2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Minimisation of Polyak-Łojasewicz Functions Using Random Zeroth-Order Oracles

Amir Ali Farzin and Iman Shames

***Abstract*— The application of a zeroth-order scheme for minimising Polyak-Łojasewicz (PL) functions is considered. The framework is based on exploiting a random oracle to estimate the function gradient. The convergence of the algorithm to a global minimum in the unconstrained case and to a neighbourhood of the global minimum in the constrained case along with their corresponding complexity bounds are presented. The theoretical results are demonstrated via numerical examples.**

I. Introduction

Zeroth-order (ZO) or derivative-free optimisation schemes are of interest when the gradient (or subgradient in case of non-differentiable cost functions) information is not readily available. A common scenario is the case where the value of the cost function, and not its higher order derivatives, is the only information available to the solver [1] and [2]. Sometimes, even if the gradient value is theoretically available, evaluating the gradient might incur high computational costs [3]. ZO methods provide a way forward for solving such optimisation problems as well. The majority of existing ZO methods aim to construct an estimate of the gradient of the function and use this estimate as a surrogate for the gradient. The method analysed in this paper is no exception to this general trend.

Various ZO optimisation methods have been designed and analysed for different problem classes. In [4], a constrained stochastic composite optimisation problem was studied where the function is possibly non-convex. The proposed algorithm in [4] relied on the existence of an unbiased variance-bounded estimator of the gradient. In [5], the authors considered the unconstrained zeroth-order optimisation problem and defined a random oracle which was an unbiased variance-bounded estimator of the gradient of a smoothed version of the original function.

The Polyak-Łojasiewicz (PL) inequality was first introduced in [6] and the convergence of gradient descent method under PL assumption was first proved there. It is observed that a range of different cost functions satisfy the PL condition [7]. Karimi et al. used the PL inequality to provide a new proof technique for analysing various first-order gradient descent methods and used proximal PL condition to analyse non-smooth cases [8]. In [9], a variant of the direct search method was employed to solve stochastic minimisation and saddle point problems. The direct search algorithm which is a derivative-free scheme and obtained the complexity bounds for the convergence of PL functions.

In this paper, we aim to fill a gap in the existing literature on random zeroth order oracles. We specifically leverage the class of random method proposed in [5] for optimisation problems, and establish its performance for the case where the cost functions satisfy the (proximal) PL condition. Specifically, we establish the convergence properties and the complexity bounds of these methods for both constrained and unconstrained cases. By doing so, we hope to shed some light on the behaviour of such algorithms when applied to a class of benignly nonconvex problems.

The outline of this paper is as follows. In Section II we introduce the necessary background information. The problems of interest are outlined in III. The convergence properties and the complexity of the algorithms for solving the problems of interest are presented in Section IV. Numerical examples are presented in V and conclusions and possible future research directions come in the end.

II. Preliminaries

In this section we provide the necessary definitions and background material required for presenting the results of this paper.

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$. The gaussian smoothed version of f, termed $f_\mu(x)$, is defined below:

$$\begin{aligned} f_\mu(x) &\stackrel{\text{def}}{=} \frac{1}{\kappa}\int_E f(x+\mu u)e^{-\frac{1}{2}||u||^2}\mathrm{d}u, \\ \kappa &\stackrel{\text{def}}{=} \int_E e^{-\frac{1}{2}||u||^2}\mathrm{d}u = \frac{(2\pi)^{n/2}}{[detB]^{\frac{1}{2}}}, \end{aligned} \tag{1}$$

where vector $u \in R^n$ is sampled from zero mean Gaussian distribution with positive definite correlation operator B^{-1} and the positive scalar smoothing parameter μ. Define the random oracle g_μ as

$$g_\mu(x) \stackrel{\text{def}}{=} \frac{f(x+\mu u)-f(x)}{\mu}Bu, \tag{2}$$

where u and B are defined above. The projection operator on a convex set $\mathcal{Z}$ is defined as

$$\mathrm{Proj}_{\mathcal{Z}}(x) \stackrel{\text{def}}{=} \arg\min_{z\in\mathcal{Z}}\|z-x\|^2,$$

where $\|\cdot\|$ is the Euclidean norm of its argument if it is a vector, and the corresponding induced norm if the argument is a matrix.

