Noname manuscript No.
(will be inserted by the editor)

A Zeroth Order Method for Stochastic Weakly Convex Optimization

V. Kungurtsev · F. Rinaldi

Received: date / Accepted: date

Abstract In this paper, we consider stochastic weakly convex optimization problems, however without the existence of a stochastic subgradient oracle. We present a derivative free algorithm that uses a two point approximation for computing a gradient estimate of the smoothed function. We prove convergence at a similar rate as state of the art methods, however with a larger constant, and report some numerical results showing the effectiveness of the approach.

Keywords Derivative Free Optimization · Zeroth Order Optimization · Stochastic Optimization · Weakly Convex Functions

Mathematics Subject Classification (2010) 90C56 · 90C15 · 65K05

1 Introduction

In this paper, we study the following class of problems:

$$\min_{x\in\mathbb{R}^n} \phi(x) := f(x) + r(x), \tag{1}$$

with $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ a stochastic, weakly convex, and potentially nonsmooth (i.e., not necessarily continuously differentiable) function, and $r(\cdot) : \mathbb{R}^n \to \bar{\mathbb{R}}$ (i.e., it is extended real valued) is convex but not necessarily even continuous,

V. Kungurtsev
Department of Computer Science, Faculty of Electrical Engineering,
Czech Technical University in Prague Prague, Czech Republic
E-mail: vyacheslav.kungurtsev@fel.cvut.cz
Research supported by the OP VVV project CZ.02.1.01/0.0/0.0/16 019/0000765 "Research Center for Informatics"

F. Rinaldi
Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova
Via Trieste, 63, 35121 Padua, Italy
Tel.: +39-049-8271424
E-mail: rinaldi@math.unipd.it

however $r(x)$ satisfies some additional conditions detailed below. Furthermore, we consider the *derivative free* or *zeroth order* context, wherein the subgradients ∂f, or unbiased estimates thereof, are not available, but only unbiased estimates of function evaluations $f(x)$ are available. We thus write

$$f(x) = \mathbb{E}_\xi[F(x;\xi)] = \int_\Xi F(x,\xi)dP(\xi),$$

with $\{F(\cdot,\xi),\ \xi \in \Xi\}$ a collection of real valued functions and P a probability distribution over the set Ξ to be precise.

We define two quantitative assumptions regarding $f(\cdot)$ and $r(\cdot)$ below. First, we define the notion of a proximal map, in particular with any constant α and any convex function h we can write $\text{prox}_{\alpha h}$ to indicate the following function:

$$\text{prox}_{\alpha h}(x) = \text{argmin}_y\{h(y) + \frac{1}{2\alpha}\|y-x\|^2\}.$$

The associated optimality condition is

$$y = \text{prox}_{\alpha h}(x) \Longleftrightarrow x - y \in \alpha\partial h(y)$$

We shall make use of the *nonexpansiveness* property of the proximal mapping in the sequel,

$$\|\text{prox}_{\alpha h}(x) - \text{prox}_{\alpha h}(y)\| \leq \|x-y\|.$$

We now state our standing assumption on the properties of (1):

Assumption 1 *1. $f(\cdot)$ is ρ-weakly convex, i.e., $f(x) + \rho\|x\|^2$ is convex for some $\rho > 0$, directionally differentiable, bounded below by $f_\star$ and locally Lipschitz with constant L_0.*
2. $r(\cdot)$ is convex (but not necessarily continuously differentiable). Furthermore, $r(x)$ is bounded below by $r_\star$.
We shall denote the lower bound of ϕ by $\phi_\star = f_\star + r_\star$.

We further assume that the proximal map of $r(x)$ can be evaluated at low computational complexity cost. We note that the ρ-weak convexity property for a given function f is equivalent to hypomonotonicity of its subdifferential map, that is

$$\langle v - w, x - y\rangle \geq -\rho\|x-y\|^2 \tag{2}$$

for $v \in \partial f(x)$ and $w \in \partial f(y)$ (see, e.g., [19, Example 12.28, page 549]).

The class of weakly convex functions is a special yet very common case of nonconvex functions, which contains all convex (possibly nonsmooth) functions and Lipschitz smooth functions. One standard subset of weakly convex functions is given by the composite function $f(x) = h(c(x))$ where h is nonsmooth and convex and $c(x)$ is continuously differentiable but non-convex (see, e.g., [9] and references therein). The additive composite class is another widely used class of weakly convex functions [10], formed from all sums $g(x) + l(x)$ with l closed and convex and g continuously differentiable.

One method for solving a weakly convex stochastic optimization problem is given as repeated iterations of,

$$x_{k+1} \in \text{argmin}_y \left\{ f_{x_k}(y; S_k) + r(y) + \frac{1}{2\alpha_k}\|y - x_k\|^2 \right\} \tag{3}$$

where $\alpha_k > 0$ is a stepsize sequence, typically taken to satisfy $\alpha_k \to 0$, and $f_{x_k}(y; S_k)$ is approximating f at x_k using a noisy estimate S_k of the data. A basic stochastic subgradient method will use the linear model

$$f_{x_k}(y; S_k) = f(x_k) + \zeta^T(y - x_k)$$

where $\zeta \approx \bar{\zeta} \in \partial f(x_k)$. When using this approach, it is common to consider the existence of some oracle of an unbiased estimate of an element of the subgradient that enables one to build up the approximation f_{x_k} with favorable properties (see,e.g., [9] or [12]). In our case we assume such an oracle is not available, and we only get access, at a point x, to a noisy function value observation $F(x, \xi)$. Stochastic problems with only functional information available often arise in optimization, machine learning and statistics. A classic example is simulation based optimization (see,e.g., [2,15] and references therein), where function evaluations usually represent the experimentally obtained behavior of a system and in practice are given by means of specific simulation tools, hence no internal or analytical knowledge for the functions is provided. Furthermore, evaluating the function at a given point is in many cases a computationally expensive task, and only a limited budget of evaluations is available in the end. Recently, suitable derivative free/zeroth order optimization methods have been proposed for handling stochastic functions (see,e.g., [6, 7, 11, 14]). For a complete overview of stochastic derivative free/zeroth order methods, we refer the interested reader to the recent review [15].

Weakly convex functions show up in the modeling of many different statistical learning applications like, e.g., (robust) phase retrieval, sparse dictionary learning, conditional value at risk (see [9] for a complete description of those problems). Other interesting applications include the training of neural networks with Exponentiated Linear Units (ELUs) activation functions [8] and machine learning problems with L-smooth loss functions (see, e.g., [13] and references therein).

In all these problems there might be cases where we only get access, at a point x, to an unbiased estimate of the loss function $F(x, \xi)$ and we thus need to resort to a stochastic derivative free/zeroth order approach in order to handle our problem. Recalling that a standard setting is wherein a function evaluation is the noisy output of some complex simulation, such a problem can appear either for an inverse problem where we are interested in using a robust nonsmooth loss function to match parameters to a nonconvex simulation, i.e., $F(x, \xi) = \sum_i \|G(x, \xi_i) - o_i\|_1$ where $\{o_i\}$ is a the set of observations and $\{\xi_i\}$ a set of samples of the simulation run, which is of the form of the composite case $h(c(x))$ described above, or even a simulation function that is convex but we are interested in, e.g., minimizing its conditional value at risk.

At the time of writing, zeroth order, or derivative free optimization for weakly convex problems has not been investigated. There are a number of works for stochastic nonconvex zeroth order optimization (e.g., [5]) and nonsmooth convex derivative free optimization (e.g., [11]).

In the case of stochastic weakly convex optimization but with access to a noisy element of the subgradient, there are a few works that have appeared fairly recently. Asymptotic convergence was shown in [12], which proves convergence with probability one for the method given in (3). Non-asymptotic convergence, as in convergence rates in expectation, is given in the two papers [9] and [16].

In this paper, we follow the approach proposed in [11] to handle nonsmoothness in our problem. We consider a smoothed version of the objective function, and we then apply a two point strategy to estimate its gradient. This tool is thus embedded in a proximal algorithm similar to the one described in [9] and enables us to get convergence at a similar rate as the original method (although with larger constants).

The rest of the paper is organized as follows. In Section 2 we describe the algorithm and provide some preliminary lemmas needed for the subsequent analysis. Section 3 contains the convergence proof. In Section 4 we show some numerical results on two standard test cases. Finally we conclude in Section 5.

