Operations Research Letters 45 (2017) 353–361

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

An indicator for the switch from derivative-free to derivative-based optimization

S. Gratton [a], N. Soualmi [b], L.N. Vicente [c,*]

[a] ENSEEIHT, INPT, rue Charles Camichel, B.P. 7122 31071, Toulouse Cedex 7, France
[b] CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 01, France
[c] CMUC, Department of Mathematics, University of Coimbra, 3001-501 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 2 February 2016
Received in revised form 2 May 2017
Accepted 7 May 2017
Available online 17 May 2017

Keywords:
Derivative-free optimization
Derivative-based optimization
Indicators
Direct-search methods
Gradient methods
Complexity and global rates

ABSTRACT

In some optimization problems found in applications, the derivatives of the objective function can be computed or approximated but at an expensive cost, and it is desirable to know when to use derivative-free methods (such as direct search, for instance) or derivative-based methods (such as gradient or quasi-Newton methods). Derivative-free methods may achieve a steady initial progress for some problems, but after some advance they may also become slower or even stagnate due to the lack of derivatives. It is thus of interest to provide a way to appropriately switch from a derivative-free method to a derivative-based one. In this paper, we develop a family of indicators for such a switch based on the decrease properties of both classes of methods (typically used when deriving worst case complexity bounds).

Published by Elsevier B.V.

1. Introduction and basic concepts

The calculation of functions involved in optimization problems appearing in computational sciences and engineering is frequently based on numerical simulation. The smoothness of the function and the access to whatever form of derivatives vary considerably across applications. While there are problems of totally black-box type where only function values can be computed in a certain (sometimes unknown) feasible region, there are other more structured problems of continuously differentiable type where both function values and gradients can be computed, an example of special interest to us being the acoustic full-waveform inverse problem in Earth imaging [11]. In such problems, the calculation of the gradient may come at a cost higher than the one for the function value, and such difference may depend on the dimension of the problem and the accuracy required.

Although Derivative-Free Optimization provides now a theory [1] to understand models and families of directions used in the various classes of methods as well as their convergence properties, a question that to our knowledge has never been addressed is when should one switch from a derivative-free method to a derivative-based one, when the gradient can be computed (or possibly approximated by finite differences). Such a general question can be posed in many different ways depending on the methods under consideration, their costs per iteration, the computational budget available, and the final accuracy desired for a solution. Other issues like non-smoothness or local vs global optimization may also play a relevant role when analyzing such an issue.

In this paper, we try to make a first contribution to the topic by considering a continuously differentiable setting where the gradient of the objective function $f : \mathbb{R}^n \to \mathbb{R}$ can be computed (or possibly approximated by finite differences), and the minimization of the function is unconstrained. The gradient of f will be considered Lipschitz continuous in $\mathbb{R}^n$ (or in a level set corresponding to an initial iterate), with constant $L_{\nabla f} > 0$. Our idea to develop an indicator for the switch under consideration will be to form appropriate ratios of lower bounds for the decreases attained in successful iterations (involving the derivative-free method and what would be expected for the derivative-based one). These lower bounds will be the ones used when deriving worst case complexity bounds (WCC) or global rates for such methods, and we will try to take advantage of the mismatches that appear in such bounds with or without using derivatives. Our ultimate goal is to detect a switch when not enough progress is being achieved compared to the one that could be done if derivatives were available, letting the derivative-free method continue otherwise (meaning to do not switch).

To illustrate and test our ideas, we will consider the following simple direct-search method which imposes a sufficient decrease condition based on a quadratic forcing function. We consider an

* Corresponding author.
E-mail addresses: serge.gratton@enseeiht.fr (S. Gratton), nacer.soualmi@cerfacs.fr (N. Soualmi), lnv@mat.uc.pt (L.N. Vicente).

http://dx.doi.org/10.1016/j.orl.2017.05.003
0167-6377/Published by Elsevier B.V.

iteration uniquely defined by a poll step which evaluates the objective function using a positive spanning set (PSS), i.e., a set of non-zero vectors that spans $\mathbb{R}^n$ with non-negative coefficients.

Algorithm 1.1. Direct-search method (polling)

Initialization: Choose a PSS D, an initial point x_0, and an initial step size $\alpha_0 > 0$. The constants $0 < \beta < 1 \leq \gamma$ are specified. Set $k = 0$.

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

2. Update iterate and step size: 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 $\alpha_{k+1} = \beta\alpha_k$. Increment k by one and go to Step 1.

It is well known that such an algorithm is well defined (for functions with Lipschitz continuous gradients) in the sense that a successful iteration is found in a finite number of step-size reductions [7]. In fact, it is possible to prove [7] that if the iteration k is unsuccessful, then

$$\|\nabla f(x_k)\| \leq \mathrm{cm}(D)^{-1}\left(L_{\nabla f}\frac{\max_{d\in D}\|d\|}{2} + \frac{1}{2\min_{d\in D}\|d\|}\right)\alpha_k, \quad (1)$$

where $\mathrm{cm}(D)$ is the cosine measure of the PSS D, defined as

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

The cosine measure of a PSS is always positive. For instance, the PSS $D_\oplus$ formed by the coordinate vectors and their negatives is such that $\mathrm{cm}(D_\oplus) = 1/\sqrt{n}$, and so is $QD_\oplus$ where Q is an orthogonal matrix. Given an $\epsilon \in (0, 1)$, it is known that such an algorithm takes at most $\mathcal{O}(n\epsilon^{-2})$ iterations and $\mathcal{O}(n^2\epsilon^{-2})$ function evaluations to drive the norm of the gradient below ϵ (see [10]).[1] The dependence of these WCC bounds on ϵ reduces to ϵ^{-1} if the function is convex and to $-\log(\epsilon)$ if the function is strongly convex (see [2]). These bounds depend quadratically on the Lipschitz constant $L_{\nabla f}$.

