Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 13–10

WORST CASE COMPLEXITY OF DIRECT SEARCH UNDER CONVEXITY

M. DODANGEH AND L. N. VICENTE

Abstract: In this paper we prove that the broad class of direct-search methods of directional type, based on imposing sufficient decrease to accept new iterates, exhibits the same global rate or worst case complexity bound of the gradient method for the unconstrained minimization of a convex and smooth function. More precisely, it will be shown that the number of iterations needed to reduce the norm of the gradient of the objective function below a certain threshold is at most proportional to the inverse of the threshold.

Our result is slightly less general than Nesterov's for the gradient method, in the sense that we require more than just convexity of the objective function and boundedness of the initial iterate to the solution set. Our additional condition can, however, be satisfied in several scenarios, such as strong or uniform convexity, boundedness of the initial level set, or boundedness of the distance from the initial contour set to the solution set. It is a mild price to pay for deriving such a global rate for zero-order methods.

Keywords: Derivative-free optimization, direct search, worst case complexity, sufficient decrease, convexity.
AMS Subject Classification (2010): 90C56, 90C25, 90C30.

1. Introduction

In this paper we focus on directional direct-search methods applied to the minimization of a real-valued, convex, and continuously differentiable objective function f, without constraints,

$$\min_{x\in\mathbb{R}^n} \ f(x). \tag{1}$$

In direct-search methods, the objective function is evaluated, at each iteration, at a finite number of points. No derivatives are required. The action of declaring an iteration successful (moving into a point of lower objective function value) or unsuccessful (staying at the same iterate) is based on objective function value comparisons. Some of these methods are directional in the sense of moving along predefined directions along which the objective function will

Date: April 8, 2013.
Support for this work was provided by FCT under grants PTDC/MAT/116736/2010 and PEst-C/MAT/UI0324/2011 and scholarship SFRH/BD/51168/2010.

eventually decrease for sufficiently small step sizes (see, e.g., [4, Chapter 9]). Those of simplicial type (see, e.g., [4, Chapter 8]), such as the Nelder-Mead method, are not considered here. There are essentially two ways of globalizing direct-search methods (of directional type), meaning making them convergent to stationary points independently of the starting point: (i) by integer lattices, insisting on generating points in grids or meshes (which refine only with the decrease of the step size), or (ii) by imposing a sufficient decrease condition, involving the size of the steps, on the acceptance of new iterates. Although we derive our results for the latter strategy, we recall that both share the essentials of these class of direct-search methods: the directional feature for the displacements, and, as in any other direct-search technique, the fact that decisions in each iteration are taken solely by comparison of objective function values.

The analyzes of global convergence of algorithms can be complemented or refined by deriving worst case complexity bounds for the number of iterations or number of function evaluations, an information which becomes valuable in many instances. Such bounds are also called global rates since no assumption on the starting point is made. In terms of the derivation of worst case complexity bounds, Nesterov [10, Page 29] first showed that the steepest descent or gradient method for unconstrained optimization takes at most $\mathcal{O}(\epsilon^{-2})$ iterations (or gradient evaluations) to drive the norm of the gradient of the objective function below $\epsilon \in (0, 1)$. Such a bound or rate has been proved sharp or tight by Cartis, Gould, and Toint [2]. There has been quite an amount of research on global rates for several other classes of algorithms in the non-convex case (see, e.g., [1, 8, 12]).

Derivative-free or zero-order methods have also been recently analyzed with the purpose of establishing their global rates. Vicente [15] has shown a global rate of $\mathcal{O}(\epsilon^{-2})$ for the number of iterations of direct-search methods (of directional type, when imposing sufficient decrease, and applied to a smooth, possibly non-convex function), which translates to $\mathcal{O}(n^2\epsilon^{-2})$ in terms of the number of function evaluations. Cartis, Gould, and Toint [3] have derived a worst case complexity bound of $\mathcal{O}(n^2\epsilon^{3/2})$ for their adaptive cubic overestimation algorithm when using finite differences to approximate derivatives. In the non-smooth case, using smoothing techniques, both Garmanjani and Vicente [7] and Nesterov [11], established a global rate of approximately $\mathcal{O}(\epsilon^{-3})$ iterations (and $\mathcal{O}(n^3\epsilon^{-3})$ function evaluations) for their for zero-order methods, where the threshold ϵ refers now to the gradient of a smoothed version of the original function. Nesterov [11] random Gaussian approach sees its worst case

[9] T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Rev.*, 45:385–482, 2003.
[10] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Dordrecht, 2004.
[11] Y. Nesterov. Random gradient-free minimization of convex functions. Technical Report 2011/1, CORE, 2011.
[12] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton's method and its global performance. *Math. Program.*, 108:177–205, 2006.
[13] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, Berlin, second edition, 2006.
[14] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.
[15] L. N. Vicente. Worst case complexity of direct search. *EURO Journal on Computational Optimization*, 2013, to appear.

M. Dodangeh
Department of Mathematics, University of Coimbra, 3001-501 Coimbra, Portugal (dodangeh@mat.uc.pt).

L. N. Vicente
CMUC, Department of Mathematics, University of Coimbra, 3001-501 Coimbra, Portugal (lnv@mat.uc.pt).

cost in terms of function evaluations reduced to $\mathcal{O}(n^2\epsilon^{-2})$ in the non-convex smooth case and to $\mathcal{O}(n^2\epsilon^{-1})$ in the convex smooth case.

Nesterov [10, Section 2.1.5] has also shown that the gradient method achieves an improved global rate of $\mathcal{O}(\epsilon^{-1})$ if the objective function is convex. It is thus natural to ask if one can achieve a similar rate for zero-order methods, and direct search offers a simple and instructive setting to answer such a question. In this paper, we will show that direct search can indeed achieve a global rate of $\mathcal{O}(\epsilon^{-1})$ under the presence of convexity. The derived worst case complexity bound measures the maximum number of iterations required to find a point where norm of the gradient of the objective function is below ϵ, and, once again, it is proved for directional direct-search methods when a sufficient decrease condition based on the size of the steps is imposed to accept new iterates. As in the non-convex case, the corresponding maximum number of objective function evaluations becomes $\mathcal{O}(n^2\epsilon^{-1})$.

The structure of the paper is as follows. In Section 2, we briefly comment on the worst case complexity (WCC) bounds or global rates of the gradient or steepest descent method. In Section 3, we describe the class of direct search under consideration and provide the known results (global asymptotics and global rates) for the smooth (continuously differentiable) and non-convex case. Then, in Section 4, we derive the WCC bound of $\mathcal{O}(\epsilon^{-1})$ iterations ($\mathcal{O}(n^2\epsilon^{-1})$ function evaluations) for such direct-search methods in the also smooth but now convex case.

This result is derived under a bound R on the distance of all unsuccessful iterates to the solution set, being the WCC bounds actually of the type $\mathcal{O}(R\epsilon^{-1})$ or $\mathcal{O}(n^2R\epsilon^{-1})$. It is proved in Section 5 that such a bound R holds under strong or uniform convexity, boundedness of the initial level set, or boundedness of the distance from the initial contour set to the solution set. In Section 6, we exhibit a parameterized family of strongly convex functions satisfying the assumption required to derive the global rate of $\mathcal{O}(\epsilon^{-1})$ for gradient-type methods in the convex case. Such an assumption requires a bound on the product of the distance of (only) the initial iterate to the solution set by the Lipschitz constant of the gradient of f. We will see, however, that the distance from the first unsuccessful iterate to the solution set grows to infinity as a function of the parameter of the family of functions (which is in turn the inverse of the stationarity threshold ϵ), in other words that $R = \mathcal{O}(\epsilon^{-1})$, showing that the bound $\mathcal{O}(\epsilon^{-1})$ cannot be in general secured and that our additional assumption is sharp.