2 Two Point Estimate and Algorithmic Scheme

We use the two point estimate presented in [11] to generate an approximation to an element of the subdifferential. In particular, consider the randomized smoothing of the function f,

$$f_u(x) = \mathbb{E}[f(x+uz)] = \int f(x+uz)dZ$$

where Z is the pdf of a standard normal variable, i.e., we take an expectation for $z \sim \mathcal{N}(0, I_n)$.

The two point estimate we use is given by considering a second smoothing, now of $f_{u_{1,t}}$ for a given $u_{1,t}$ indexed by iteration t, i.e.,

$$f_{u_{1,t},u_{2,t}}(x) = \mathbb{E}[f_{u_{1,t}}(x+u_{2,t}z)] = \int f_{u_{1,t}}(x+u_{2,t}z)dZ.$$

To derive the specific step computed, let us consider the derivative of this function with respect to x. We first write,

$$\begin{aligned} f_{u_{1,t}u_{2,t}}(x) &= \textstyle\int f_{u_{1,t}}(x+u_{2,t}z)dZ = \frac{1}{\kappa}\int f_{u_{1,t}}(x+u_{2,t}v)e^{-\frac{\|v\|^2}{2}}dv \\ &= \textstyle\frac{1}{\kappa u_{2,t}^n}\int f_{u_{1,t}}(y)e^{-\frac{\|y-x\|^2}{2u_{2,t}^2}}dy, \end{aligned}$$

where

$$\kappa := \int e^{-\frac{\|v\|^2}{2}}dv = (2\pi)^{n/2}$$

and we used the change of variables $y = x + u_{2,t}v$. Now we write,

$$\begin{aligned}\nabla f_{u_{1,t},u_{2,t}}(x) &= \tfrac{1}{\kappa u_{2,t}^{n+2}} \int f_{u_{1,t}}(y) e^{-\frac{\|y-x\|^2}{2u_{2,t}^2}} (y-x)dy \\ &= \tfrac{1}{\kappa u_{2,t}} \int f_{u_{1,t}}(x+u_{2,t}v) e^{-\frac{\|v\|^2}{2}} v dv \\ &= \tfrac{1}{\kappa} \int \tfrac{f_{u_{1,t}}(x+u_{2,t}v)-f(x)}{u_{2,t}} e^{-\frac{\|v\|^2}{2}} v dv \\ &= \int \tfrac{f_{u_{1,t}}(x+u_{2,t}z)-f(x)}{u_{2,t}} z dZ.\end{aligned} \tag{4}$$

where the third equality comes from the fact that the function $ve^{-\frac{\|v\|^2}{2}}$ is even so integration over $f(x)$ is zero.

Now let $\{u_{1,t}\}_{t=1}^{\infty}$, $\{u_{2,t}\}_{t=1}^{\infty}$ be two nonincreasing sequences of positive parameters such that $u_{2,t} \leq u_{1,t}/2$, x_t is the given point, ξ_t is a sample of the stochastic oracle Ξ, $Z_1 \sim \mu_1$ and $Z_2 \sim \mu_2$ are two vectors independently sampled from distributions $\mu_1 \sim \mathcal{N}(0, I_n)$ and $\mu_2 \sim \mathcal{N}(0, I_n)$. From the derivation above, we can see that the quantity,

$$\begin{aligned}g_t(x) &= G(x, u_{1,t}, u_{2,t}, Z_{1,t}, Z_{2,t}, \xi_t) = \\ &= \frac{F(x + u_{1,t}Z_{1,t} + u_{2,t}Z_{2,t}; \xi_t) - F(x + u_{1,t}Z_{1,t}; \xi_t)}{u_{2,t}} Z_{2,t},\end{aligned} \tag{5}$$

is an unbiased estimator of $\nabla f_{u_{1,t},u_{2,t}}(x)$. Thus, effectively, the first random variable $u_{1,t}Z_{1,t}$ smooths out the nonsmooth function F and the second $u_{2,t}Z_{2,t}$ obtains a zeroth order estimate, using noisy function computations, of its derivative. We shall use $g_t(x)$ specifically in our algorithm at each iteration. We highlight the importance of using an adequate random number generator to compute $Z_{1,t}$, $Z_{2,t}$ and stochastic function realization ξ_t at every iteration. We hence have that the two samples used for ξ_t and $Z_{1,t}$ are the same in $F(x + u_{1,t}Z_{1,t} + u_{2,t}Z_{t,2}; \xi_t)$ and $F(x + u_{1,t}Z_{1,t}; \xi_t)$, making the two point estimator essentially a common random number device.

We now report some results that provide theoretical guarantees on the error in the estimate. These results appear in [18], however we include some of their (short) proofs for completeness.

Lemma 1 *[18, Lemma 1] It holds that,*

$$\frac{1}{\kappa} \int \|v\|^p e^{-\frac{\|v\|^2}{2}} dv \leq n^{p/2}, \tag{6}$$

with $p \in [0, 2]$, and

$$n^{p/2} \leq \frac{1}{\kappa} \int \|v\|^p e^{-\frac{\|v\|^2}{2}} dv \leq (p+n)^{p/2}, \tag{7}$$

with $p \geq 2$.

Lemma 2 *[18, Theorem 1] It holds that,*

$$\left| f_{u_{1,t}}(x) - f(x) \right| \leq u_{1,t} L_0 \sqrt{n}, \tag{8}$$

with L_0 Lipschitz constant for f.

Proof Indeed,

$$\begin{aligned}\left|f_{u_{1,t}}(x) - f(x)\right| &\leq \tfrac{1}{\kappa}\int |f(x+u_{1,t}v) - f(x)|\, e^{-\frac{\|v\|^2}{2}} dv \leq \tfrac{u_{1,t}L_0}{\kappa}\int \|v\|\, e^{-\frac{\|v\|^2}{2}} dv \\ &\leq u_{1,t}L_0\sqrt{n},\end{aligned}$$

where we have used the Lipschitz constant L_0 for f as given in Assumption 1 and the last inequality follows from equation (6) in Lemma 1.

Lemma 3 *[18, Lemma 2] The function $f_{u_{1,t}}$ is Lipschitz continuously differentiable with constant $\frac{L_0\sqrt{n}}{u_{1,t}}$.*

Proof

$$\begin{aligned}\left\|\nabla f_{u_{1,t}}(x) - \nabla f_{u_{1,t}}(y)\right\| &\leq \tfrac{1}{u_{1,t}\kappa}\int |f(x+u_{1,t}v) - f(y+u_{1,t}v)|\, e^{-\frac{\|v\|^2}{2}}\|v\| dv \\ &\leq \tfrac{L_0}{u_{1,t}\kappa}\|x-y\|\int e^{-\frac{\|v\|^2}{2}}\|v\| dv \leq \tfrac{L_0\sqrt{n}}{u_{1,t}}\|x-y\|.\end{aligned}$$

The condition proved in Lemma 3 is equivalent to the following inequality (see, e.g., [18]):

$$\left|f_{u_{1,t}}(y) - f_{u_{1,t}}(x) - \langle\nabla f_{u_{1,t}}(x), (y-x)\rangle\right| \leq \frac{L_0\sqrt{n}}{u_{1,t}}\|x-y\|^2. \tag{9}$$

Lemma 4 *[18, Lemma 3] It holds that*

$$\left\|\nabla f_{u_{1,t},u_{2,t}}(x) - \nabla f_{u_{1,t}}(x)\right\| \leq \frac{u_{2,t}L_0\sqrt{n}(n+3)^{3/2}}{2u_{1,t}} \leq \frac{u_{2,t}}{u_{1,t}}\bar{\sigma}, \tag{10}$$

with $\bar{\sigma} = \frac{L_0\sqrt{n}(n+3)^{3/2}}{2}$.