The authors are with CIICADA Lab, School of Engineering, ANU {amirali.farzin,iman.shames}@anu.edu.au

Copyright ©2024 EUCA

Definition 1 ($C^{1,1}$ Functions)**.** *The differentiable function $f(x) : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}^n$ as its domain is in $C^{1,1}$ if its gradient is Lipschitz, i.e.,*

$$||\nabla f(x) - \nabla f(y)|| \leq L_1(f)||x - y||, \forall x, y \in D, \quad (3)$$

where $L_1(f) > 0$ is the gradient Lipschitz constant*.*

Remark 1. *Any $f(x) \in C^{1,1}$ satisfies*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1(f)}{2}||y - x||^2. \quad (4)$$

Definition 2 (PL Functions [6])**.** *The differentiable function $f(x) : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}^n$ as its domain is termed a* PL function *if it satisfies the Polyak-Łojasewicz (PL) condition, i.e.,*

$$\frac{1}{2}||\nabla f(x)||^2 \geq l(f(x) - f^*), \ \forall x \in D, \quad (5)$$

where $l > 0$ is the PL constant *and $f^* = f(x^*)$.*

We have the following result for PL functions.

Lemma 1. *PL inequality implies that every stationary point of the function is a global minimum.*

Proof: Assume that x is an arbitrary stationary point, i.e., $||\nabla f(x)|| = 0$. Substituting x in (5) yields

$$f(x) - f^* \leq 0 \implies f(x) = f^*.$$

Hence, any stationary point corresponds to the global minimum.

Definition 3 (Proximal PL Functions)**.** *Consider the function $F(x) = f(x) + h(x)$ where $f(x) : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}^n$ is a differentiable function over as its domain and $h(x)$ is possibly nondiffrentiable. The function $F(x)$ is a proximal PL (PPL) function if it satisfies the PPL condition on a set $\mathcal{X}$ with positive constant l, for all $x \in \mathcal{X}$, i.e.,*

$$\frac{1}{2}Q(x, L_1(f)) \geq l(F(x) - F(x^*)), \quad (6)$$

where

$$Q(x, a) \stackrel{\text{def}}{=} -2a\min_{z \in \mathcal{X}}\{\frac{a}{2}||z - x||^2 + \langle \nabla f(x), z - x \rangle + h(z) - h(x)\} \quad (7)$$

with a being a positive scalar.

The Proximal PL condition is a generalisation of the PL condition (5). To see more examples, refer to [8].

To state complexity results we use the big-O notation in the sense defined below.

Definition 4 (The big O-notation)**.** *Suppose $f(x)$ and $g(x)$ are two positive scalar functions defined on some subset of the real numbers. We write $f(x) = \mathcal{O}(g(x))$, and say $f(x)$ is in the order of $g(x)$, if and only if there exist constants $\bar{K}$ and M such that $f(x) \leq Mg(x)$ for all $x > K$.*

III. Problems of Interest

In this paper we study the performance of a random zeroth-order method for optimising PL functions for both unconstrained and constrained cases. Specifically, first, we study the following unconstrained problem

$$\min_{x \in R^n} \quad f(x) \quad (8)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is in $C^{1,1}$, satisfies the PL condition (5), and is bounded below.

Later, we will focus on the following constrained problem:

$$\min_{x \in \mathcal{X}} \quad f(x), \quad (9)$$

where $\mathcal{X}$ is a compact convex set in $\mathbb{R}^n$ with diameter d_x and $f(x)$ is in $C^{1,1}$ and $F(x) = f(x) + \mathcal{I}_{\mathcal{X}}(x)$ satisfies (6) where $\mathcal{I}_{\mathcal{X}}(x)$ is the indicator function of set $\mathcal{X}$, i.e.,

$$\mathcal{I}_{\mathcal{X}}(x) = \begin{cases} 0 & x \in \mathcal{X} \\ \infty & x \notin \mathcal{X}. \end{cases}$$

The boundedness of $\mathcal{X}$ and continuity of $f(x)$ guarantees that the problem has a solution.

IV. Main Results

In this section, we establish the convergence properties and the complexity bounds of a well-known class of random zeroth-order algorithms proposed by Nesterov and Spokoiny in [5] for minimising (proximal) PL functions. We consider both unconstrained and constrained cases.

A. Unconstrained Problem

In this subsection, we study the unconstrained problem (8). The framework introduced in [5] is recalled in Algorithm 1, where x_0 is the initial guess, h_k is the step size and N is the number of iterations.

Algorithm 1 RS_μ

1: Input: x_0, μ, h_k, N
2: **for** $k = 1, \ldots, N$ **do**
3: Generate u
4: Calculate $g_\mu(x_k)$ using (2)
5: $x_{k+1} = x_k - h_k g_\mu(x_k)$
6: **end for**
7: return x_{k+1}

The following theorem characterises the convergence properties of Algorithm 1 applied to problem (8).