In Section 7 we draw some concluding remarks based on the specifics of the material covered during the paper. We note that the notation $\mathcal{O}(M)$ has meant and will mean a multiple of M, where the constant multiplying M does not depend on the iteration counter k of the method under analysis (thus depending only on f or on algorithmic constants set at the initialization of the method). The dependence of M on the dimension n of the problem will be made explicit whenever appropriate. The vector norms will be ℓ_2 ones. Given an open subset Ω of $\mathbb{R}^n$, we denote by $\mathcal{C}^1_\nu(\Omega)$ the set of continuously differentiable functions in Ω with Lipschitz continuous gradient in Ω, where ν is the Lipschitz constant of the gradient. We use the notation $\mathcal{F}(\Omega)$ to represent the space of convex functions defined on a convex set Ω. The intersection of both is denoted by $\mathcal{F}^1_\nu(\Omega) = \mathcal{F}(\Omega) \cap \mathcal{C}^1_\nu(\Omega)$, where Ω is open and convex.

2. WCC of gradient-type methods

Given a starting point $x_0 \in \mathbb{R}^n$, the gradient or steepest descent method takes the form $x_{k+1} = x_k - h_k \nabla f(x_k)$, where $h_k > 0$ defines the step size. The algorithm can be applied whenever the function f is continuously differentiable, and the well known fact that $-\nabla f(x_k)$ is a descent direction provides the basis for the convergence properties of the method. The update of the step size h_k is also a crucial point in this class of minimization algorithms. There are improper choices of the step size that make such gradient-type algorithms diverge [13, Chapter 3]. The proper update of the step size is thus central in achieving global convergence (see, e.g., [10, 13]).

For a number of the well known strategies to update the step size, it is possible to prove that, when $f \in \mathcal{C}^1_\nu(\mathbb{R}^n)$, there is a constant $C = C(\nu) > 0$ such

$$f(x_k) - f(x_{k+1}) \;\geq\; C(\nu)\|\nabla f(x_k)\|^2, \tag{2}$$

where $C(\nu)$ is essentially a multiple of $1/\nu$, with ν the Lipschitz constant of the gradient of f, (being the multiple dependent on the parameters involved in the update of the step size). In such cases, assuming that f is also bounded from below in $\mathbb{R}^n$, one can show that the gradient method takes at most $\mathcal{O}(\epsilon^{-2})$ iterations to reduce the gradient below $\epsilon \in (0,1)$ (see [10, Page 29]), to be more specific

$$\left(\frac{f(x_0) - f_{low}}{C(\nu)}\right)\frac{1}{\epsilon^2}.$$

that a maximum of $\mathcal{O}(\epsilon^{-1})$ iterations and $\mathcal{O}(n^2\epsilon^{-1})$ function evaluations are required to compute a point for which the norm of the gradient of the objective function f is smaller than ϵ (see Corollaries 4.1–4.2). Such a global rate translates into a decrease of $\mathcal{O}(1/k)$ for $f(x_k) - f_*$ for the sequence $\{x_k\}_{k\in\mathcal{S}}$ of successful iterations (see Lemma 4.1).

This result is not obtained for all convex functions (for which the solution set is non-empty), as it is the case for the gradient method. In fact, one has seen that one cannot accommodate excessively flatness in the function while approaching the solution set.

The type of non-asymptotic analysis developed in this paper provides cost or complexity bounds in the worst case, involving an order of accuracy. The multiple of this accuracy depends in turn on a significant number of constants, some coming from the algorithm, others problem dependant. In Section 6, for instance, we looked at one of such constants who played a major role in our analysis.

However, the combined effect of all the constants might make a global rate look either worse or not so bad. Moreover, the practical or average behavior of the algorithm might be much below such worst case complexity bounds. In the forthcoming PhD thesis [5], a number of numerical experiments will be reported to better illustrate these two issues for direct search on convex functions.

References

[1] N. I. M. Gould C. Cartis and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity. *Math. Program.*, 130:295–319, 2011.

[2] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. *SIAM J. Optim.*, 20:2833–2852, 2010.

[3] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM J. Optim.*, 22:66–86, 2012.

[4] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, 2009.

[5] M. Dodangeh. *Worst Case Complexity of Direct Search under Convexity*. PhD thesis, Dept. Mathematics, Univ. Coimbra, 2014, forthcoming.

[6] E. D. Dolan, R. M. Lewis, and V. Torczon. On the local convergence of pattern search. *SIAM J. Optim.*, 14:567–583, 2003.

[7] R. Garmanjani and L. N. Vicente. Smoothing and worst-case complexity for direct-search methods in nonsmooth optimization. *IMA J. Numer. Anal.*, to appear.

[8] S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM J. Optim.*, 19:414–444, 2008.

where in the last inequality we used $2\bar{k}+1 \geq \varepsilon$, or equivalently, $2-(2\bar{k}+1)\varepsilon^{-1} \leq 1$. Then we test the direction $(0, y_0\varepsilon^{-1})$

$$\begin{aligned}
& f(x_{\bar{k}}) - f(x_{\bar{k}} + (0, y_0\varepsilon^{-1})) \\
= \; & f(-\varepsilon+\bar{k}, y_0(1-\bar{k}\varepsilon^{-1})) - f(-\varepsilon+\bar{k}, y_0(1-(\bar{k}-1)\varepsilon^{-1})) \\
= \; & [y_0^2(1-\bar{k}\varepsilon^{-1})^2 + 0.5(-1+\bar{k}\varepsilon^{-1})^2 - 1 + \bar{k}\varepsilon^{-1}] \\
& -[y_0^2(1-(\bar{k}-1)\varepsilon^{-1})^2 + 0.5(-1+\bar{k}\varepsilon^{-1})^2 - 1 + \bar{k}\varepsilon^{-1}] \\
= \; & y_0^2[(1-\bar{k}\varepsilon^{-1})^2 - (1+(1-\bar{k})\varepsilon^{-1})^2] \\
= \; & y_0^2[1-\bar{k}\varepsilon^{-1} - (1+(1-\bar{k})\varepsilon^{-1})][1-\bar{k}\varepsilon^{-1} + 1 + (1-\bar{k})\varepsilon^{-1}] \\
= \; & -y_0^2\varepsilon^{-1}[2-(2\bar{k}-1)\varepsilon^{-1}] \; < \; 0,
\end{aligned}$$

where in the last inequality we used $2\bar{k}-1 < \varepsilon$. Finally, we look at the decrease along the direction $(-1, 0)$

$$\begin{aligned}
& f(z_{\bar{k}}) - f(z_{\bar{k}} + (-1, 0)) \\
= \; & f(-\varepsilon+\bar{k}, y_0(1-\bar{k}\varepsilon^{-1})) - f(-\varepsilon+\bar{k}-1, y_0(1-\bar{k}\varepsilon^{-1})) \\
= \; & [y_0^2(1-\bar{k}\varepsilon^{-1})^2 + 0.5(-1+\bar{k}\varepsilon^{-1})^2 - 1 + \bar{k}\varepsilon^{-1}] \\
& -[y_0^2(1-\bar{k}\varepsilon^{-1})^2 + 0.5(-1+(\bar{k}-1)\varepsilon^{-1})^2 - 1 + (\bar{k}-1)\varepsilon^{-1}] \\
= \; & 0.5[-1+\bar{k}\varepsilon^{-1} + 1 - (\bar{k}-1)\varepsilon^{-1}][-1+\bar{k}\varepsilon^{-1} - 1 + (\bar{k}-1)\varepsilon^{-1}] + \varepsilon^{-1} \\
= \; & 0.5\varepsilon^{-1}[-2 + 2\bar{k}\varepsilon^{-1} - \varepsilon^{-1}] + \varepsilon^{-1} \\
< \; & 0.5\varepsilon^{-1} \; < \; \mathcal{C},
\end{aligned}$$

where, again, in the first inequality we used $2\bar{k} - 1 < \varepsilon$.