Proof First, note that

$$\nabla f_{u_{1,t}}(x) = \frac{1}{\kappa}\int \langle\nabla f_{u_{1,t}}(x), v\rangle e^{-\frac{\|v\|^2}{2}} v dv.$$

And so,

$$\begin{aligned}&\left\|\nabla f_{u_{1,t},u_{2,t}}(x) - \nabla f_{u_{1,t}}(x)\right\| \\ &\quad= \left\|\tfrac{1}{\kappa}\int\left(\tfrac{f_{u_{1,t}}(x+u_{2,t}v) - f_{u_{1,t}}(x)}{u_{2,t}} - \langle\nabla f_{u_{1,t}}(x), v\rangle\right) v e^{-\frac{\|v\|^2}{2}} dv\right\| \\ &\quad\leq \tfrac{1}{\kappa u_{2,t}}\int\left|f_{u_{1,t}}(x+u_{2,t}v) - f_{u_{1,t}}(x) - u_{2,t}\langle\nabla f_{u_{1,t}}(x), v\rangle\right|\|v\| e^{-\frac{\|v\|^2}{2}} dv \\ &\quad\leq \tfrac{u_{2,t}L_0\sqrt{n}}{2\kappa u_{1,t}}\int\|v\|^3 e^{\frac{-\|v\|^2}{2}} dv \leq \tfrac{u_{2,t}L_0\sqrt{n}(n+3)^{3/2}}{2u_{1,t}},\end{aligned}$$

where the first inequality uses some basic property of the integrals, the second inequality uses equation (9) coming from Lemma 3, and the last inequality uses equation (7) in Lemma 1.

We further report one more useful preliminary result.

Lemma 5 *The following inequality holds:*

$$\langle \nabla f_u(x) - \nabla f_u(y), x - y \rangle \geq -\rho \|x - y\|^2 - 4L_0 u \|x - y\|.$$

Proof By using the definition of $f_{u_{1,t}}(x)$, we have

$$\langle \nabla f_u(x) - \nabla f_u(y), x - y \rangle = \left\langle \nabla \left(\int \left(f(x + uz) - f(y + uz) \right) dZ \right), x - y \right\rangle$$

After a proper rewriting, we use (2) to get a lower bound on the considered term, for any given vector e_x of n components and any one element equal to one, we have,

$$\begin{aligned}
&\left\langle \left(\lim_{t \to 0} \tfrac{\int (f(x+uz+te_x) - f(x+uz) - f(y+uz+te_x) + f(y+uz))}{t} dZ \right), x - y \right\rangle \\
&\geq -\rho \|x - y\|^2 + \\
&+ \left\langle \left(\lim_{t \to 0} \tfrac{\int (f(x+uz+te_x) - f(x+te_x) - f(x+uz) + f(x) - f(y+uz+te_x) + f(y+te_x) + f(y+uz) - f(y))}{t} dZ \right), x - y \right\rangle \\
&\quad \geq -\rho \|x - y\|^2 - 4L_0 u \|x - y\|,
\end{aligned}$$

where the last inequality is obtained from the Lipschitz property of f (Assumption 1).

We make the following Assumption on f:

Assumption 2 *It holds that $F(\cdot, \xi)$ is $L(\xi)$-Lipschitz and $L(P) := \sqrt{\mathbb{E}[L(\xi)^2]}$ is finite.*

The following lemma uses previous results to characterizes an important condition on the error of the estimate.

Lemma 6 *Given a point x s.t. $\|x\| \leq M$, with M a finite positive value, then it holds that*

$$\mathbb{E}[\|g_t(x)\|^2] \leq \hat{C}. \tag{11}$$

where $\hat{C}$ depends on M, $L(P)$ and n but is independent of x.

Proof Define $\hat{f}(x) = f(x) + \rho \|x\|^2$ for $\|x\| \leq M$ and a continuous linearly growing extension otherwise (e.g., for any x take the greatest norm subgradient $g(x)$ at $\frac{Mx}{\|x\|}$ and linearize, $\hat{f}(x) = \hat{f}\left(\frac{Mx}{\|x\|}\right) + g(x)^T (x - \frac{Mx}{\|x\|})$). Note that by this construction and the assumptions on $f(x)$, it holds that $\hat{f}(x)$ is convex and Lipschitz. Let $\hat{g}_t(x)$ be the two point gradient approximation of $\hat{f}(x)$, defining $\hat{f}_{u_{1,t}}(x)$ accordingly. Furthermore, let $h(x) = \hat{f}(x) - f(x)$, $\hat{g}_t^h(x)$ its two point gradient approximation, and $h_{u_{1,t}}(x)$ its smoothed function. We have,

$$\|g_t(x)\| = \|\hat{g}_t(x) - \hat{g}_t^h(x)\| \leq \|\hat{g}_t(x)\| + \|\hat{g}_t^h(x)\|.$$

Since $\hat{f}_{u_{1,t}}$ and $h_{u_{1,t}}$ are both Lipschitz and convex, we now directly apply [11, Lemma 2] to both errors on the right hand side to obtain the final result.

Note that the last lemma combined with the previous results implies a tighter bound on $\|\nabla f_{u_{1,t}}(x)\|^2$, specifically,

$$\begin{aligned}
&\|\nabla f_{u_{1,t}}(x)\|^2 \\
&\leq 3\|\nabla f_{u_{1,t}}(x) - \nabla f_{u_{1,t},u_{2,t}}(x)\|^2 + 3\mathbb{E}\|g_t(x) - \nabla f_{u_{1,t},u_{2,t}}(x)\|^2 + 3\mathbb{E}\|g_t(x)\|^2 \\
&\leq 3u_{2,t}^2\bar{\sigma}^2/u_{1,t}^2 - 6\mathbb{E}\left\langle g_t(x), \nabla f_{u_{1,t},u_{2,t}}(x)\right\rangle + 3\|\nabla f_{u_{1,t},u_{2,t}}(x)\|^2 + 6\mathbb{E}\|g_t(x)\|^2 \\
&\leq 3u_{2,t}^2\bar{\sigma}^2/u_{1,t}^2 + 6\hat{C} \leq \bar{C}.
\end{aligned} \tag{12}$$

In order to get the first inequality, we used some basic properties of the expectation and the inequality $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$. Then we used Lemma 4 to upper bound the first term in the summation and suitably rewrote the second one thus getting the RHS of the second inequality. The third one was finally obtained by taking into account unbiasedness of $g_t(x)$ (i.e., $\mathbb{E}[g_t(x)] = \nabla f_{u_{1,t}u_{2,t}}(x)$) and Lemma 6.

The algorithmic scheme used in the paper is reported in Algorithm 1. At each iteration t we simply build a two point estimate g_t of the gradient related to the smoothed function and then apply a proximal map to the point $x_t - \alpha_t g_t$, with $\alpha_t > 0$ a suitably chosen stepsize.

We let α_t be a diminishing step-size and set

$$u_{1,t} = \alpha_t^2 \quad \text{and} \quad u_{2,t} = \alpha_t^3. \tag{13}$$

Algorithm 1 Proximal Stochastic Derivative Free Algorithm

Input: $x_0 \in \mathbb{R}^n$, sequence $\{\alpha_t\}_{t\geq 0}$, and iteration count T.

For $t = 0, 1, \ldots, T$

Step 1) Sample $\xi_t \sim P$, $Z_1 \sim \mu_1$ and $Z_2 \sim \mu_2$.

Step 2) Set $u_{1,t} = \alpha_t^2$ and $u_{2,t} = \alpha_t^3$.

Step 3) Build the two point estimate $g_t = G(x_t, u_{1,t}, u_{2,t}, Z_{1,t}, Z_{2,t}, \xi_t)$.

Step 4) Set $x_{t+1} = \text{prox}_{\alpha_t r}(x_t - \alpha_t g_t)$.

End For

Sample $t^* \in \{0, \ldots, T\}$ according to $\mathbb{P}(t^* = t) = \alpha_t / \sum_{i=0}^{T} \alpha_i$.

Return x_t^*.

We thus have in our scheme a derivative free version of Algorithm 3.1 reported in [9].

3 Convergence of the Derivative Free Algorithm

We now analyze the convergence properties of Algorithm 1. We follow [9, Section 3.2] in the proof of our results. We consider a value $\bar{\rho} > \rho$, and assume $\alpha_t < \min\left\{\frac{1}{\bar{\rho}}, \frac{\bar{\rho}-\rho}{2}\right\}$ for all t.