Theorem 1. *Let the sequence $\{x_k\}_{k \geq 0}$ be generated by Algorithm 1 (RS_μ), where $f(x) \in C^{1,1}$ and satisfies PL condition. Then, for any $N \geq 0$, with $h_k = \frac{1}{4(n+4)L_1(f)}$, we have*

$$\frac{1}{N+1}\sum_{k=0}^{N}(\Phi_k - f^*) \leq \frac{8(n+4)L_1(f)}{l}\left[\frac{f(x_0) - f^*}{N+1} + \frac{3\mu^2(n+4)}{32}L_1(f)\right] + \frac{\mu^2}{4l}L_1^2(f)(n+6)^3, \quad (10)$$

where $\Phi_k \overset{def}{=} E_{\mathcal{U}_k}[f(x_k)]$ *for all* $k \geq 1$ *and* $\Phi_0 = f(x_0)$. *Also,* $\mathcal{U}_k = \{u_0, u_1, \ldots, u_{k-1}\}$ *and* n *is the dimension of* x.

Proof: From [5, Equation (12)], we know that if $f \in C^{1,1}$, then $f_\mu \in C^{1,1}$ and $L_1(f_\mu) \leq L_1(f)$. Thus writing (4) for $f_\mu(x)$ at points x_k and x_{k+1}

$$\begin{aligned} f_\mu(x_{k+1}) \leq & f_\mu(x_k) - h_k \langle \nabla f_\mu(x_k), g_\mu(x_k) \rangle \\ & + \frac{1}{2} h_k^2 L_1(f_\mu) ||g_\mu(x_k)||^2. \end{aligned} \tag{11}$$

From [5, Equation (21)], we know for a function $f(x)$ we have

$$\nabla f_\mu(x) = \frac{1}{\kappa} \int \frac{f(x + \mu u) - f(x)}{\mu} e^{-\frac{1}{2}||u||^2} B u du.$$

From the definition of $g_\mu(x)$ and the term obtained for $\nabla f_\mu(x)$, we have $E_u[g_\mu(x)] = \nabla f_\mu(x)$ (it means $g_\mu(x)$ is an unbiased estimator of $\nabla f_\mu(x)$). Taking the expectation in u_k yields

$$\begin{aligned} E_{u_k}(f_\mu(x_{k+1})) \leq & f_\mu(x_k) - h_k ||\nabla f_\mu(x_k)||^2 \\ & + \frac{1}{2} h_k^2 L_1(f_\mu) E_{u_k}(||g_\mu(x_k)||^2). \end{aligned} \tag{12}$$

For a function $f \in C^{1,1}$ from [5, Lemma 5], we have

$$\begin{aligned} E_u[||g_\mu(x)||^2] \leq & 4(n+4)||\nabla f_\mu(x)||^2 \\ & + 3\mu^2 L_1^2(f)(n+4)^3 \end{aligned} \tag{13}$$

Substituting (13) in (12) and noting $L_1(f_\mu) \leq L_1(f)$, we obtain

$$\begin{aligned} E_{u_k}(f_\mu(x_{k+1})) \leq & f_\mu(x_k) - h_k ||\nabla f_\mu(x_k)||^2 + \\ & \frac{1}{2} h_k^2 L_1(f) \left(4(n+4)||\nabla f_\mu(x_k)||^2 \right. \\ & \left. + 3\mu^2 L_1^2(f)(n+4)^3 \right). \end{aligned} \tag{14}$$

Fixing $h_k = \hat{h} \overset{def}{=} \frac{1}{4(n+4)L_1(f)}$:

$$\begin{aligned} E_{u_k}(f_\mu(x_{k+1})) \leq & f_\mu(x_k) - \frac{1}{2} \hat{h} ||\nabla f_\mu(x_k)||^2 \\ & + \frac{3\mu^2}{32} L_1(f)(n+4). \end{aligned} \tag{15}$$

Taking the expectation of this inequality with respect to $\mathcal{U}_{k-1} = \{u_0, u_1, \ldots, u_{k-1}\}$, we obtain

$$\Phi_{k+1} \leq \Phi_k - \frac{1}{2} \hat{h} \Xi_k^2 + \frac{3\mu^2(n+4)}{32} L_1(f), \tag{16}$$

where $\Xi_k^2 = E_{\mathcal{U}_k}(||\nabla f_\mu(x_k)||^2)$. Considering $f^* \leq f(x_{N+1})$ and summing (16) from $k = 0$ to $k = N$ and divide it by $N+1$, we get

$$\begin{aligned} \frac{1}{N+1} \sum_{k=0}^{N} \Xi_k^2 \leq & 8(n+4)L_1(f)[\frac{f(x_0) - f^*}{N+1} \\ & + \frac{3\mu^2(n+4)}{32} L_1(f)], \end{aligned} \tag{17}$$

From [5, Lemma 4], it is known that for a function $f \in C^{1,1}$

$$||\nabla f(x)||^2 \leq 2||\nabla f_\mu(x)||^2 + \frac{\mu^2}{2} L_1^2(f)(n+6)^3. \tag{18}$$