One can see how does this example illustrate our theory. Recalling that the distance from the unsuccessful iterate $x_{\bar{k}}$ to X_*^f is arbitrarily large, $\|x_{\bar{k}} - x_*\| \geq (\varepsilon-1)/2$, one has that $R = \mathcal{O}(\varepsilon)$ in Assumption 4.1, and by setting $\epsilon = 1/\varepsilon$, $R = \mathcal{O}(\epsilon^{-1})$. From Theorem 4.1 and regardless of how large ω in (7) can be, one sees immediately that our theory cannot predict better than $\mathcal{O}(\epsilon^{-2})$ (when $p = 2$) as a worst case complexity bound. If one inspects better how $\epsilon = \varepsilon^{-1}$ influences ω (note that $\mathcal{C} = \varepsilon^{-1}$ and $d_{\min} = y_0\varepsilon^{-1}$), one comes to the conclusion that not even better than $\mathcal{O}(\epsilon^{-3})$ can be predicted in this example.

7. Conclusions

To our knowledge it is the second time that a derivative-free method is shown to exhibit a worst case complexity bound or global rate of $\mathcal{O}(\epsilon^{-1})$ in the convex case, following the random Gaussian approach [11]. In fact we have proved

The constant multiplying ϵ^{-2} depends thus only on ν, on the parameters involved in the update of the step size, on and the lower bound f_{low} for f in $L_f(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$.

If, additionally, f is assumed convex, i.e., $f \in \mathcal{F}_\nu^1(\mathbb{R}^n)$, then Nesterov [10, Section 2.1.5] showed that one can achieve a better worst case complexity bound in terms of the negative power of ϵ. First, based on the geometric properties of smooth convex functions (essentially [10, Equation (2.1.7)]), he proved, for simplicity using $h_k = 1/\nu$, that

$$f(x_m) - f_* \; \leq \; \frac{2\nu\|x_0 - x_*\|^2}{m+4}, \tag{3}$$

where f_* is the value of the function at a (global) minimizer (see [10, Corollary 2.1.2]), assumed to exist. But then one can easily see, by repeatedly applying (2), that for $m < k$

$$\frac{2\nu\|x_0 - x_*\|^2}{m+4} \; \geq \; C(\nu)\sum_{\ell=m}^{k} \|\nabla f(x_\ell)\|^2.$$

By choosing $k = 2m$ the gradient method is then proved to only take at most $\mathcal{O}(\epsilon^{-1})$ iterations to achieve a threshold of ϵ on the norm of the gradient. The constant multiplying ϵ^{-1} is essentially a multiple of

$$\nu\|x_0 - x_*\|.$$

3. WCC of direct search

The direct-search method under analysis is described in Algorithm 3.1, following the presentation in [4, Chapter 7]. The directional feature is presented in the poll step, where points of the form $x_k + \alpha_k d$, for directions d belonging to the positive spanning set D_k, are tested for sufficient decrease. For this purpose, following the terminology in [9], $\rho : (0,\infty) \to (0,\infty)$ will represent a forcing function, i.e., a non-decreasing (continuous) function satisfying $\lim_{t\to 0} \frac{\rho(t)}{t} = 0$. Typical examples of forcing functions are $\rho(t) = \mathcal{C}t^p$, for $p > 1$ and $\mathcal{C} > 0$. The poll step is successful if the value of the objective function is sufficiently decreased relatively to the step size α_k, in the sense of $f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k)$, in which case the step size is possibly increased. The algorithm opportunistically moves to the first of such points found. Failure in doing so defines an unsuccessful iteration, and the step size is decreased by a factor strictly less than 1 that changes between two bounds which need to be fixed during the course of the iterations. The search step is purposely

left open since it does not interfere in any of the convergence properties of the algorithm, and it is solely used to improve the practical performance of the overall algorithm.

Algorithm 3.1 (Directional direct-search method).

Initialization

Choose x_0 with $f(x_0) < +\infty$, $\alpha_0 > 0$, $0 < \beta_1 \leq \beta_2 < 1$, and $\gamma \geq 1$.

For $k = 0, 1, 2, \ldots$

(1) **Search step:** Try to compute a point with $f(x) < f(x_k) - \rho(\alpha_k)$ by evaluating the function f at a finite number of points. If such a point is found, then set $x_{k+1} = x$, declare the iteration and the search step successful, and skip the poll step.

(2) **Poll step:** Choose a positive spanning set D_k. Order the set of poll points $P_k = \{x_k + \alpha_k d : d \in D_k\}$. Start evaluating f at the poll points following the chosen order. If a poll point $x_k + \alpha_k d_k$ is found such that $f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k)$, then stop polling, set $x_{k+1} = x_k + \alpha_k d_k$, and declare the iteration and the poll step successful. Otherwise, declare the iteration (and the poll step) unsuccessful and set $x_{k+1} = x_k$.

(3) **Mesh parameter update:** If the iteration was successful, then maintain or increase the step size parameter: $\alpha_{k+1} \in [\alpha_k, \gamma\alpha_k]$. Otherwise, decrease the step size parameter: $\alpha_{k+1} \in [\beta_1\alpha_k, \beta_2\alpha_k]$.

When the objective function is bounded from below one can prove that there exists a subsequence of unsuccessful iterates driving the step size parameter to zero (see [9] or [4, Theorems 7.1 and 7.11 and Corollary 7.2]).

Lemma 3.1. *Let f be bounded from below on $L(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$. Then Algorithm 3.1 generates an infinite subsequence K of unsuccessful iterates for which* $\lim_{k \in K} \alpha_k = 0$.

Note that when the function f is convex and has a minimizer, it is necessarily bounded from below (see, e.g., [10, Theorem 2.1.1]).

To continue towards the global properties (asymptotic convergence and rates) for this class of direct search, one must look at the key feature of a positive spanning set, its cosine measure [9]. Given a positive spanning set D (with

one observes that the sufficient decrease condition always holds

$$\begin{aligned}
& f(x_{k-1}) - f(x_k) \\
&= [y_0^2(1-(k-1)\varepsilon^{-1})^2 + 0.5(-1+(k-1)\varepsilon^{-1})^2 - 1 + (k-1)\varepsilon^{-1}] \\
&\quad -[y_0^2(1-k\varepsilon^{-1})^2 + 0.5(-1+k\varepsilon^{-1})^2 - 1 + k\varepsilon^{-1}] \\
&= y_0^2[1-(k-1)\varepsilon^{-1} - (1-k\varepsilon^{-1})][1-(k-1)\varepsilon^{-1} + 1 - k\varepsilon^{-1}] \\
&\quad +0.5[-1+(k-1)\varepsilon^{-1} + 1 - k\varepsilon^{-1}][-1+(k-1)\varepsilon^{-1} - 1 + k\varepsilon^{-1}] - \varepsilon^{-1} \\
&= y_0^2\varepsilon^{-1}[2-(2k-1)\varepsilon^{-1}] - 0.5\varepsilon^{-1}[-2+(2k-1)\varepsilon^{-1}] - \varepsilon^{-1} \\
&> \varepsilon^{-1} = \mathcal{C}\alpha_k,
\end{aligned}$$

where in the last inequality we used $2k-1 < \varepsilon$, or equivalently, $2-(2k-1)\varepsilon^{-1} > 1$.