We first define the function

$$\phi^{u,t}(x) = f_{u_{1,t}}(x) + r(x),$$

and introduce the Moreau envelope function

$$\phi^{u,t}_{1/\lambda}(x) = \min_y \phi^{u,t}(y) + \frac{\lambda}{2}\|y - x\|^2 \ ,$$

with the proximal map

$$\text{prox}_{\phi^{u,t}/\lambda}(x) = \text{argmin}_y\{\phi^{u,t}(y) + \frac{\lambda}{2}\|y - x\|^2\}.$$

We use the corresponding definition of $\phi_{1/\lambda}(x)$ as well in the convergence theory,

$$\phi_{1/\lambda}(x) = \min_y \phi(y) + \frac{\lambda}{2}\|y - x\|^2 = \min_y f(y) + r(y) + \frac{\lambda}{2}\|y - x\|^2.$$

To begin with let

$$\hat{x}_t = \text{prox}_{\phi^{u,t}/\bar{\rho}}(x_t).$$

Some of the steps follow along the same lines given in [9, Lemma 3.5], owing to the smoothness of $f_{u_{1,t}}(x)$.

We derive the following recursion lemma, which establishes an important descent property for the iterates. We denote by $\mathbb{E}_t$ the conditional expectation with respect to the σ-algebra of random events up to iteration t, i.e., all of $Z_{1,s}$, $Z_{2,s}$ and ξ_s are given for $s < t$, and for $s \geq t$ are random variables. In order to derive this lemma, we require an additional assumption that is reasonable in this setting.

Assumption 3 *The sequence $\{x_t\}$ generated by the algorithm is bounded (i.e., there exists an $M > 0$ s.t., $\|x_t\| \leq M$ for all t).*

Note that this assumption can be satisfied if, for instance, $r(\cdot) = \sum_{j=1}^{J} r_j(\cdot)$ and for at least one $j \in \{1, ...J\}$, $r_j(\cdot)$ is an indicator for a compact set $\mathcal{X}$ (i.e., $r(x) = 0$ if $x \in \mathcal{X}$ and $r(x) = \infty$ otherwise).

Lemma 7 *Let α_t satisfy,*

$$\alpha_t \leq \frac{\bar{\rho} - \rho}{(1 + \bar{\rho}^2 - 2\bar{\rho}\rho + 4\delta_0 L_0)} \ . \tag{14}$$

where $\delta_0 = 1 - \alpha_0\bar{\rho}$.

Then it holds that there exists a B independent of t such that

$$\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 \leq \|x_t - \hat{x}_t\|^2 + \alpha_t^2 B - \alpha_t(\bar{\rho} - \rho)\|x_t - \hat{x}_t\|^2.$$

Proof First we see that $\hat{x}_t$ can be obtained as a proximal point of r:

$$\hat{x}_t = \text{prox}_{\phi^{u,t}/\bar{\rho}}(x_t) \Longleftrightarrow$$

$$\bar{\rho}(x_t - \hat{x}_t) \in \partial r(\hat{x}_t) + \nabla f_{u_{1,t}}(\hat{x}_t) \Longleftrightarrow$$

$$\alpha_t\bar{\rho}(x_t - \hat{x}_t) \in \alpha_t\partial r(\hat{x}_t) + \alpha_t\nabla f_{u_{1,t}}(\hat{x}_t) \Longleftrightarrow$$

$$\alpha_t\bar{\rho}x_t - \alpha_t\nabla f_{u_{1,t}}(\hat{x}_t) + (1-\alpha_t\bar{\rho})\hat{x}_t \in \hat{x}_t + \alpha_t\partial r(\hat{x}_t)$$

$$\Longleftrightarrow \hat{x}_t = \text{prox}_{\alpha_t r}\left(\alpha_t\bar{\rho}x_t - \alpha_t\nabla f_{u_{1,t}}(\hat{x}_t) + (1-\alpha_t\bar{\rho})\hat{x}_t\right).$$

We notice that the last equivalence follows from the optimality conditions related to the proximal subproblem. Letting $\delta_t = 1 - \alpha_t\bar{\rho}$, we get,

$$\begin{aligned}\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 &= \mathbb{E}_t\|\text{prox}_{\alpha_t r}(x_t - \alpha_t g_t) - \text{prox}_{\alpha_t r}(\alpha_t\bar{\rho}x_t - \alpha_t\nabla f_{u_{1,t}}(x_t) + \delta_t\hat{x}_t)\|^2 \\ &\leq \mathbb{E}_t\left\|x_t - \alpha_t g_t - (\alpha_t\bar{\rho}x_t - \alpha_t\nabla f_{u_{1,t}}(\hat{x}_t) + \delta_t\hat{x}_t)\right\|^2,\end{aligned}$$

where the inequality is obtained by considering the non-expansiveness property of the proximal map $\text{prox}_{\alpha_t r}(x)$. We thus can write the following chain of equalities:

$$\begin{aligned}&\mathbb{E}_t\left\|x_t - \alpha_t g_t - (\alpha_t\bar{\rho}x_t - \alpha_t\nabla f_{u,1}(\hat{x}_t) + \delta_t\hat{x}_t)\right\|^2 = \\ &= \mathbb{E}_t\left\|\delta_t(x_t - \hat{x}_t) - \alpha_t(g_t - \nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2 = \\ &= \mathbb{E}_t\left\|\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t)) - \alpha_t(g_t - \nabla f_{u_{1,t}}(x_t))\right\|^2 = \\ &= \mathbb{E}_t\left\|\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2 \\ &\quad -2\alpha_t\mathbb{E}_t\left[\left\langle\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t)), g_t - \nabla f_{u_{1,t}}(x_t)\right\rangle\right] \\ &\quad +\alpha_t^2\mathbb{E}_t\left\|g_t - \nabla f_{u_{1,t}}(x_t)\right\|^2,\end{aligned}$$

with the first equality obtained by rearranging the terms inside the norm, the second one by simply adding and subtracting $\alpha_t\nabla f_{u_{1,t}}(x_t)$ to those terms, and the third one by taking into account the definition of Euclidean norm and the basic properties of the expectation. Now, we get the following

$$\begin{aligned}&\mathbb{E}_t\left\|\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2 \\ &\quad -2\alpha_t\mathbb{E}_t\left[\left\langle\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t)), g_t - \nabla f_{u_{1,t}}(x_t)\right\rangle\right] \\ &\quad +\alpha_t^2\mathbb{E}_t\left\|g_t - \nabla f_{u_{1,t}}(x_t)\right\|^2 \\ &= \left\|\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2 \\ &\quad -2\alpha_t\left[\left\langle\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t)), \mathbb{E}[g_t] - \nabla f_{u_{1,t}}(x_t))\right\rangle\right] \\ &\quad +\alpha_t^2\mathbb{E}_t\left\|g_t - \nabla f_{u_{1,t}}(x_t)\right\|^2 \\ &= \left\|\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2 \\ &\quad -2\alpha_t\left[\left\langle\delta_t(x_t - \hat{x}_t) - \alpha_t(\nabla f_{u_{1,t}}(x_t) - \nabla f_{u_{1,t}}(\hat{x}_t)), \nabla f_{u_{1,t}u_{2,t}}(x_t) - \nabla f_{u_{1,t}}(x_t))\right\rangle\right] \\ &\quad +\alpha_t^2\mathbb{E}_t\left\|g_t - \nabla f_{u_{1,t}}(x_t)\right\|^2.\end{aligned}$$