From (5) and (18) one concludes

$$\begin{aligned} 2l(f(x_k) - f^*) \leq & ||\nabla f(x_k)||^2 \\ \leq & 2||\nabla f_\mu(x_k)||^2 + \frac{\mu^2}{2} L_1^2(f)(n+6)^3. \end{aligned} \tag{19}$$

Taking the expectation of this inequality with respect to $\mathcal{U}_k$

$$E_{\mathcal{U}_k}(2l(f(x_k) - f^*)) \leq \theta_k^2 \leq 2\Xi_k^2 + \frac{\mu^2}{2} L_1^2(f)(n+6)^3, \tag{20}$$

where $\theta_k^2 = E_{\mathcal{U}_k}(||\nabla f(x_k)||^2)$. Summing the inequality from $k = 0$ to $k = N$ and dividing it by $N+1$, yields

$$\begin{aligned} \frac{1}{N+1} \sum_{k=0}^{N} (\Phi_k - f^*) \leq & \frac{1}{2l(N+1)} \sum_{k=0}^{N} \theta_k^2 \\ \leq & \frac{1}{l(N+1)} \sum_{k=0}^{N} \Xi_k^2 + \frac{\mu^2}{4l} L_1^2(f)(n+6)^3. \end{aligned} \tag{21}$$

In a practical implementation, we might be interested in identifying the "best" solution guess at any step N. To this aim, define $\hat{x}_N \overset{\text{def}}{=} \arg\min_x [f(x) : x \in \{x_0, \ldots, x_N\}]$. Hence,

$$E_{\mathcal{U}_{N-1}}(f(\hat{x}_N) - f^*) \leq \frac{1}{N+1} \sum_{k=0}^{N} (\Phi_k - f^*). \tag{22}$$

Remark 2 (RS_μ Complexity, Parameter Selection, and Solution Error Bound)**.** *If* μ *is in the order of* $\mathcal{O}(\frac{\sqrt{l\epsilon}}{n^{3/2} L_1(f)})$ *and* N *is in the order of* $\mathcal{O}(\frac{n L_1(f)}{l\epsilon})$*, it is guaranteed that* $E_{\mathcal{U}_{N-1}}(f(\hat{x}_N)) - f^* \leq \epsilon$ *for some positive scalar* ϵ.

In comparison with the non-convex smooth case in [5], from the properties of PL functions the convergence is to a global minimum (see Lemma 1). Also, comparing the bounds in these two cases, for the case where $l > 1$ one can obtain the same error bound, i.e. ϵ, with fewer iterations. The number of iterations is inversely proportional with l.

B. Constrained Problem

Now we will focus on the problem (9). This problem can be reformulated as

$$\min_{x \in R^n} f(x) + \mathcal{I}_{\mathcal{X}}(x),$$

where $\mathcal{I}_{\mathcal{X}}(x)$ is the indicator function of $\mathcal{X}$. The new scheme for constrained problem is defined in Algorithm 2. In

Algorithm 2 RSc_μ

1: Input: x_0, h_k, μ, N
2: **for** $k = 1, \ldots, N$ **do**
3: Generate u
4: Calculate $g_\mu(x_k)$ using (2)
5: $\bar{x}_{k+1} = x_k - h_k g_\mu(x_k)$
6: $x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(\bar{x}_{k+1})$
7: **end for**
8: return x_{k+1}

Algorithm 2, the projection is used to ensure that the output sequence is completely in the feasible set.

Before proceeding further, we define the following auxiliary variables:

$$P_{\mathcal{X}}(x, g(x), h) \overset{\text{def}}{=} \frac{1}{h}[x - \text{Proj}_{\mathcal{X}}(x - hg(x))], \tag{23}$$

$$s_k \overset{\text{def}}{=} P_{\mathcal{X}}(x_k, g_\mu(x_k), h_k), \tag{24}$$

$$v_k \overset{\text{def}}{=} P_{\mathcal{X}}(x_k, \nabla f_\mu(x_k), h_k). \tag{25}$$

In the unconstrained case we had $x_{k+1} = x_k - h_k g_\mu(x_k)$. In the constrained case, from Algorithm 2, (23), and (24), we can see that $x_{k+1} = x_k - h_k s_k$.

Before stating the main result, in what follows, we propose a lower bound for the value of an operator that plays a crucial role in proving the main result of this section.

Lemma 2. *Consider problem (9) where $f(x) \in C^{1,1}$ is a proximal PL function in the sense of Definition 3, we have*

$$\begin{aligned} T(x_k, L_1(f)) \geq\ & 2l(f(x_k) - f(x^*)) \\ & - \mu L_1(f)^2 (n+3)^{3/2} d_x - 2L_1(f) d_x ||\xi_k||, \end{aligned} \tag{26}$$

where

$$\begin{aligned} T(x, a) \overset{def}{=} -2a \min_{z \in \mathcal{X}} \{ & \frac{a}{2}||z - x||^2 \\ & + \langle g_\mu(x), z - x \rangle + \mathcal{I}_{\mathcal{X}}(z) - \mathcal{I}_{\mathcal{X}}(x) \}, \end{aligned} \tag{27}$$

for a positive scalar a.