The gradient of f at x_k is given by

$$\nabla f(x_k) = (k\varepsilon^{-2}, 2y_0(1-k\varepsilon^{-1})),$$

and, recalling $\varepsilon > 2$ and $2k-1 < \varepsilon$, one has $\|\nabla f(x_k)\| > \varepsilon^{-1}$. The distant to the minimizer is,

$$\|x_k - x_*\| = \sqrt{k^2 + y_0^2(1-k\varepsilon^{-1})^2} \geq k.$$

Let $\bar{k}$ be the largest integer number such that $2k-1 < \varepsilon$. Thus $2\bar{k}+1 \geq \varepsilon$, and consequently, $\bar{k} \geq \frac{1}{2}(\varepsilon - 1)$. Hence $\|x_{\bar{k}} - x_*\| \geq (\varepsilon-1)/2$.

To show that $\bar{k}$ is unsuccessful, we start by showing that the sufficient decrease condition is not satisfied along the direction $(1, -y_0\varepsilon^{-1})$

$$\begin{aligned}
& f(x_{\bar{k}}) - f(x_{\bar{k}} + (1, -y_0\varepsilon^{-1})) \\
&= f(-\varepsilon + \bar{k}, y_0(1-\bar{k}\varepsilon^{-1})) - f(-\varepsilon + \bar{k} + 1, y_0(1-(\bar{k}+1)\varepsilon^{-1})) \\
&= [y_0^2(1-\bar{k}\varepsilon^{-1})^2 + 0.5(-1+\bar{k}\varepsilon^{-1})^2 + (-1+\bar{k}\varepsilon^{-1})] \\
&\quad -[y_0^2(1-(\bar{k}+1)\varepsilon^{-1})^2 + 0.5(-1+(\bar{k}+1)\varepsilon^{-1})^2 + (-1+(\bar{k}+1)\varepsilon^{-1})] \\
&= y_0^2[1-\bar{k}\varepsilon^{-1} - 1 + (\bar{k}+1)\varepsilon^{-1}][1-\bar{k}\varepsilon^{-1} + 1 - (\bar{k}+1)\varepsilon^{-1}] \\
&\quad +0.5[-1+\bar{k}\varepsilon^{-1} + 1 - (\bar{k}+1)\varepsilon^{-1}][-1+\bar{k}\varepsilon^{-1} - 1 + (\bar{k}+1)\varepsilon^{-1}] - \varepsilon^{-1} \\
&= y_0^2\varepsilon^{-1}[2-(2\bar{k}+1)\varepsilon^{-1}] - 0.5\varepsilon^{-1}[-2+(2\bar{k}+1)\varepsilon^{-1}] - \varepsilon^{-1} \\
&\leq (y_0^2 - 0.5)\varepsilon^{-1} = \mathcal{C},
\end{aligned}$$

6. A convex example

We have already mentioned in Section 2 that the gradient method exhibits a worst case complexity bound of $\mathcal{O}(\epsilon^{-1})$ iterations as long as $\nu\|x_0 - x_*\|$, for some $x_* \in X_*^f$, is independent of ϵ, where ν is the Lipschitz constant of the gradient. Moreover, Nesterov [10, Theorem 2.1.13] showed that the gradient method for a constant step size of at most $\frac{2}{\nu}$ generates iterates such that $\|x_k - x_*\| \leq \|x_0 - x_*\|$.

Making directional direct-search methods achieve the same global rate of $\mathcal{O}(\epsilon^{-1})$ required a bound (independent of ϵ) on the distance of all unsuccessful iterations to the solution set (see Assumption 4.1). In fact the lack of knowledge of the gradient makes the control of the distance to the solution set harder, which as the following example will demonstrate can become arbitrarily large.

Let $\varepsilon \in (2, \infty)$ and consider the application of Algorithm 3.1 to the strongly convex function $f : \mathbb{R}^2 \to \mathbb{R}$ parameterized by ε

$$f(x, y) \;=\; y^2 + \frac{1}{2}(\varepsilon^{-1}x)^2 + \varepsilon^{-1}x,$$

using $p > 1$, $\gamma = 1$, $x_0 = (-\varepsilon, y_0)$, $\alpha_0 = 1$, $\mathcal{C} = \varepsilon^{-1}$, and

$$D \;=\; \begin{bmatrix} 1 & 0 & -1 \\ -y_0\varepsilon^{-1} & y_0\varepsilon^{-1} & 0 \end{bmatrix},$$

with $y_0^2 = 1.5$. The choices of β_1 and β_2 are irrelevant for our argument. The unique minimizer of f is $x_* = (-\varepsilon, 0)$. The Lipschitz constant ν of the gradient of f is at most 2.

Note that for such an initial point x_0, one has $\nu\|x_0 - x_*\| \leq 2y_0 = \sqrt{6}$ and thus one does expect the global rate $\mathcal{O}(\epsilon^{-1})$ to hold for gradient methods.

Let us see what happens with the above instance of Algorithm 3.1 on this example. First notice that as long as $2k - 1 < \varepsilon$ the algorithm keeps making successful iterations along the direction $(1, -y_0\varepsilon^{-1})$ maintaining the initial step size 1. In fact, by setting

$$x_k \;=\; (-\varepsilon + k, y_0(1 - k\varepsilon^{-1})),$$

nonzero vectors), its cosine measure is given by

$$\mathrm{cm}(D) \;=\; \min_{0 \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{v^\top d}{\|v\|\|d\|}.$$

Any positive spanning set with nonzero vectors has a positive cosine measure. This fact means for any non-zero vector, in particular the negative gradient at a given point, there is at least one direction in D making an acute angle with it. Such a property enables us to derive that the norm of the gradient is of the order of the step size when an unsuccessful iteration occurs [6, 9] (see also [4, Theorem 2.4 and Equation (7.14)]).

Theorem 3.1. *Let D_k be a positive spanning set and $\alpha_k > 0$ be given. Assume that ∇f is Lipschitz continuous (with constant $\nu > 0$) in an open set containing all the poll points in P_k. If $f(x_k) \leq f(x_k + \alpha_k d) + \rho(\alpha_k)$, for all $d \in D_k$, i.e., the iteration k is unsuccessful, then*

$$\|\nabla f(x_k)\| \;\leq\; \mathrm{cm}(D_k)^{-1} \left(\frac{\nu}{2}\alpha_k \max_{d \in D_k} \|d\| + \frac{\rho(\alpha_k)}{\alpha_k \min_{d \in D_k} \|d\|} \right). \tag{4}$$

It becomes then obvious that one needs to avoid degenerate positive spanning sets.