The first equality, in this case, was obtained by explicitly taking expectation wrt to ξ_t, while we used the unbiasedness of g_t (i.e., $\mathbb{E}[g_t] = \nabla f_{u_{1,t}u_{2,t}}(x_t)$) to

get the second one. We now upper bound the terms in the summation:

$$
\begin{aligned}
&\left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2\\
&\quad -2\alpha_t\left[\left\langle\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t)),\nabla f_{u_{1,t}u_{2,t}}(x_t)-\nabla f_{u_{1,t}}(x_t)\right\rangle\right]\\
&\quad +\alpha_t^2\mathbb{E}_t\left\|g_t-\nabla f_{u_{1,t}}(x_t)\right\|^2\\
&\leq \left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2\\
&\quad -2\alpha_t\left[\left\langle\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t)),\nabla f_{u_{1,t}u_{2,t}}(x_t)-\nabla f_{u_{1,t}}(x_t)\right\rangle\right]\\
&\quad +2\alpha_t^2\mathbb{E}_t\left\|g_t-\nabla f_{u_{1,t},u_{2,t}}(x_t)\right\|^2+2\alpha_t^2\mathbb{E}_t\left\|\nabla f_{u_{1,t},u_{2,t}}(x_t)-\nabla f_{u_{1,t}}(x_t)\right\|^2\\
&\quad \leq \left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2\\
&\quad +2\left(\alpha_t\left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|\right)\left\|\nabla f_{u_{1,t}u_{2,t}}(x_t)-\nabla f_{u_{1,t}}(x_t)\right\|\\
&\quad +2\alpha_t^2\mathbb{E}_t\left\|g_t\right\|^2-4\alpha_t^2\mathbb{E}_t\left\langle g_t(x_t),\nabla f_{u_{1,t},u_{2,t}}(x_t)\right\rangle+2\alpha_t^2\left\|\nabla f_{u_{1,t},u_{2,t}}(x_t)\right\|^2\\
&\qquad +2\alpha_t^2\mathbb{E}_t\left\|\nabla f_{u_{1,t},u_{2,t}}(x_t)-\nabla f_{u_{1,t}}(x_t)\right\|^2\\
&\quad \leq \left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2\\
&\quad +\alpha_t^2\left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2+\alpha_t^2\bar{\sigma}^2\\
&\qquad +2\alpha_t^2\hat{C}-2\alpha_t^2\|\nabla f_{u_{1,t},u_{2,t}}(x)\|^2+2\alpha_t^4\bar{\sigma}^2.\\
&\quad \leq \left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2\\
&\quad +\alpha_t^2\left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2+\alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2+2\alpha_t^2\hat{C}
\end{aligned}
$$

We first split the last term from the previous displayed equation using $(a+b)^2\leq 2a^2+2b^2$ and some basic properties of the expectation. The first inequality was obtained by using Cauchy-Schwarz and by suitably rewriting the third term in the summation. We then used the inequality $2a\cdot b\leq a^2+b^2$ combined with Lemma 4 (or equation (10)) to bound the resulting second term in the summation, that is $\left\|\nabla f_{u_{1,t}u_{2,t}}(x_t)-\nabla f_{u_{1,t}}(x_t)\right\|^2$, inputting equation (13) to obtain the explicit constant and relation with respect to α_t, and Lemma 6 to upper bound the third term, and finally applying the unbiased estimate property of g_t,thus getting the next inequality. Hence we write

$$
\begin{aligned}
&\left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2+\\
&\quad +\alpha_t^2\left\|\delta_t(x_t-\hat{x}_t)-\alpha_t(\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t))\right\|^2+\alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2+2\alpha_t^2\hat{C}\\
&= (1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2-2(1+\alpha_t^2)\delta_t\alpha_t\langle x_t-\hat{x}_t,\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t)\rangle\\
&\quad +(1+\alpha_t^2)\alpha_t^2\|\nabla f_{u_{1,t}}(x_t)-\nabla f_{u_{1,t}}(\hat{x}_t)\|^2+\alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2+2\alpha_t^2\hat{C}\\
&\leq (1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2+2(1+\alpha_t^2)\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|^2+8(1+\alpha_t^2)\delta_t L_0\alpha_t^3\|x_t-\hat{x}_t\|\\
&\quad +(1+\alpha_t^2)\alpha_t^2\left(\|\nabla f_{u_{1,t}}(x_t)\|^2-2\langle\nabla f_{u_{1,t}}(x_t),\nabla f_{u_{1,t}}(\hat{x}_t)\rangle+\|\nabla f_{u_{1,t}}(\hat{x}_t)\|^2\right)\\
&\quad +\alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2+2\alpha_t^2\hat{C}\\
&\leq (1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2+2(1+\alpha_t^2)\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|^2+8(1+\alpha_t^2)\delta_t L_0\alpha_t^3\|x_t-\hat{x}_t\|\\
&\quad +(1+\alpha_t^2)\alpha_t^2\left(\|\nabla f_{u_{1,t}}(x_t)\|^2+2\|\nabla f_{u_{1,t}}(x_t)\|\|\nabla f_{u_{1,t}}(\hat{x}_t)\|+\|\nabla f_{u_{1,t}}(\hat{x}_t)\|^2\right)\\
&\quad +\alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2+2\alpha_t^2\hat{C}\\
&\leq (1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2+2(1+\alpha_t^2)\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|^2+8(1+\alpha_t^2)\delta_t L_0\alpha_t^3\|x_t-\hat{x}_t\|\\
&\quad +4(1+\alpha_t^2)\alpha_t^2\bar{C}+\alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2+2\alpha_t^2\hat{C},
\end{aligned}
$$

where the equality is given by rearranging the terms in the summation and taking into account the definition of Euclidean norm. The inequality is obtained by upper bounding the scalar product by means of Lemma 5 and the

third term in the summation by combining the triangle inequality, the Cauchy-Schwartz inequality and (12). Continuing:

$$
\begin{aligned}
&(1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2 + 2(1+\alpha_t^2)\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|^2 + 8(1+\alpha_t^2)\delta_t L_0\alpha_t^3\|x_t-\hat{x}_t\| \\
&\quad +4(1+\alpha_t^2)\alpha_t^2\bar{C} + \alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2 + 2\alpha_t^2\hat{C} \\
&= (1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2 + 2(1+\alpha_t^2)\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|^2 \\
&\quad +8(1+\alpha_t^2)\delta_t L_0\left(\alpha_t^2\right)(\alpha_t\|x_t-\hat{x}_t\|) \\
&\quad +4(1+\alpha_t^2)\alpha_t^2\bar{C} + \alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2 + 2\alpha_t^2\hat{C} \\
&\leq (1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2 + 2(1+\alpha_t^2)\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|^2 + 4(1+\alpha_t^2)\delta_t L_0\alpha_t^4 \\
&\quad +4(1+\alpha_t^2)\delta_t L_0\alpha_t^2\|x_t-\hat{x}_t\|^2 + 4(1+\alpha_t^2)\alpha_t^2\bar{C} + \alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2 + 2\alpha_t^2\hat{C} \\
&= (1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2 + 2(1+\alpha_t^2)\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|^2 \\
&\quad +4(1+\alpha_t^2)\delta_t L_0\alpha_t^2\|x_t-\hat{x}_t\|^2 + 4(1+\alpha_t^2)\delta_t L_0\alpha_t^4 + 4(1+\alpha_t^2)\alpha_t^2\bar{C} \\
&\quad +\alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2 + 2\alpha_t^2\hat{C}.
\end{aligned}
$$

The first and last equality are simply obtained by rearranging the terms in the summation. The inequality is obtained by upper bounding the third term in the summation using inequality $2a\cdot b \leq a^2+b^2$. Finally, recalling the definition of $\delta_t = 1-\alpha_t\bar{\rho}$, we have