Proof: From the definition of PPL functions, we have

$$\begin{aligned} & 2l(f(x_k) - f(x^*)) \leq Q(x_k, L_1(f)) \\ & \leq -2L_1(f) \min_{z \in \mathcal{X}} (\frac{L_1(f)}{2}||z - x_k||^2 \\ & + \langle \nabla f(x_k), z - x_k \rangle) \\ & \leq -2L_1(f) \min_{z \in \mathcal{X}} (\frac{L_1(f)}{2}||z - x_k||^2 + \langle \nabla f_\mu(x_k) \\ & , z - x_k \rangle + \langle \nabla f(x_k) - \nabla f_\mu(x_k), z - x_k \rangle) \\ & \leq -2L_1(f) \min_{z \in \mathcal{X}} (\frac{L_1(f)}{2}||z - x_k||^2 + \langle \nabla f_\mu(x_k) \\ & , z - x_k \rangle - ||\nabla f(x_k) - \nabla f_\mu(x_k)||||z - x_k||), \end{aligned}$$

where the second inequality is a consequence of evaluating (7) for $x_k \in \mathcal{X}$ and noting that $\mathcal{I}_{\mathcal{X}}(x_k) = 0$. From [5, Lemma 3], for a function $f \in C^{1,1}$ we have

$$||\nabla f(x) - \nabla f_\mu(x)|| \leq \frac{\mu}{2} L_1(f)(n+3)^{3/2}. \tag{28}$$

From (28) and $||z - x_k|| \leq d_x$ for all $z \in \mathcal{X}$, we have

$$\begin{aligned} & 2l(f(x_k) - f(x^*)) \leq Q(x_k, L_1(f)) \\ & \leq -2L_1(f) \min_{z \in \mathcal{X}} (\frac{L_1(f)}{2}||z - x_k||^2 \\ & + \langle \nabla f_\mu(x_k), z - x_k \rangle - \frac{\mu L_1(f)(n+3)^{3/2} d_x}{2}) \\ & \leq -2L_1(f) \min_{z \in \mathcal{X}} (\frac{L_1(f)}{2}||z - x_k||^2 + \langle g_\mu(x_k) \\ & , z - x_k \rangle - ||\xi_k|| d_x - \frac{\mu L_1(f)(n+3)^{3/2} d_x}{2}). \end{aligned}$$

where $\xi_k \overset{\text{def}}{=} g_\mu(x_k) - \nabla f_\mu(x_k)$. Rearranging above terms completes the proof.

Before stating the main result regarding the performance of Algorithm 2, we present the following lemma on the variance of the random oracle $g_\mu(x)$.

Lemma 3. *Random oracle $g_\mu(x)$ is a variance bounded unbiased estimator of $\nabla f_\mu(x)$. That is*

$$E_u[||g_\mu(x_k) - \nabla f_\mu(x_k)||^2] \leq \sigma_k^2, \ \ \sigma_k \geq 0.$$

Proof: We know that $E_u[g_\mu(x_k)] = \nabla f_\mu(x_k)$, so we have $E_u[||g_\mu(x_k) - \nabla f_\mu(x_k)||^2] \leq E_u[||g_\mu(x_k)||^2] \leq \sigma_k^2$.

Remark 3. *An upper bound for $E_u[||g_\mu(x_k)||^2]$ can be obtained. For example, from [5, Theorem 4] we know for a function $f \in C^{0,0}$ we have $E_u[||g_\mu(x)||^2] \leq L_0(f)^2(n+4)^2$ and for a function $f \in C^{1,1}$ we have $E_u[||g_\mu(x)||^2] \leq \frac{\mu^2}{2} L_1^2(f)(n+6)^3 + 2(n+4)||\nabla f(x)||^2$. These upper bounds can be used as candidates for σ_k^2.*