Assumption 3.1. *All positive spanning sets D_k used for polling (for all k) must satisfy $\mathrm{cm}(D_k) \geq \mathrm{cm}_{\min}$ and $d_{\min} \leq \|d\| \leq d_{max}$ for all $d \in D_k$ (where $\mathrm{cm}_{\min} > 0$ and $0 < d_{\min} < d_{\max}$ are constants).*

A first global asymptotic result is then easily obtained by combining Lemma 3.1 and Theorem 3.1 (under Assumption 3.1), and assures the convergence to zero of the gradient at a subsequence of unsuccessful iterates. Moreover, we have the following worst case complexity bounds in this general non-convex, smooth setting [15].

Theorem 3.2. *Consider the application of Algorithm 3.1 when $\rho(\alpha) = \mathcal{C}\alpha^p$, $p > 1$, $\mathcal{C} > 0$, and D_k satisfies Assumption 3.1. Let f be bounded from below in $L_f(x_0)$ and $f \in \mathcal{C}^1_\nu(\Omega)$ where Ω is an open set containing $L_f(x_0)$.*

Under these assumptions, to reduce the gradient below $\epsilon \in (0, 1)$, Algorithm 3.1 takes at most

$$\mathcal{O}((\sqrt{n}\nu\epsilon^{-1})^{\hat{p}}),$$

iterations, and at most

$$\mathcal{O}(n(\sqrt{n}\nu\epsilon^{-1})^{\hat{p}}).$$

function evaluations, where $\hat{p} = \frac{p}{\min(1,p-1)}$.

When $p = 2$, *these numbers are of* $\mathcal{O}(n\nu^2\epsilon^{-2})$ *and* $\mathcal{O}(n^2\nu^2\epsilon^{-2})$, *respectively.*

The constant in $\mathcal{O}(\cdot)$ *depends only on* $d_{\min}$, $d_{\max}$, $\mathrm{cm}_{\min}$, $\mathcal{C}$, p, β_1, β_2, γ, α_0, *and on the lower bound of* f *in* $L_f(x_0)$.

How the step size α_k is updated impacts in several ways the global rates given above for Algorithm 3.1. In fact, the choice of $\mathcal{C}$ in the forcing function and the choice of the parameters β_1, β_2, and γ in the step size updating formulas influence the constant in the bound (4). Increasing $\mathcal{C}$, for instance, will decrease the number of successful iterations [15, Theorem 3.1], possibly leading to more unnecessary unsuccessful iterations and consequently more unnecessary function evaluations. Increasing the value of the expansion factor $\gamma \geq 1$ will increase the maximum number of unsuccessful iterations compared to the number of successful ones [15, Theorem 3.2], again possibly leading to more unnecessary unsuccessful iterations and consequently more unnecessary function evaluations. Setting $\gamma = 1$ leads to an optimal choice in this respect. One practical strategy to accommodate $\gamma > 1$ is by considering an upper bound for the step size itself.

Assumption 3.2. *There is a positive constant* M *such that* $\alpha_k \leq M$ *for* $\forall k \geq 0$.

Under this assumption Theorem 3.1 simplifies to the following:

Corollary 3.1. *Consider* $\rho(\alpha_k) = \mathcal{C}\alpha_k^p$, $p > 1$, $\mathcal{C} > 0$. *Under the assumptions of Theorem 3.1 and Assumptions 3.1 and 3.2, if* $f(x_k) \leq f(x_k + \alpha_k d) + \rho(\alpha_k)$, *for all* $d \in D_k$, *i.e., the iteration* k *is unsuccessful, then*

$$\|\nabla f(x_k)\| \;\leq\; \mathrm{cm}_{\min}^{-1} \frac{\frac{\nu}{2} d_{\max} M + \mathcal{C} d_{\min}^{-1} M^{p-1}}{M^{\min(1,p-1)}} {\alpha_k}^{\min(1,p-1)}. \tag{5}$$

The step size upper bound M will appear thus in the upper bound for the gradient in unsuccessful iterations. When $p = 2$, the upper bound on the gradient does not depends on M,

$$\|\nabla f(x_k)\| \;\leq\; \mathrm{cm}_{\min}^{-1} \left(\frac{\nu}{2} d_{\max} + \mathcal{C} d_{\min}^{-1}\right) \alpha_k.$$

The analysis of worst case complexity for the convex case when $p \neq 2$ will, however, depend on the upper bound M for the step size.

Next we show that Assumption 4.1 will hold for strongly convex functions, being the constant R there of $\mathcal{O}(1/\sqrt{\mu})$.

Proposition 5.2. *Let* f *be a continuous and strongly convex function in* $\mathbb{R}^n$ *with constant* μ. *Then*

$$\sup_{y \in L_f(x_0)} \mathrm{dist}(y, X_*^f) \;\leq\; \sqrt{\frac{2}{\mu}(f(x_0) - f_*)}.$$

Proof: Under the assumptions of the proposition, X_*^f is a singleton and let x_* be the minimizer of f. It is simple to see that we must also have

$$\frac{1}{2}\mu\|y - x_*\|^2 \;\leq\; f(y) - f(x_*), \quad \forall y \in \mathbb{R}^n.$$

Otherwise, there must exist a $y \in \mathbb{R}^n$ violating the above inequality. Then, by using the definition of strongly convexity with

$$0 \;\leq\; t < \; 1 - \frac{f(y) - f(x_*)}{\frac{\mu}{2}\|y - x_*\|^2},$$

one would obtain

$$\begin{aligned} f(ty + (1-t)x_*) \;&\leq\; tf(y) + (1-t)f(x_*) - t(1-t)\frac{\mu}{2}\|y - x_*\|^2 \\ &=\; f(x_*) + t\left[f(y) - f(x_*) - \frac{\mu}{2}\|y - x_*\|^2\right] + \frac{t^2}{2}\mu\|y - x_*\|^2 \\ &<\; f(x_*), \end{aligned}$$

which is a contradiction. In particular, for the level set $L_f(x_0)$, one has $\frac{1}{2}\mu\|y - x_*\|^2 \leq f(y) - f(x_*) \leq f(x_0) - f(x_*)$, $\forall y \in L_f(x_0)$, as it was desired. ∎

Another possible situation under which Assumption 4.1 is satisfied is when the initial level set is bounded. The proof is trivial and omitted.

Proposition 5.3. *Let* $f \in \mathcal{F}(\mathbb{R}^n)$ *and* X_f^* *be non-empty and assume that* $L_f(x_0)$ *is bounded. Then*

$$\sup_{y \in L_f(x_0)} \mathrm{dist}(y, X_*^f) \;\leq\; 2 \sup_{x \in L_f(x_0)} \|x\|.$$

Note that there are convex functions f such that $\sup_{y \in L_f(x_0)} \mathrm{dist}(y, X_*^f)$ is finite but neither f is strongly convex nor $L_f(x)$ is bounded for any x, being such an instance the two-dimensional function $f(x, y) = y^2$.

The constant in $\mathcal{O}(\cdot)$ depends only on $d_{\min}$, $d_{\max}$, $\mathrm{cm}_{\min}$, $\mathcal{C}$, p, β_1, β_2, γ, α_0, and on the constant R of Assumption 4.1.

5. Discussion of the assumptions

Now we are going to exhibit three situations under which Assumption 4.1 is verified. First we clarify that the distance from the initial level set to the solution set is never larger than the distance from the first contour set to the same solution set.

In addition to being assumed non-empty, note that the solution set X_*^f is also convex, since $f \in \mathcal{F}(\mathbb{R}^n)$. Such a solution set will always be closed in our context since we will assume that the function is continuous (an assumption that would also result from the observation that f is real-valued and convex in $\mathbb{R}^n$, see [14, Theorem 10.1]). Thus, projecting onto X_*^f results in an unique point.