For the sake of simplicity, we take the gradient method with backtracking as our derivative-based method.

Algorithm 1.2. Gradient method (backtracking)

Initialization: Choose initial point x_0. Let $c \in (0, 1)$ and $b > 0$ be specified. Set $k = 0$.

1. Backtrack: Let α_k be the first scalar in $b, b/2, b/4, \ldots$ such that

$$f(x_k - \alpha_k\nabla f(x_k)) \leq f(x_k) - c\alpha_k\|\nabla f(x_k)\|^2. \quad (2)$$

2. Update iterate: Compute $x_{k+1} = x_k - \alpha_k\nabla f(x_k)$. Increment k by one and go to Step 1.

It is known that Algorithm 1.2 is well defined in the sense that it is always possible to find α_k of the form given in the algorithm such that (2) is satisfied (see, e.g., [9]). Moreover, each iteration of Algorithm 1.2 satisfies

$$f(x_k) - f(x_{k+1}) \geq C\|\nabla f(x_k)\|^2, \quad (3)$$

with

$$C = c\max\left(\frac{1-c}{L_{\nabla f}}, b\right). \quad (4)$$

[1] The notation $\mathcal{O}(A)$ will mean a scalar times A, where the scalar does not depend on the iteration counter of the method under analysis (thus depending only on the problem or on algorithmic constants).

It is also well known (see [8]) that the WCC effort for the gradient method (to reduce the norm of the gradient below ϵ) is of $\mathcal{O}(\epsilon^{-2})$ in general, reducing to $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(-\log(\epsilon))$ in the convex and strongly convex cases, respectively. Note that these bounds depend linearly on $L_{\nabla f}$.

The remaining of this paper is organized in three sections. In Section 2, we will describe our main idea to develop a family of indicators for the switch from derivative-free to derivative-based iterations using direct search and the gradient method as motivation. Two concrete indicators are then proposed in Section 3 and their numerical performance (using Algorithms 1.1 and 1.2) is reported. Finally, in Section 4, we will further discuss the scope of our approach.

2. Elements for the indicators

At each iteration of a gradient-based method, one typically has

$$f(x_k) - f(x_{k+1}) \geq \frac{G}{L_{\nabla f}}\|\nabla f(x_k)\|^2, \quad (5)$$

where $G > 0$ is a fixed constant independent of f or of the iteration counter. This is the case for the gradient method with line search satisfying both Wolfe conditions (sufficient decrease condition (2) and curvature condition), or for the gradient method with backtracking line search imposing only the sufficient decrease condition (2); (see Algorithm 1.2 and (3)–(4)).

At each successful iteration of a derivative-free method based on sufficient decrease, one typically has

$$f(x_k) - f(x_{k+1}) \geq D_1 t_k^2,$$

where $D_1 > 0$ is a fixed algorithmic parameter independent of f and of the iteration counter. The step-size parameter t_k represents the trust-region radius δ_k in derivative-free trust-region methods or the step size α_k in direct-search methods (see Algorithm 1.1). If ℓ is an iteration where the step-size parameter t_ℓ is reduced (for either class of methods), one has

$$\|\nabla f(x_\ell)\| \leq D_2(n)L_{\nabla f}\, t_\ell, \quad (6)$$

where $D_2(n) > 0$ is a fixed constant independent of f and of the iteration counter (but typically dependent on n); see (1) for the direct-search case and [3] for the trust-region one.

Let k be a given successful iteration and r_k the last iteration before k where the step size has been reduced. Let C_k be the set of indices corresponding to successful iterations between r_k and k where some approximation to the gradient is known.

The decrease produced by a gradient-based method would have been at least

$$f(r_k) - f(x_{k+1}) \geq \sum_{j\in C_k}\frac{G}{L_{\nabla f}}\|\nabla f(x_j)\|^2. \quad (7)$$

On the other hand, the decrease produced by a derivative-free method would have been at least

$$f(r_k) - f(x_{k+1}) \geq \sum_{j\in C_k} D_1 t_j^2 \geq \frac{D_1|C_k|}{D_2(n)^2 L_{\nabla f}^2}\|\nabla f(x_{r_k})\|^2. \quad (8)$$

Establishing a ratio between the decreases in (8) with those in (7), yields two quantities

$$\frac{D_1 L_{\nabla f}\sum_{j\in C_k} t_j^2}{G\sum_{j\in C_k}\|\nabla f(x_j)\|^2} \quad \text{and} \quad \frac{D_1|C_k|\|\nabla f(x_{r_k})\|^2}{GD_2(n)^2 L_{\nabla f}\sum_{j\in C_k}\|\nabla f(x_j)\|^2}.$$

This motivates the introduction of the following two indicators

$$I_k^1 = \frac{\sum_{j\in C_k} t_j^2}{\sum_{j\in C_k}\|g_j\|^2} \quad \text{and} \quad I_k^2 = \frac{|C_k|\|g_{r_k}\|^2}{\sum_{j\in C_k}\|g_j\|^2} \quad (9)$$