Theorem 2. *Consider problem (9). Let the sequence $\{x_k\}_{k \geq 0}$ be generated by RSc$_\mu$, when $f(x) \in C^{1,1}$ satisfies the proximal PL condition in the sense of Definition 3. Then, for any $N \geq 0$, with $h_k = \frac{1}{L_1(f)}$, we have*

$$\begin{aligned} \frac{1}{N+1} \sum_{k=0}^{N} \Phi_k - f(x^*) \leq\ & \frac{L_1(f)}{l} \frac{f(x_0) - f(x^*)}{N+1} \\ & + \frac{\mu d_x L_1(f)^2 (n+3)^{3/2}}{2l} + \frac{L_1(f) d_x}{l(N+1)} \sum_{k=0}^{N} \sigma_k \\ & + \frac{1}{l(N+1)} \sum_{k=0}^{N} \sigma_k^2, \end{aligned} \tag{29}$$

where $\Phi_k \overset{\text{def}}{=} E_{\mathcal{U}_k}[f(x_k)]$ for all $k \geq 1$, $\Phi_0 = f(x_0)$, σ_k is given in Lemma 3, $\mathcal{U}_k = \{u_0, u_1, \ldots, u_{k-1}\}$, n is the dimension of x, and d_x is the diameter of the feasible set.

Proof: From [5, Equation (12)] we know $f_\mu(x) \in C^{1,1}$. Writing (4) for $f_\mu(x)$ at points x_k and x_{k+1}, yields

$$\begin{aligned} & f_\mu(x_{k+1}) \leq f_\mu(x_k) + \frac{L_1(f_\mu)}{2}||x_{k+1} - x_k||^2 \\ & + \langle \nabla f_\mu(x_k), x_{k+1} - x_k \rangle \\ & \leq f_\mu(x_k) + \frac{L_1(f_\mu)}{2}||x_{k+1} - x_k||^2 \\ & + \langle g_\mu(x_k), x_{k+1} - x_k \rangle - \langle \xi_k, x_{k+1} - x_k \rangle \\ & \leq f_\mu(x_k) - \frac{1}{2L_1(f)} T(x_k, L_1(f)) + h_k \langle \xi_k, s_k \rangle \\ & \leq f_\mu(x_k) - \frac{1}{2L_1(f)} T(x_k, L_1(f)) \\ & + h_k \langle \xi_k, s_k - v_k \rangle + h_k \langle \xi_k, v_k \rangle, \end{aligned}$$

where the third inequality follows from (27), $x_k \in \mathcal{X}$, and the fact that

$$\begin{aligned} T(x_k, L_1(f)) = -2L_1(f) \min_{z \in \mathcal{X}} \{ & \frac{L_1(f)}{2}||z - x_k||^2 \\ & + \langle g_\mu(x_k), z - x_k \rangle \} \end{aligned}$$

$$= -2L_1(f)\Big(\frac{L_1(f)}{2}||x_{k+1}-x_k||^2 + \langle g_\mu(x_k), x_{k+1}-x_k\rangle\Big).$$

The last equality above follows from the definition of the projection operator. Taking expected value with respect to u_k and considering Lemmas 3 and 4 in the appendix, and $h_k = \frac{1}{L_1(f)}$, leads to

$$E_{u_k}[f_\mu(x_{k+1})] \leq f_\mu(x_k) - \frac{1}{2L_1(f)}E_{u_k}[T(x_k, L_1(f))] + \frac{\sigma_k^2}{L_1(f)}.$$

Taking the expectation with respect to $\mathcal{U}_{k-1}$, we have

$$E_{\mathcal{U}_k}[T(x_k, L_1(f))] \leq 2L_1(f)(\Phi_{k+1} - \Phi_k) + 2\sigma_k^2.$$

Summing over $k = 0$ to $k = N$ and dividing it by $N+1$, results in

$$\frac{1}{N+1}\sum_{k=0}^{N} E_{\mathcal{U}_k}[T(x_k, L_1(f))] \leq \frac{2}{N+1}\sum_{k=0}^{N}\sigma_k^2 + 2L_1(f)\frac{f(x_0)-f(x^*)}{N+1}.$$

Also, taking the expectation of (26) with respect to u_k and then $\mathcal{U}_{k-1}$, summing over $k = 0$ to $k = N$, dividing it by $N+1$ and using Lemma 5 from the appendix, yield

$$\frac{2l}{N+1}\sum_{k=0}^{N}\Phi_k - f(x^*) \leq \mu d_x L_1(f)^2(n+3)^{3/2} + \frac{1}{N+1}\sum_{k=0}^{N} E_{\mathcal{U}_k}[T(x_k, L_1(f))] + \frac{2L_1(f)d_x}{N+1}\sum_{k=0}^{N}\sigma_k.$$

Thus, we have

$$\frac{1}{N+1}\sum_{k=0}^{N}\Phi_k - f(x^*) \leq \frac{L_1(f)}{l}\frac{f(x_0)-f(x^*)}{N+1} + \frac{\mu d_x L_1(f)^2(n+3)^{3/2}}{2l} + \frac{L_1(f)d_x}{l(N+1)}\sum_{k=0}^{N}\sigma_k + \frac{1}{l(N+1)}\sum_{k=0}^{N}\sigma_k^2.$$