Proposition 5.1. *Let $f \in \mathcal{F}^1(\mathbb{R}^n)$ and X_f^* be non-empty. Then*

$$\sup_{y \in L_f(x_0)} \operatorname{dist}(y, X_*^f) \;=\; \sup_{y \in f^{-1}(f(x_0))} \operatorname{dist}(y, X_*^f).$$

Proof: For a given z in $L_f(x_0)$ but not in X_*^f, there is a unique $x_* \in X_*^f$ such that

$$\operatorname{dist}(z, X_*^f) \;=\; \min_{x \in X_*^f} \|z - x\| \;=\; \|z - x_*\|.$$

Let us consider the real function $g(t) = f(x_* + t(z - x_*))$ for $t \in [0, \infty)$, which belongs to $\mathcal{F}^1([0,\infty))$ and attains a (global) minimum at $t_* = 0$. In addition, it is monotonically increasing as

$$\frac{\partial g}{\partial t}(t) \;=\; \langle \nabla f(y_t), z - x_* \rangle \;=\; \frac{1}{t} \langle \nabla f(y_t) - \nabla f(x_*), y_t - x_* \rangle \;\geq\; 0,$$

where $y_t = x_* + t(z - x_*)$. The function g is also unbounded ($\lim_{t \to +\infty} g(t) = +\infty$) since otherwise, by [14, Theorem 32.1], it would be constant and this would contradict the fact that $z \notin X_*^f$. As a result, by the Mean Value Theorem, there exists a $t_z > 0$ such that $g(t_z) = f(y_{t_z}) = f(x_0)$.

It is not possible to have $t_z < 1$ because this would lead to $f(z) = g(1) > g(t_z) = f(x_0)$ which would imply $z \notin L_f(x_0)$. Since $t_z \geq 1$,

$$\operatorname{dist}(z, X_*^f) \;=\; \|z - x_*\| \;=\; \frac{1}{t_z} \|y_{t_z} - x_*\| \;\leq\; \|y_{t_z} - x_*\| \;=\; \operatorname{dist}(y_{t_z}, X_*^f).$$

Thus $\sup_{y \in L_f(x_0)} \operatorname{dist}(y, X_*^f) \leq \sup_{y \in f^{-1}(f(x_0))} \operatorname{dist}(y, X_*^f)$. ∎

4. WCC of direct search for a class of convex functions

The solution set for problem (1) is denoted by

$$X_*^f \;=\; \{x \in \mathbb{R}^n : \; x \text{ is a minimizer of } f\}.$$

In this paper we will always consider the case when X_*^f is non-empty.

We will analyze the worst case complexity of direct search when the objective function is smooth and convex under the assumption of a bound on the distance from all unsuccessful iterates to the solution set. We will denote by $\mathcal{S}$ and $\mathcal{U}$ the sets of indices corresponding to all successful and unsuccessful iterations, respectively.

Assumption 4.1. *There exists a positive constant R such that*

$$\sup_{j \in \mathcal{U}} \operatorname{dist}(x_j, X_*^f) \;\leq\; R.$$

We will discuss in Section 5 several scenarios under which this assumption is satisfied. We will also see in Section 6 that such an assumption seems necessary.

We will start by measuring the decrease obtained in the objective function until a given iteration as a function of the number of unsuccessful iterations occurred until then. Recall that $f_* = f(x_*)$ for some $x_* \in X_*^f$ and $\hat{p} = \frac{p}{\min(1,p-1)} > 2$ for $p > 1$.

Lemma 4.1. *Consider the application of Algorithm 3.1 when $\rho(t) = \mathcal{C}\, t^p$, $p > 1$, $\mathcal{C} > 0$, and D_k satisfies Assumption 3.1. Let Assumptions 3.2 and 4.1 also hold. Let $f \in \mathcal{F}_\nu^1(\Omega)$, where Ω is an open set containing $L_f(x_0)$, and X_*^f be non-empty.*

Let k_0 be the index of the first unsuccessful iteration (which must exist from Lemma 3.1). Then Algorithm 3.1 generates a sequence $\{x_k\}_{k \geq k_0}$ such that

$$(f(x_k) - f_*)^{\hat{p}-1} \;<\; \frac{R^{\hat{p}}}{\omega(k - k_0 - m - 1)}, \tag{6}$$

where

$$\omega \;=\; \omega_g^{\hat{p}} \beta_1^p \mathcal{C}, \quad \omega_g \;=\; \frac{2\, \mathrm{cm}_{\min}\, M^{\min(1,p-1)}}{\nu d_{\max} M + 2\mathcal{C} d_{\min}^{-1} M^{p-1}}, \tag{7}$$

and $m = m(k, k_0)$ is the number of unsuccessful iterations between k_0 and k.

Proof: Let $\{k_i\}_{i=0}^m$ represent the set of unsuccessful iterations which occur between iteration k_0, inclusively, and iteration k. Since all iterations between k_m

and k are successful and k_m is unsuccessful, we have that

$$\begin{aligned} f(x_k) &< f(x_{k-1}) - \mathcal{C}\alpha_{k-1}^p \\ &\vdots \\ &< f(x_{k_m+1}) - \mathcal{C}\sum_{j=k_m+1}^{k-1} \alpha_j^p \\ &\leq f(x_{k_m+1}) - \mathcal{C}(k-k_m-1)\alpha_{k_m+1}^p \\ &\leq f(x_{k_m}) - \beta_1^p \mathcal{C}(k-k_m-1)\alpha_{k_m}^p. \end{aligned}$$

Now, by Corollary 3.1,

$$f(x_k) < f(x_{k_m}) - (k-k_m-1)\omega\|\nabla f(x_{k_m})\|^{\hat{p}}. \tag{8}$$

By applying a similar argument, but now starting from x_{k_i}, $i = m, \ldots, 1$, we deduce that

$$f(x_{k_i}) < f(x_{k_{i-1}}) - (k_i - k_{i-1} - 1)\omega\|\nabla f(x_{k_{i-1}})\|^{\hat{p}}. \tag{9}$$

Denote $\Delta f_i = f(x_{k_i}) - f_*$, for $i = 0, \ldots, m$ and $\Delta f_{m+1} = f(x_k) - f_*$. Then, using the property stated in [10, Equation (2.1.7)] for $f \in \mathcal{F}_\nu^1(\mathbb{R}^n)$,

$$\begin{aligned} f_* &= f(x_*^i) \\ &\geq f(x_{k_i}) + \langle \nabla f(x_{k_i}), x_*^i - x_{k_i}\rangle + \frac{1}{2\nu}\|\nabla f(x_*^i) - \nabla f(x_{k_i})\|^2 \\ &\geq f(x_{k_i}) + \langle \nabla f(x_{k_i}), x_*^i - x_{k_i}\rangle, \end{aligned}$$

for some $x_*^i \in X_*^f$, $i = 0, \ldots, m$. Thus, using Assumption 4.1,

$$\begin{aligned} \Delta f_i &\leq \langle \nabla f(x_{k_i}), x_{k_i} - x_*^i\rangle \\ &\leq \|\nabla f(x_{k_i})\|\|x_{k_i} - x_*^i\| \\ &\leq R\|\nabla f(x_{k_i})\|, \quad i = 0, \ldots, m. \end{aligned} \tag{10}$$

