n, initialization θ_n^0, lower bound $\tilde{f}_0$ 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: $\bar{\theta}$

Examples and Experiments

Generalized Linear Model

We assume $(Y_1, X_1), \ldots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$ satisfy

$$Y_i = g\left(X_i^\top \theta^*\right) + \varepsilon_i, \quad \forall i \in [n]$$

where $g : \mathbb{R} \to \mathbb{R}$ is a given link function, θ^* is a true but unknown parameter, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. noises from $\mathcal{N}\left(0, \sigma^2\right)$ where $\sigma > 0$ is a given variance parameter. 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 $\mathcal{N}(0, I_d)$. We want to optimize the empirical loss:

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

Generalized Linear Model

For the generalized linear model 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 $\min_{1 \leq k \leq t} \left\|\theta_n^k - \theta^*\right\| \leq \bar{c}_1 \sqrt{\frac{d+\log(1/\delta)}{n}}$ for $t \geq \bar{c}_2 \log\left(\frac{n}{d+\log(1/\delta)}\right)$

(ii) Low signal-to-noise regime: When $\theta^* = 0$, we find that $\min_{1 \leq k \leq t} \left\|\theta_n^k - \theta^*\right\| \leq c_1' \left(\frac{d+\log(1/\delta)}{n}\right)^{1/(2p)}$ 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.

Generalized Linear Model

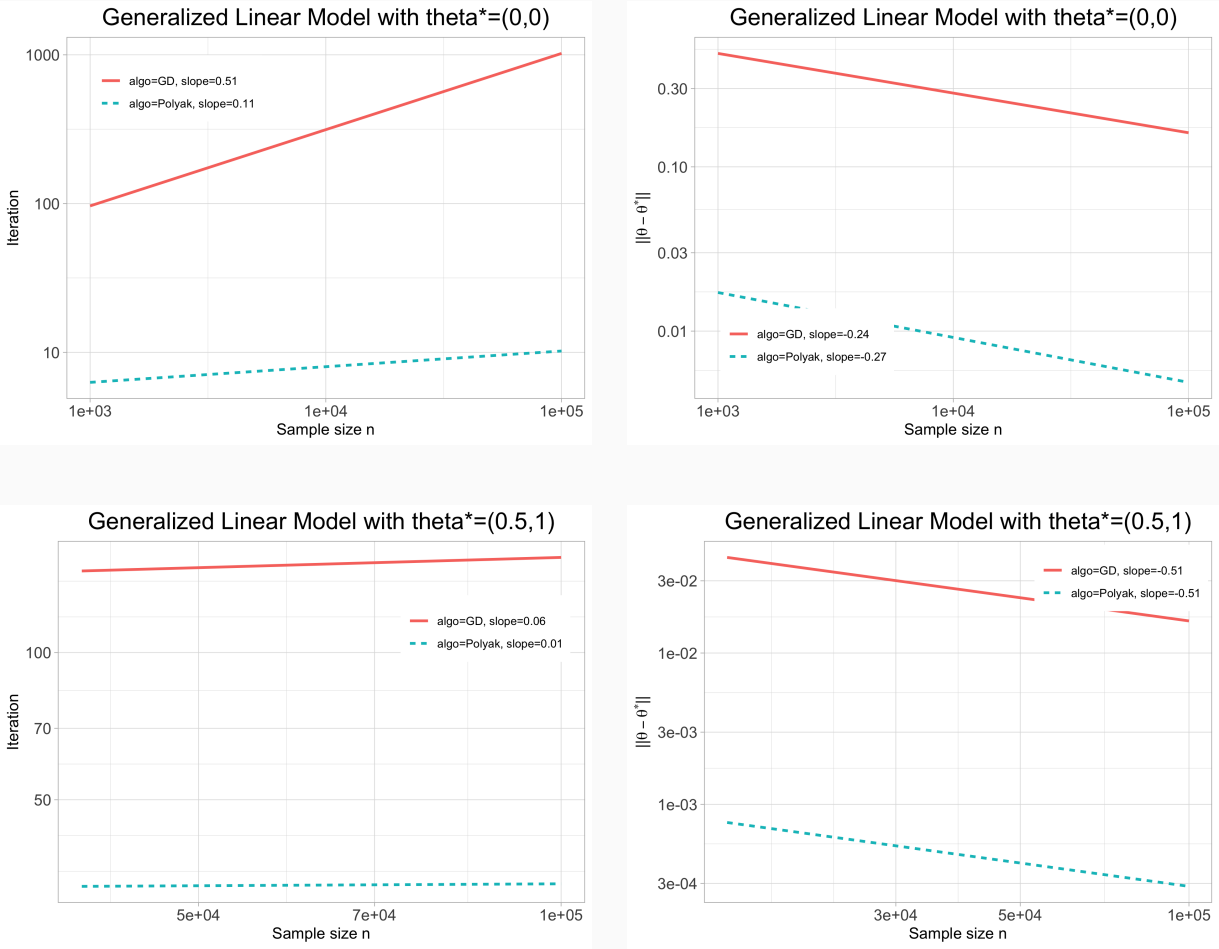

Figure 1: GLM with the link function $g(r) = r^2$.