$$
\begin{aligned}
&(1+\alpha_t^2)\delta_t^2\|x_t-\hat{x}_t\|^2 + 2(1+\alpha_t^2)\delta_t\alpha_t\rho\|x_t-\hat{x}_t\|^2 + 4(1+\alpha_t^2)\delta_t L_0\alpha_t^2\|x_t-\hat{x}_t\|^2 \\
&\quad +4(1+\alpha_t^2)\delta_t L_0\alpha_t^4 + 4(1+\alpha_t^2)\alpha_t^2\bar{C} + \alpha_t^2(1+2\alpha_t^2)\bar{\sigma}^2 + 2\alpha_t^2\hat{C} \\
&= \left[1-2\alpha_t\bar{\rho}+\alpha_t^2\bar{\rho}^2+\alpha_t^2-2\alpha_t^3\bar{\rho}+\alpha_t^4\bar{\rho}^2+2\alpha_t\rho-2\alpha_t^2\bar{\rho}\rho+2\alpha_t^3\rho-2\alpha_t^4\bar{\rho}\rho\right. \\
&\quad \left.+4\delta_t L_0\alpha_t^2+4\delta_t L_0\alpha_t^4\right]\|x_t-\hat{x}_t\|^2 \\
&\quad +\left[4(1+\alpha_t^2)\delta_t L_0\alpha_t^2 + 4(1+\alpha_t^2)\bar{C} + (1+2\alpha_t^2)\bar{\sigma}^2 + 2\hat{C}\right]\alpha_t^2 \\
&= \left[1-2\alpha_t(\bar{\rho}-\rho)+\alpha_t^2(1+\bar{\rho}^2-2\bar{\rho}\rho+4\delta_t L_0)-2\alpha_t^3(\bar{\rho}-\rho)\right]\|x_t-\hat{x}_t\|^2 \\
&\quad +\alpha_t^4(\bar{\rho}^2-2\bar{\rho}\rho+4\delta_t L_0)\|x_t-\hat{x}_t\|^2 \\
&\quad +\left[4(1+\alpha_t^2)\delta_t L_0\alpha_t^2 + 4(1+\alpha_t^2)\bar{C} + (1+2\alpha_t^2)\bar{\sigma}^2 + 2\hat{C}\right]\alpha_t^2 \\
&\leq \left[1-2\alpha_t(\bar{\rho}-\rho)+\alpha_t(\bar{\rho}-\rho)-2\alpha_t^3(\bar{\rho}-\rho)\right]\|x_t-\hat{x}_t\|^2 \\
&\quad +\alpha_t^3(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|^2 \\
&\quad +\left[4(1+\alpha_t^2)\delta_t L_0\alpha_t^2 + 4(1+\alpha_t^2)\bar{C} + (1+2\alpha_t^2)\bar{\sigma}^2 + 2\hat{C}\right]\alpha_t^2 \\
&\leq \left[1-\alpha_t(\bar{\rho}-\rho)\right]\|x_t-\hat{x}_t\|^2 \\
&\qquad +\left[4(1+\alpha_t^2)\delta_t L_0\alpha_t^2 + 4(1+\alpha_t^2)\bar{C} + (1+2\alpha_t^2)\bar{\sigma}^2 + 2\hat{C}\right]\alpha_t^2 \\
&:= \left[1-\alpha_t(\bar{\rho}-\rho)\right]\|x_t-\hat{x}_t\|^2 + B\alpha_t^2,
\end{aligned}
$$

where the second to last inequality is obtained by simply considering the expression of α_t in equation (14). We combine the sequence of inequalities shown in the lemma to obtain the result.

After proving Lemma 7, we can now state the main convergence result for Algorithm 1.

Theorem 1 *The sequence generated by Algorithm 1 satisfies,*

$$\mathbb{E}[\phi_{1/\bar{\rho}}(x_{t+1})] \leq \mathbb{E}[\phi_{1/\bar{\rho}}(x_t)] + (2L_0\sqrt{n} + \tfrac{B\bar{\rho}}{2})\alpha_t^2 - \tfrac{\alpha_t(\bar{\rho}-\rho)}{2\bar{\rho}}\mathbb{E}[\|\nabla\phi_{1/\bar{\rho}}^{u,t}(x_t)\|^2]$$

and thus,

$$\begin{aligned}\mathbb{E}[\|\nabla\phi_{1/\bar{\rho}}^{u,t^*}(x_{t^*})\|^2] &= \tfrac{1}{\sum_{t=0}^{T}\alpha_t}\sum_{t=0}^{T}\alpha_t\mathbb{E}[\|\nabla\phi_{1/\bar{\rho}}^{u,t}(x_t)\|^2] \\ &\leq \tfrac{2\bar{\rho}}{\bar{\rho}-\rho}\frac{\phi_{1/\bar{\rho}}(x_0)-\phi_\star+(2L_0\sqrt{n}+\frac{B\bar{\rho}}{2})\sum_{t=0}^{T}\alpha_t^2}{\sum_{t=0}^{T}\alpha_t}.\end{aligned} \tag{15}$$

In particular, if we define α_t to be

$$\alpha_t = \min\left\{\frac{1}{\bar{\rho}}, \frac{\bar{\rho}-\rho}{2}, \sqrt{\frac{\phi_{1/\bar{\rho}}(x_0)-\phi_*}{2L_0\sqrt{n}+\frac{B\bar{\rho}}{2}}}\right\}\frac{1}{\sqrt{T+1}} \tag{16}$$

then,

$$\mathbb{E}[\|\nabla\phi_{1/\bar{\rho}}^{u,t^*}(x_{t^*})\|^2] \leq \frac{4\bar{\rho}}{\bar{\rho}-\rho}\sqrt{\frac{(\phi_{1/\bar{\rho}}(x_0)-\phi_*)(2L_0\sqrt{n}+\frac{B\bar{\rho}}{2})}{T+1}}. \tag{17}$$

Proof We have,

$$\begin{aligned}\mathbb{E}_t[\phi_{1/\bar{\rho}}(x_{t+1})] &\leq \mathbb{E}_t\left[\phi(\hat{x}_t)+\tfrac{\bar{\rho}}{2}\|\hat{x}_t-x_{t+1}\|^2\right] \\ &\leq \phi(\hat{x}_t)+\tfrac{\bar{\rho}}{2}\left(\|x_t-\hat{x}_t\|^2+B\alpha_t^2-\alpha_t(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|^2\right) \\ &\leq \phi^{u,t}(\hat{x}_t)+u_{1,t}L_0\sqrt{n}+\tfrac{\bar{\rho}}{2}\left(\|x_t-\hat{x}_t\|^2+B\alpha_t^2-\alpha_t(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|^2\right),\end{aligned}$$

where the first inequality comes from the definition of the proximal map, the second by considering the result proved in Lemma 7, and the third by Lemma 2. Continuing,

$$\begin{aligned}&\phi^{u,t}(\hat{x}_t)+u_{1,t}L_0\sqrt{n}+\tfrac{\bar{\rho}}{2}\left(\|x_t-\hat{x}_t\|^2+B\alpha_t^2-\alpha_t(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|^2\right) \\ &\quad= \phi_{1/\bar{\rho}}^{u,t}(x_t)+u_{1,t}L_0\sqrt{n}+\tfrac{B\bar{\rho}}{2}\alpha_t^2-\tfrac{\bar{\rho}\alpha_t}{2}(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|^2 \\ &\quad\leq \phi_{1/\bar{\rho}}(x_t)+2u_{1,t}L_0\sqrt{n}+\tfrac{B\bar{\rho}}{2}\alpha_t^2-\tfrac{\bar{\rho}\alpha_t}{2}(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|^2 \\ &\quad\leq \phi_{1/\bar{\rho}}(x_t)+2\alpha_t^2L_0\sqrt{n}+\tfrac{B\bar{\rho}}{2}\alpha_t^2-\tfrac{\bar{\rho}\alpha_t}{2}(\bar{\rho}-\rho)\|x_t-\hat{x}_t\|^2 \\ &\quad= \phi_{1/\bar{\rho}}(x_t)+2\alpha_t^2L_0\sqrt{n}+\tfrac{B\bar{\rho}}{2}\alpha_t^2-\tfrac{\alpha_t(\bar{\rho}-\rho)}{2\bar{\rho}}\|\nabla\phi_{1/\bar{\rho}}^{u,t}(x_t)\|^2,\end{aligned}$$

with the first inequality obtained by Lemma 2. Indeed, let us now call $\bar{x}_t$ the minimizer of $\phi(x_t)+\frac{\bar{\rho}}{2}\|x-x_t\|^2$ and recall that $\hat{x}_t$ is the minimizer of $\phi^{u,t}(x_t)+\frac{\bar{\rho}}{2}\|x-x_t\|^2$. We have

$$\begin{aligned}\phi_{1/\bar{\rho}}^{u,t}(x_t) = \phi^{u,t}(x_t)+\frac{\bar{\rho}}{2}\|\hat{x}_t-x_t\|^2 &\leq \phi^{u,t}(x_t)+\frac{\bar{\rho}}{2}\|\bar{x}_t-x_t\|^2 \\ &\leq \phi(x_t)+u_{1,t}L_0\sqrt{n}+\frac{\bar{\rho}}{2}\|\bar{x}_t-x_t\|^2 \\ &= \phi_{1/\bar{\rho}}(x_t)+u_{1,t}L_0\sqrt{n}.\end{aligned}$$