By combining inequalities (8), (9), and (10) and setting here for simplicity $k_{m+1} = k$, we obtain, for $i = 1, \ldots, m, m+1$,

$$\Delta f_i \leq \Delta f_{i-1} - \frac{\omega}{R^{\hat{p}}}(k_i - k_{i-1} - 1)\Delta f_{i-1}^{\hat{p}} \leq \Delta f_{i-1}. \tag{11}$$

Hence, $\Delta f_{i-1}/\Delta f_i \geq 1$, $i = 1, \ldots, m, m+1$. Now we divide the first inequality in (11) by $\Delta f_i \Delta f_{i-1}$, then use $\hat{p} > 2$ and $\Delta f_{i-1} \geq \Delta f_{m+1}$, and later

which then implies

$$k_0 < \frac{f(x_0) - f(x_{k_0})}{\mathcal{C}\alpha_0^p} \leq \frac{f(x_0) - f^*}{\mathcal{C}\alpha_0^p}.$$

Note that for any $p > 1$, $\hat{p}$ is bigger than 2 and so $1 - \hat{p} < -1$. Hence for any given $\epsilon \in (0,1)$, $\epsilon^{1-\hat{p}} > 1$, and therefore one can establish that the number of iterations required to achieve the first unsuccessful one is bounded by

$$\left\lceil \frac{f(x_0) - f_*}{\mathcal{C}\alpha_0^p}\epsilon^{1-\hat{p}} \right\rceil.$$

We are finally ready to state the worst case complexity bound for Algorithm 3.1 when the objective function is convex.

Corollary 4.1. *Let all assumptions of Theorem 4.1 hold.*
To reduce the gradient below $\epsilon \in (0,1)$, Algorithm 3.1 takes at most

$$\mathcal{O}\left(\nu^{\hat{p}}\epsilon^{1-\hat{p}}\right) \tag{17}$$

iterations. When $p = 2$, this number is of $\mathcal{O}(\nu^2\epsilon^{-1})$.
The constant in $\mathcal{O}(\cdot)$ depends only on $d_{\min}$, $d_{\max}$, $\mathrm{cm}_{\min}$, $\mathcal{C}$, p, β_1, β_2, γ, α_0, and on the constant R of Assumption 4.1.

To count the corresponding number of function evaluations we need first to factor out the dependence of n in the above bound. We know from [15] that, in this bound, only the minimum cosine measure of the positive spanning sets depends explicitly on n. One also knows from the positive spanning set formed by the coordinate vectors and their negatives that such minimum cosine measure can be set greater than or equal to $1/\sqrt{n}$, and thus $1/\omega \leq \mathcal{O}(n^{\frac{\hat{p}}{2}})$, where ω is given in (7). On the other hand, each poll step when using such positive spanning sets costs at most $\mathcal{O}(n)$ function evaluations. One then assumes, for compatibility with the cost of such poll steps, that the search step, when non-empty, takes at most $\mathcal{O}(n)$ function evaluations.

Corollary 4.2. *Let all assumptions of Theorem 4.1 hold. Let $\mathrm{cm}_{\min}$ be at least a multiple of $1/\sqrt{n}$ and the number of function evaluations per iteration be at most a multiple of n.*
To reduce the gradient below $\epsilon \in (0,1)$, Algorithm 3.1 takes at most

$$\mathcal{O}\left(n^{\frac{\hat{p}+2}{2}}\nu^{\hat{p}}\epsilon^{1-\hat{p}}\right)$$

function evaluations. When $p = 2$, this number is of $\mathcal{O}(n^2\nu^2\epsilon^{-1})$.

ϵ, starting from k_0, Algorithm 3.1 takes at most $|\mathcal{U}_{\bar{k}}(k_0)|$ unsuccessful iterations, where

$$|\mathcal{U}_{\bar{k}}(k_0)| \ \leq \ \left\lceil \omega_1 |\mathcal{S}_{\bar{k}}(k_0)| + \omega_2 + \frac{1}{\min(p-1,1)} \log_{\beta_2}(\omega_g \epsilon) \right\rceil,$$

$$\omega_1 \ = \ -\log_{\beta_2}(\gamma), \quad \omega_2 \ = \ -\log_{\beta_2}(\alpha_{k_0}) + \log_{\beta_2}(\beta_1),$$

ω_g is given in (7), and $|\mathcal{S}_{\bar{k}}(k_0)|$ in the number of successful iterations between k_0 and $\bar{k}$.

Proof: From Corollary 3.1 and the definition of ω_g in (7), we have, for each unsuccessful iteration k, that

$$\|\nabla f(x_k)\| \ \leq \ \omega_g^{-1} \alpha_k^{\min(1,p-1)}.$$

As before, we can backtrack from any successful iteration to the nearest unsuccessful iteration and, due to the step size updating rules, we have the following inequality for any iteration after k_0

$$\alpha_k \ \geq \ \beta_1 (\omega_g \epsilon)^{\frac{1}{\min(1,p-1)}}, \quad k = k_0, k_0+1, \ldots, \bar{k}-1.$$

On the other hand, one knows that either $\alpha_k \leq \beta_2 \alpha_{k-1}$ or $\alpha_k \leq \gamma \alpha_{k-1}$. Hence, by induction,

$$\alpha_{\bar{k}-1} \ \leq \ \alpha_{k_0} \gamma^{|\mathcal{S}_{\bar{k}}(k_0)|} \beta_2^{|\mathcal{U}_{\bar{k}}(k_0)|}.$$

Now, since $\beta_2 < 1$, the function $\log_{\beta_2}(\cdot)$ is monotonically decreasing, and one obtains (the coefficient ω_1 is nonnegative due to $\gamma \geq 1$)

$$|\mathcal{U}_{\bar{k}}(k_0)| \ \leq \ \omega_1 |\mathcal{S}_{\bar{k}}(k_0)| + \omega_2 + \frac{1}{\min(p-1,1)} \log_{\beta_2}(\omega_g \epsilon).$$

■

Theorem 4.1 and 4.2 show that Algorithm 3.1 takes at most $\mathcal{O}(\epsilon^{1-\hat{p}})$ iterations after the first unsuccessful one to bring the norm of the gradient below $\epsilon \in (0,1)$. Thus the only part missing is to bound the number of successful iterations until the first unsuccessful one. As in [15], one can do this easily since, from the fact that k_0 is the first unsuccessful iteration,

$$f(x_{k_0}) \ < \ f(x_0) - \mathcal{C} \sum_{j=0}^{k_0-1} \alpha_j^p \ \leq \ f(x_0) - \mathcal{C} k_0 \alpha_0^p,$$

$\Delta f_{i-1}/\Delta f_i \geq 1$,

$$\begin{aligned} \frac{1}{\Delta f_i} \ &\geq \ \frac{1}{\Delta f_{i-1}} + \frac{\omega}{R^{\hat{p}}}(k_i - k_{i-1} - 1) \frac{\Delta f_{i-1}^{\hat{p}-1}}{\Delta f_i} \\ &\geq \ \frac{1}{\Delta f_{i-1}} + \frac{\omega \Delta f_{m+1}^{\hat{p}-2}}{R^{\hat{p}}}(k_i - k_{i-1} - 1) \frac{\Delta f_{i-1}}{\Delta f_i} \\ &\geq \ \frac{1}{\Delta f_{i-1}} + \frac{\omega \Delta f_{m+1}^{\hat{p}-2}}{R^{\hat{p}}}(k_i - k_{i-1} - 1). \end{aligned} \tag{12}$$