In a practical implementation of the algorithm we might be interested in keeping track of the best guess for the optimum solution at any given step N. As in the unconstrained case, let this best guess be denoted by $\hat{x}_N = \arg\min_x[f(x) : x \in \{x_0, \ldots, x_N\}]$. Thus,

$$E_{\mathcal{U}_{N-1}}(f(\hat{x}_N) - f^*) \leq \frac{1}{N+1}\sum_{k=0}^{N}(\Phi_k - f^*).$$

Remark 4 (RSc_μ Complexity, Parameter Selection, and Solution Error Bound)**.** *If $\mu \leq \frac{l\epsilon}{d_x L_1(f)^2(n+3)^{3/2}}$, N is in the order of $\mathcal{O}\left(\frac{L_1(f)}{l\epsilon}\right)$, and denoting $\sigma = \max_k[\sigma_k : k \in \{0, \ldots, N\}]$, then*

$$E_{\mathcal{U}_{N-1}}(f(\hat{x}_N) - f^*) \leq \epsilon + \frac{L_1(f)d_x\sigma}{l} + \frac{\sigma^2}{l}, \qquad (30)$$

for some positive scalar ϵ. Thus, we can guarantee that there exists an integer $\bar{N}$ such that for all $N \geq \bar{N}$, $E_{\mathcal{U}_{N-1}}f(x_k)$ is in a neighbourhood of f^* for an appropriate choice of μ.

Similar phenomena are observed consistently in the literature on constrained non-convex problems, where this effect has been reported for different algorithms, see e.g. [4], [10]. Note that similarly to the unconstrained case l appears in the denominator of the iteration number order term. Additionally, the two terms on the right-hand side of (30) are inversely proportional with l. Also, it can be seen in this case due to the choice of h_k, the number of iterations is not dependent on the dimension of x, but we know that $h_k \in (0, \frac{1}{L_1(f)}]$ and in fact N is in the order of $\mathcal{O}(\frac{1}{lh_k\epsilon})$ for the more general case.

V. Numerical Examples

In this section, we consider two scenarios. In both scenarios we study the performance of Algorithms 1 and 2 applied to unconstrained and constrained least square problems. Specifically, the objective function is assumed to be $||Ax-b||^2$, where $A \in R^{m\times n}$ $(n \geq m)$, $x \in R^n$, and $b \in R^m$. The function is not strongly convex but it satisfies the PL condition with $l = 2||A^T||^2$ and is in $C^{1,1}$ with $L_1(f) = 2||A^TA||$.

Scenario 1: In the first scenario, we consider an unconstrained least-squares problem of the form introduced above with $m = 100$ and $n = 1000$. In light of Remark 2, we choose $\epsilon = 0.01$ and consequently we set $\mu = 10^{-7}$ and $N = 200000$. Matrix A rows are sampled from $\mathcal{N}(0, I_m)$ and $b = A\bar{x} + \omega$, where $\bar{x}$ sampled from $\mathcal{N}(0,1)$ and ω from $\mathcal{N}(0, 0.01)$. Moreover, the initial condition vector is sampled from $\mathcal{N}(0,1)$. We explore the performance of the algorithm for two step sizes $\frac{1}{4(n+4)L_1(f)} \approx 10^{-7}$ and 10^{-6}. We repeat the example for 25 times. The empirical mean of best guess for the optimum solution $(f(\hat{x}_N))$ over 25 runs and upper bound calculated in (22) is presented in Fig. 1. As it was shown in Theorem 1, by increasing the step size the convergence to a neighbourhood of the solution would be faster, but this comes at the expense of larger error bounds.

Scenario 2: In this scenario we consider a constrained least squares problem with the same problem parameter choices as the previous scenario except for μ. From Remark 4, given the same value for ϵ as the previous case, we choose $\mu = 10^{-10}$ and $N = 200000$. The constraint set $\mathcal{X}$ is assumed to be $\mathcal{X} = \{x \in \mathbb{R}^n | x_i \in [-0.5, 0.5], \forall i \in \{1, \ldots, n\}\}$. We explore the performance of the algorithm for two step sizes $\frac{1}{nL_1(f)} \approx 10^{-7}$ and 10^{-6} (both are less than $\frac{1}{L_1(f)}$). The empirical mean of objective function value in iterations generated by Algorithm 2 over 25 runs are depicted in Fig. 2. The effect of additional error terms in the constrained case can be seen by comparing the figures at point 10^4 iterations for $h_k = 10^{-6}$ cases and at point 10^5 iterations for $h_k = 10^{-7}$cases.