The second inequality is obtained by using definition of $u_{1,t}$ in (13). The last equality is due basic properties of the Moreau envelope and to the definition of $\hat{x}_t$ (see the beginning of Lemma 7). Now, we take full expectations and obtain:

$$\mathbb{E}[\phi_{1/\bar{\rho}}(x_{t+1})] \leq \mathbb{E}[\phi_{1/\bar{\rho}}(x_t)] + 2\alpha_t^2 L_0\sqrt{n} + \frac{B\bar{\rho}}{2}\alpha_t^2 - \frac{\alpha_t(\bar{\rho}-\rho)}{2\bar{\rho}}\mathbb{E}[\|\nabla\phi_{1/\bar{\rho}}^{u,t}(x_t)\|^2].$$

The rest of the proof is as in [9, Theorem 3.4]. In particular, summing the recursion, we get,

$$\mathbb{E}[\phi_{1/\bar{\rho}}(x_{T+1})] \leq \mathbb{E}[\phi_{1/\bar{\rho}}(x_0)] + (2L_0\sqrt{n} + \frac{B\bar{\rho}}{2})\sum_{t=0}^{T}\alpha_t^2 - \frac{(\bar{\rho}-\rho)}{2\bar{\rho}}\sum_{t=0}^{T}\alpha_t\mathbb{E}[\|\nabla\phi_{1/\bar{\rho}}^{u,t}(x_t)\|^2].$$

Now, noting that

$$\phi_{1/\lambda}(x) = \min_y f(y) + r(y) + \tfrac{\lambda}{2}\|y-x\|^2 \geq \phi_\star,$$

where we used the lower boundedness of $f+r$ in Assumption 1, we can finally state that

$$\frac{1}{\sum_{t=0}^{T}\alpha_t}\sum_{t=0}^{T}\alpha_t\mathbb{E}[\|\nabla\phi_{1/\bar{\rho}}^{u,t}(x_t)\|^2] \leq \leq \frac{2\bar{\rho}}{\bar{\rho}-\rho}\frac{\phi_{1/\bar{\rho}}(x_0)-\phi_\star+(2L_0\sqrt{n}+\frac{B\bar{\rho}}{2})\sum_{t=0}^{T}\alpha_t^2}{\sum_{t=0}^{T}\alpha_t}.$$

Since the left-hand side is by definition $\mathbb{E}[\|\nabla\phi_{1/\bar{\rho}}^{u,t^*}(x_{t^*})\|^2]$, we get the inequality (15). Furthermore, by plugging the expression of α_t given in (16) into (15), we get the final inequality (17).

Theorem 1 gives an overall bound on the weighted expected norm of the proximal map as the statistical measure of distance to convergence with respect to the number of iterations. The worst case bound is weighted by the possible range of the function the algorithm must traverse, i.e., from the starting value to the global minimum, as well as the error in the iterates in traversing this range due to the inaccuracy in the zeroth order function and noisy subgradient approximations. The order of the convergence is the same as the one reported in [9], however, the constant is larger, given the additional error in the quality of the steps. Note that as the convergence result is stated in a similar formalism, using the gradient of the Moreau envelope, we can interpret this approximate stationarity concept as given in [9, pages 3-4], namely that a small value of $\|\nabla\phi_\lambda(x)\|$ implies that x is near some point $\hat{x}$ (specifically $\|x-\hat{x}\| = \lambda\|\nabla\phi_\lambda(x)\|$) satisfying a bound to the distance to stationarity, $\text{dist}(0,\partial\phi(\hat{x})) \leq \|\nabla\phi_\lambda(x)\|$. In this case an additional level of approximation to stationarity is added as we are taking the gradient of the smoothed proximal function, which is itself a perturbation of the original function.

4 Numerical Results

In this section, we investigate the numerical performance of Algorithm 1 on a set of standard weakly convex optimization problems defined in [9]. In particular, we consider phase retrieval, which seeks to minimize the function,

$$\min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \left|\langle a_i, x\rangle^2 - b_i\right| \tag{18}$$

and blind deconvolution, which seeks to minimize

$$\min_{(x,y)\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \left|\langle u_i, x\rangle\langle v_i, y\rangle - b_i\right|. \tag{19}$$

Both of these applications are ones in which Common Random Numbers (CRNs) are a reasonable assumption, making two-point gradient estimates relevant. In particular, in (18), the pairs (a_i, b_i) can be held constant between two function evaluations, and in (19), triplets (u_i, v_i, b_i) can be fixed as well.

4.1 Comparison with methods using a stochastic subgradient oracle

We first compare Algorithm 1 with the stochastic subgradient method and the stochastic proximal method in [9]. The goal in this set of experiments is understanding if our approach is competitive with those ones that use a stochastic subgradient oracle and how the practical behavior of the method fits with the theoretical analysis.

We generate random Gaussian measurements in $N(0, I_{d\times d})$ and a target signal $\bar{x}$ uniformly on the random sphere to compute b_i with dimensions $(d, m) = (10, 30), (20, 60), (40, 120)$. We use fixed stepsizes α_t in the range $[10^{-6}, 10^{-1}]$. We generate ten runs of each algorithm for every dimension and stepsize, and pick the best one according to the final objective value. The total number of iterations used in all cases is 100000.

We show the gap of the different methods when varying the stepsize for both phase retrieval (Figure 1) and blind deconvolution (Figure 2).

It is interesting that the zeroth order algorithm performs on par with the ones that use the stochastic subgradient oracle. In particular, our method is more robust to the choice of the stepsize than the stochastic subgradient method and it is competitive with the proximal method. In Figure 3 and 4, we report the path of the objective values obtained with the stepsize equal to 10^{-4} for the instances $(d, m) = (10, 30)$. These are nice examples of how good the zeroth order algorithm works when the stepsize is properly chosen.

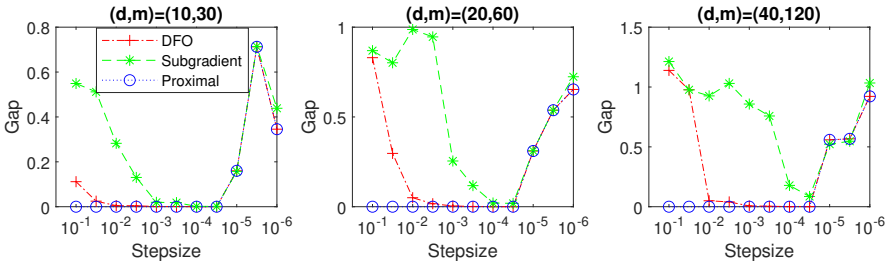

Fig. 1: Gap values for phase retrieval.

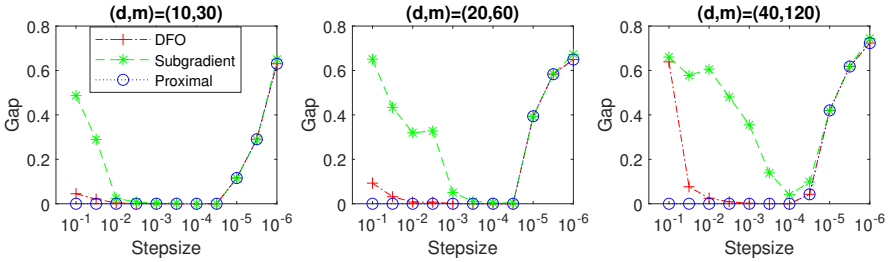

Fig. 2: Gap values for blind deconvolution.

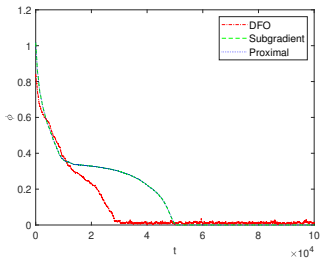

Fig. 3: Convergence of the function values - phase retrieval (d,m)=(10,30).

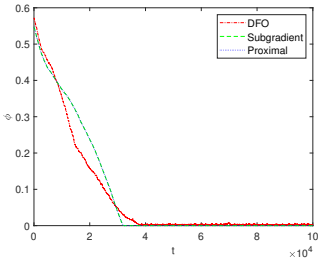

Fig. 4: Convergence of the function values - blind deconvolution (d,m)=(10,30).