By summing the inequality (12) for $i = 1, \ldots, m, m+1$, we arrive at

$$\begin{aligned} \frac{1}{\Delta f_{m+1}} \ &\geq \ \frac{1}{\Delta f_0} + \frac{\omega \Delta f_{m+1}^{\hat{p}-2}}{R^{\hat{p}}}(k_{m+1} - k_0 - m - 1) \\ &= \ \frac{\omega \Delta f_{m+1}^{\hat{p}-2}}{R^{\hat{p}}}(k_{m+1} - k_0 - m - 1), \end{aligned}$$

or, equivalently,

$$\begin{aligned} (f(x_k) - f^*)^{\hat{p}-1} \ &= \ \Delta f_{m+1}^{\hat{p}-1} \\ &\leq \ \frac{R^{\hat{p}}}{\omega(k_{m+1} - k_0 - m - 1)} \\ &= \ \frac{R^{\hat{p}}}{\omega(k - k_0 - m - 1)}, \end{aligned}$$

as we wanted to prove. ■

In the following theorem by using the result of Lemma 4.1 we will derive an upper bound for the number of successful iterations after the first unsuccessful one needed to achieve a point for which the norm of the gradient is below a given threshold.

Theorem 4.1. *Consider the application of Algorithm 3.1 when $\rho(t) = \mathcal{C}\, t^p$, $p > 1$, $\mathcal{C} > 0$, and D_k satisfies Assumption 3.1. Let Assumptions 3.2 and 4.1 also hold. Let $f \in \mathcal{F}_\nu^1(\Omega)$, where Ω is an open set containing $L_f(x_0)$, and X_*^f be non-empty.*

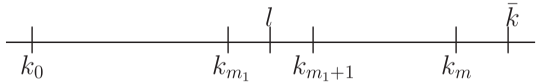
Let k_0 be the index of the first unsuccessful iteration (which must exist from Lemma 3.1). Given any $\epsilon \in (0,1)$, assume that $\|\nabla f(x_{k_0})\| > \epsilon$ and let $\bar{k}$ be the first iteration after k_0 such that $\|\nabla f(x_{\bar{k}})\| \leq \epsilon$. Then, to achieve $\|\nabla f(x_{\bar{k}})\| \leq$

ϵ, starting from k_0, Algorithm 3.1 takes at most $|\mathcal{S}_{\bar{k}}(k_0)|$ successful iterations, where

$$|\mathcal{S}_{\bar{k}}(k_0)| \ \leq \ \left\lceil 2\frac{R}{\omega}\epsilon^{1-\hat{p}}+1 \right\rceil \tag{13}$$

and ω is given in (7).

Proof: Let l, with $k_0 < l < \bar{k}$, be the index of a successful iteration occurring before k, m be number of unsuccessful iterations between k_0 and $\bar{k}$, m_1 be the number of unsuccessful iterations between k_0 and l, and $k_1, k_2, \ldots, k_m$ be the sequence of unsuccessful iterations between k_0 and $\bar{k}$.



Let us assume first that there are unsuccessful iterations between l and $\bar{k}$ (like in the figure above). Exactly as in the derivation of inequalities (8)–(9), applying also Corollary 3.1 and the step size updating rules, we have

$$f(x_{\bar{k}}) \ < \ f(x_{k_m}) - (\bar{k}-k_m-1)\omega\|\nabla f(x_{k_m})\|^{\hat{p}}$$

and,

$$\begin{aligned} f(x_{k_i}) \ &< \ f(x_{k_{i-1}}) - (k_i-k_{i-1}-1)\omega\|\nabla f(x_{k_{i-1}})\|^{\hat{p}}, \quad m_1+2 \leq i \leq m, \\ f(x_{k_{m_1+1}}) \ &< \ f(x_l) - (k_{m_1+1}-l)\omega\|\nabla f(x_{k_{m_1}})\|^{\hat{p}}. \end{aligned}$$

Summing up these inequalities and considering $\|\nabla f(x_k)\| > \epsilon$ for $k < \bar{k}$ lead us to

$$f(x_l) \ > \ f(x_{\bar{k}}) + (\bar{k}-l-m+m_1)\omega\epsilon^{\hat{p}}.$$

If there are no unsuccessful iterations between l and $\bar{k}$, $m = m_1$ and this inequality is also true by a similar argument. On the other hand, by Lemma 4.1

$$(f(x_l)-f_*)^{\hat{p}-1} \ \leq \ \frac{R^{\hat{p}}}{\omega(l-k_0-m_1-1)}.$$

So, in conclusion

$$\begin{aligned} (\bar{k}-l-m+m_1)\omega\epsilon^{\hat{p}} \ &\leq \ (\bar{k}-l-m+m_1)\omega\epsilon^{\hat{p}} + f(x_{\bar{k}}) - f_* \\ &\leq \ f(x_l) - f_* \\ &\leq \ \left(\frac{R^{\hat{p}}}{\omega(l-k_0-m_1-1)}\right)^{\frac{1}{\hat{p}-1}}. \end{aligned} \tag{14}$$

Now we choose l such that the number of successful iterations after l is at most one times higher than the number of successful iterations until l. To explicitly describe l we divide the number of successful iterations into two parts $\frac{1}{2}(\bar{k}-k_0-m-1)$, then add the number m_1 of unsuccessful iterations until the middle point, and finally shift by k_0. Hence l is given by

$$l \ = \ \left\lfloor \frac{\bar{k}-k_0-m-1}{2} \right\rfloor + k_0 + m_1 + 1.$$

With such a choice of l, the number κ of successful iterations between k_0 and l is

$$\kappa \ = \ l-k_0-m_1-1$$

and a simple argument shows that

$$\kappa \ = \ l-k_0-m_1-1 \ \leq \ \bar{k}-l-m+m_1 \ \leq \ \kappa+1, \tag{15}$$

as expected.

Now, from (14),

$$\begin{aligned} (\omega\kappa)^{\frac{\hat{p}}{\hat{p}-1}} \ &\leq \ \omega(\bar{k}-l-m+m_1)[\omega(l-k_0-m_1-1)]^{\frac{1}{\hat{p}-1}} \\ &\leq \ R^{\frac{\hat{p}}{\hat{p}-1}}\epsilon^{-\hat{p}}, \end{aligned}$$

and

$$\kappa \ \leq \ \frac{R}{\omega}\epsilon^{1-\hat{p}}. \tag{16}$$

But due to equation (15), $2\kappa+1$ is bigger than the number of successful iterations between k_0 and k,

$$\begin{aligned} 2\kappa+1 \ &= \ \kappa+1+\kappa \\ &\geq \ (\bar{k}-l-m+m_1)+(l-k_0-m_1-1) \\ &= \ \bar{k}-k_0-m-1, \end{aligned}$$

which finishes the proof. ∎

Following [15, Theorem 3.2] one can also guarantee that the number of unsuccessful iterations is of the same order as the number of successful ones. The proof is given for sake of clearness and completeness.

Theorem 4.2. *Let all assumptions of Theorem 4.1 hold.*

Let k_0 be the index of the first unsuccessful iteration (which must exist from Lemma 3.1). Given any $\epsilon \in (0,1)$, assume that $\|\nabla f(x_{k_0})\| > \epsilon$ and let $\bar{k}$ be the first iteration after k_0 such that $\|\nabla f(x_{\bar{k}})\| \leq \epsilon$. Then, to achieve $\|\nabla f(x_{\bar{k}})\| \leq$