VI. Conclusion and Future Research Directions

The application of a zeroth-order scheme using random oracles for minimising PL functions with or without

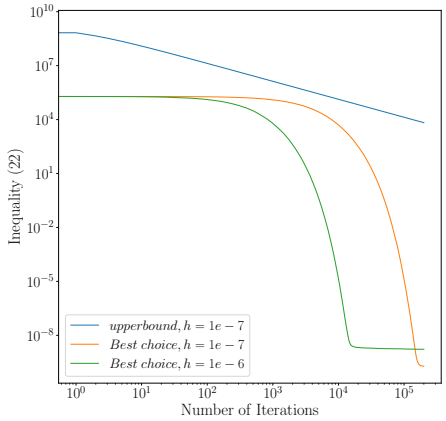

Fig. 1. The evolution of the empirical mean of $f(\hat{x}_k)$ and the calculated upper bound versus the number of iterations in Scenario 1.

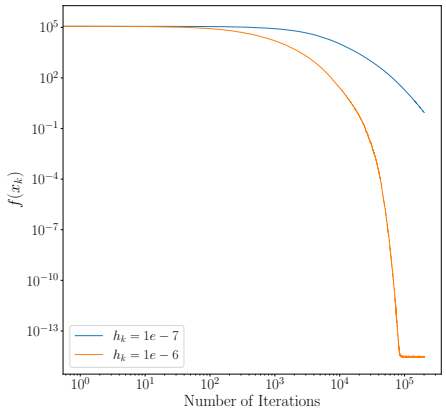

Fig. 2. The evolution of $f(x_k)$ versus the number of iterations. Note that in this case $f(x^*) = 0$ in Scenario 2.

constraints was discussed. For the unconstrained problem, the convergence properties of the proposed algorithm were studied. Additionally, the complexity bounds for the algorithm along with guidelines for selecting algorithm parameters were introduced. Next, a generalisation of the PL inequality was exploited to establish the convergence of the algorithm for solving constrained problems. Similar to the unconstrained case, complexity bounds were derived. Numerical examples were presented to illustrate the theoretical results. An immediate future step is extending the analysis of the constrained case to the case where the constraint set is unbounded. Another possible future direction is applying similar techniques for solving minimax problems using zeroth-order oracles of the type studied in this paper.

APPENDIX

A. Auxiliary Lemmas

Lemma 4. *Let* $\xi_k \overset{\text{def}}{=} g_\mu(x_k) - \nabla f_\mu(x_k)$. *From [4, Proposition 1], it implies that* $\langle \xi_k, s_k - v_k \rangle \leq ||\xi_k||^2$.

Lemma 5. *For a function* $f(x)$ *with smoothed version* $f_\mu(x)$ *and random oracle* $g_\mu(x)$, *we have*

$$E_{u_k}[||\xi_k||] \leq \sigma_k, \quad \sigma_k \geq 0,$$

where $\xi_k = g_\mu(x_k) - \nabla f_\mu(x_k)$.

Proof: Due to Lemma 3, $E_{u_k}[||\xi_k||^2] \leq \sigma_k^2$. From Jensen's inequality, we have

$$E_{u_k}[||\xi_k||]^2 \leq E_{u_k}[||\xi_k||^2].$$

Thus, $E_{u_k}[||\xi_k||] \leq \sigma_k$.

REFERENCES

[1] Alejandro I Maass et al. "Zeroth-Order Optimization on Subsets of Symmetric Matrices With Application to MPC Tuning". In: *IEEE Transactions on Control Systems Technology* 30.4 (2021), pp. 1654–1667.

[2] Alonso Marco et al. "Automatic LQR tuning based on Gaussian process global optimization". In: *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2016, pp. 270–277.

[3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *SIAM review* 60.2 (2018), pp. 223–311.

[4] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization". In: *Mathematical Programming* 155.1-2 (2016), pp. 267–305.

[5] Yurii Nesterov and Vladimir Spokoiny. "Random Gradient-Free Minimization of Convex Functions". In: *Foundations of Computational Mathematics* 17.2 (Apr. 2017), pp. 527–566. ISSN: 1615-3375, 1615-3383. DOI: 10.1007/s10208-015-9296-2. URL: http://link.springer.com/10.1007/s10208-015-9296-2 (visited on 06/05/2023).

[6] Boris T Polyak. "Gradient methods for solving equations and inequalities". In: *USSR Computational Mathematics and Mathematical Physics* 4.6 (1964), pp. 17–32.

[7] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks". In: *Applied and Computational Harmonic Analysis* 59 (2022), pp. 85–116.

[8] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer. 2016, pp. 795–811.

[9] Sotirios-Konstantinos Anagnostidis, Aurelien Lucchi, and Youssef Diouane. "Direct-search for a class of stochastic min-max problems". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3772–3780.

[10] Sijia Liu et al. "Zeroth-order stochastic variance reduction for nonconvex optimization". In: *Advances in Neural Information Processing Systems* 31 (2018).