4.2 Comparison with a naive stochastic variant of NOMAD

Now, in order to understand if Algorithm 1 is somehow competitive with other (stochastic) non-smooth methods from the DFO literature, we report here a preliminary comparison with a naive stochastic variant of NOMAD [1, 3]. More specifically, we consider a mesh adaptive direct-search (MADS) that

uses a unit-size sample for each evaluation of the zeroth order oracle[1]. We use 100 randomly generated instances in our tests for both phase retrieval and blind deconvolution problems. We generate random Gaussian measurements in $N(0, I_{d\times d})$ and a target signal $\bar{x}$ uniformly on the random sphere to compute b_i with dimensions $(d, m) = (4, 10)$. The choice of restricting the analysis to small dimensional instances is mainly due to the fact that this naive version of `NOMAD` gives very poor performances on larger dimensional instances. We use fixed stepsizes $\alpha_t \in \{10^{-3}, 10^{-2}\}$ in our algorithm. We generate ten runs of each algorithm for every problem and pick the best one according to the final objective value. The total number of function values used in all cases is 10000. We considered data and performance profiles [17] when comparing the methods. Specifically, let S be a set of algorithms and P a set of problems. For each $s \in S$ and $p \in P$, let $t_{p,s}$ be the number of function evaluations required by algorithm s on problem p to satisfy the condition

$$f(x_k) \leq f_L + \tau(f(x_0) - f_L) \tag{20}$$

where $0 < \tau < 1$ and f_L is the best objective function value achieved by any solver on problem p. Then, performance and data profiles of solver s are the following functions

$$\rho_s(\alpha) = \frac{1}{|P|}\left|\left\{p \in P : \frac{t_{p,s}}{\min\{t_{p,s'} : s' \in S\}} \leq \alpha\right\}\right|,$$
$$d_s(\kappa) = \frac{1}{|P|}\left|\{p \in P : t_{p,s} \leq \kappa(n_p + 1)\}\right|$$

where n_p is the dimension of problem p.
We report, in Figure 5 and Figure 6, the data and performance profiles for the experiments on phase retrieval and blind deconvolution problems, respectively. From the plots it can be seen that our algorithm (with suitable choices of the stepsize) outperforms the naive version of `NOMAD` for all precisions. We notice that, when $\tau = 10^{-3}$, `NOMAD` does not appear in the plots, hence it never satisfies the condition (20) for this precision. We further report, in Figure 7 and Figure 8, the box plots related to the function gaps obtained with the algorithms over the 100 instances considered in the tests. Those plots show that our algorithm gets very close to the optimal value for suitable choices of the stepsize.

[1] We would like to notice that the MADS algorithm was originally developed for deterministic blackbox optimization. Recently a stochastic variant of this approach was proposed in [4].

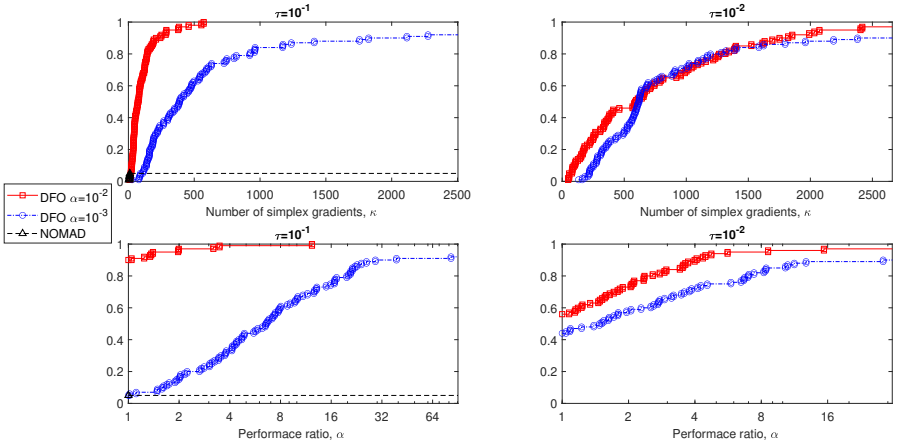

Fig. 5: Comparison between DFO and NOMAD for phase retrieval - Performance and data profiles

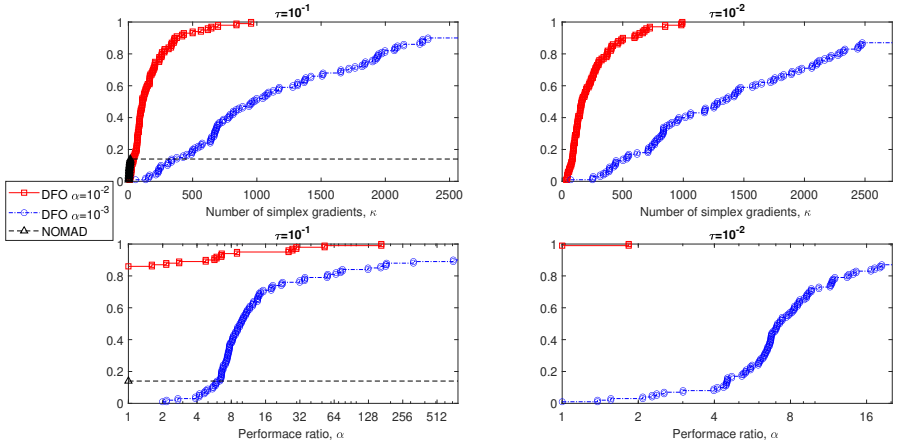

Fig. 6: Comparison between DFO and NOMAD for blind deconvolution - Performance and data profiles

5 Conclusion

In this paper we studied, for the first time, minimization of a stochastic weakly convex function without the presence of an oracle of a noisy estimate of the subgradient of the function, i.e., in the context of derivative-free or zeroth order optimization. We were able to derive theoretical convergence rate results on par with the standard methods for stochastic weakly convex optimization, and demonstrated the algorithm's efficacy on a couple of standard test cases. In expanding the scope of zeroth order optimization, we hope that this work highlights the potential of derivative free methods in general, and the two point

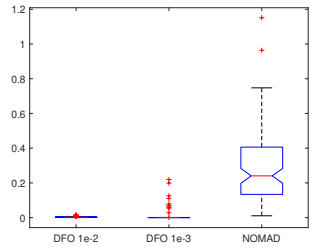

Fig. 7: Box plots function gap - phase retrieval.

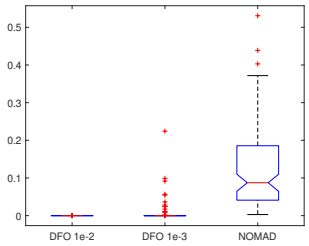

Fig. 8: Box plots function gap - blind deconvolution.

smoothed function approximation technique in particular, to an increasingly wider class of problems.

Acknowledgements. The authors are indebted to the referees for their constructive suggestions which helped to improve on the earlier version of this article.

References

1. C. Audet, S. Le Digabel, C. Tribes, and V. Rochon Montplaisir. The NOMAD project. Software available at `https://www.gerad.ca/nomad/`.
2. Satyajith Amaran, Nikolaos V Sahinidis, Bikram Sharda, and Scott J Bury. Simulation optimization: a review of algorithms and applications. *Annals of Operations Research*, 240(1):351–380, 2016.
3. C. Audet, S. Le Digabel, and C. Tribes. NOMAD user guide. Technical Report G-2009-37, Les cahiers du GERAD, 2009.
4. Charles Audet, Kwassi Joseph Dzahini, Michael Kokkolaras, and Sébastien Le Digabel. Stomads: Stochastic blackbox optimization using probabilistic estimates. *arXiv preprint arXiv:1911.01012*, 2019.
5. Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality, and saddle-points. *arXiv preprint arXiv:1809.06474*, pages 651–676, 2019.
6. Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS journal on optimization*, 1(2):92–119, 2019.
7. Ruobing Chen, Matt Menickelly, and Katya Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, 2018.
8. Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.
9. Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.
10. Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.
11. John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

12. John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.
13. Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
14. Jeffrey Larson and Stephen C Billups. Stochastic derivative-free optimization using a trust region framework. *Computational Optimization and applications*, 64(3):619–645, 2016.
15. Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.
16. Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Jason D Lee. Incremental methods for weakly convex optimization. *arXiv preprint arXiv:1907.11687*, 2019.
17. Jorge J Moré and Stefan M Wild. Benchmarking derivative-free optimization algorithms. *SIAM J. Optim.*, 20(1):172–191, 2009.
18. Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.
19. R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.