Operations Research Letters 44 (2016) 790–795

Contents lists available at ScienceDirect

Operations Research Letters

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

The value function approach to convergence analysis in composite optimization

Edouard Pauwels

IRIT-UPS, 118 route de Narbonne, 31062 Toulouse, France

ARTICLE INFO

Article history:
Received 6 April 2016
Received in revised form
6 October 2016
Accepted 6 October 2016
Available online 14 October 2016

Keywords:
Composite optimization
Gauss–Newton method
KL inequality
Value function
Convergence

ABSTRACT

This works aims at understanding further convergence properties of first order local search methods with complex geometries. We focus on the composite optimization model which unifies within a simple formalism many problems of this type. We provide a general convergence analysis of the composite Gauss–Newton method as introduced in Burke and Ferris (1995) (studied further in Chong and Wang, 2002; Chong and Ng, 2007; Lewis and Wright, 2015) under tameness assumptions (an extension of semi-algebraicity). Tameness is a very general condition satisfied by virtually all problems solved in practice. The analysis is based on recent progresses in understanding convergence properties of sequential convex programming methods through the value function as introduced in Bolte and Pauwels (2016).

1. Introduction

In composite optimization, convergence of Gauss–Newton methods is a question that has attracted a lot of research efforts in the past decades. Let us mention a few milestones: criticality of accumulation points was proved in [10], convergence under sharpness assumption around accumulation points is given in [11], and extensions to weaker regularity conditions are described in [13,12]. Asymptotic behavior under prox-regularity and identification under partial smoothness is investigated in [18]. These results attest to the difficulty of this undertaking. Although the composite model is strongly structured and Gauss–Newton method is explicitly designed to take advantage of it, convergence of iterates always relies on strong local growth conditions around accumulation points. These are often difficult to check in advance for general problems due to the complexity of the optimization model. To our knowledge, a simple and flexible global convergence analysis is still lacking for these methods.

Departing from existing approaches to address such complex geometries, we rely on tameness assumptions. In the nonsmooth nonconvex world, this assumption allows to use a powerful geometric property, the so-called nonsmooth Kurdyka–Łojasiewicz (KL) inequality, which holds true for many classes of functions [19,17,6,7]. We require problem data to be definable, a generalization of the property of being semi-algebraic [25,14]. This rules out non favorable pathological situations such as wild oscillations (e.g. fractals). This framework is general enough to model the vast majority of functions that can be handled numerically with a classical computer, while providing a sufficient condition for KL inequality to hold [7]. For a smoother understanding, the reader non familiar with tame geometry may replace "definable" by "semi-algebraic". Recall that an object is said to be real semi-algebraic if it can be defined as "the solution set of one of several systems of polynomial equalities and inequalities".

The use of KL inequality in nonconvex optimization provided significant advances in understanding convergence of first order methods [1–4,6,9]. However, the application of these techniques in complex geometric settings, such as composite optimization, remains an important challenge. A recent breakthrough has been made in [8], which describes a general convergence analysis of Sequential Quadratic Programming methods [15,5,16]. This is an important example of complex geometric structures with challenging convergence analysis. To overcome the difficulty of dealing with problems with complex geometries in this context, [8] has introduced a new methodology based on the so-called value function.

We propose a general convergence guaranty for a variant of the composite Gauss–Newton method [10,11]. The main idea consists in viewing Gauss–Newton method along the lines of [8] through the value function approach. An important improvement brought to [8] is the integration of a general backtracking search in the analysis. This allows to deal with smooth functions whose gradients are merely *locally* Lipschitz continuous. This flexibility is extremely important from a practical point of view and requires

E-mail address: edouard.pauwels@irit.fr.

http://dx.doi.org/10.1016/j.orl.2016.10.003
0167-6377/

Composite Gauss–Newton

Choose $x_0 \in D$, $\mu_0 > 0$, $\tau > 1$ and iterate

Step 1. Set $\mu_k = \mu_0$ and compute the candidate iterate:

$$\tilde{x}_{k+1} \leftarrow \operatorname*{argmin}_{y\in D} g(F(x_k) + \nabla F(x_k)(y - x_k)) + \frac{\mu_k}{2}\|y - x_k\|^2$$

Step 2. While $g\left(F(\tilde{x}_{k+1})\right) > g(F(x_k) + \nabla F(x_k)(\tilde{x}_{k+1} - x_k)) + \frac{\mu_k}{2}\|\tilde{x}_{k+1} - x_k\|^2$

$$\mu_k \leftarrow \tau\mu_k$$
$$\tilde{x}_{k+1} \leftarrow \operatorname*{argmin}_{y\in D} g(F(x_k) + \nabla F(x_k)(y - x_k)) + \frac{\mu_k}{2}\|y - x_k\|^2$$

Step 3. Update

$$x_{k+1} \leftarrow \tilde{x}_{k+1} \tag{1}$$

non trivial extensions (see [21] for works in this direction). To the best of our knowledge this result is new, it relies on easily verifiable assumptions and it is flexible enough to encompass many problems encountered in practice. In addition, we emphasize that it provides a simple and intuitive way to highlight the potential of the value function approach designed in [8].

In Section 2, we describe the problem of interest, the main assumptions and the algorithm. We also state our main convergence result. We introduce notations, important definitions and results from nonsmooth analysis and geometry in Section 3. The value function and its most important properties are described in Section 4. Section 5 contains the proof of the main result.

2. Problem setting and main result

We consider the composite optimization problem.

$$\min_{x\in D\subset\mathbb{R}^n} g(F(x)). \tag{2}$$

Our main standing assumption is the following.

Assumption 1. $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^2$ and $g\colon \mathbb{R}^m \to \mathbb{R}$ is convex and finite valued. $D \subset \mathbb{R}^n$ is convex and closed. F, g and D are definable in the same o-minimal structure on the field of real numbers (fixed throughout the text).

Note that Assumption 1 ensures that g is locally Lipschitz continuous [22, Theorem 10.4]. For any $i = 1, 2, \ldots, m$, we use the notation f_i for the $\mathcal{C}^2$ function that corresponds to coordinate i of F. We denote by $\nabla F(x)$ the Jacobian matrix of F at x:

$$\nabla F(x) = \left[\frac{\partial f_i}{\partial x_j}(x)\right] \in \mathbb{R}^{m\times n}.$$

We will analyze the numerical scheme (1) which is a backtracking variant of the composite Gauss–Newton descent method [10,11,13,12,18].

Remark 1. The dynamical feature of the step-size parameter μ_k is akin to a backtracking procedure. Indeed, Assumption 1 ensures that F is locally smooth and g is locally Lipschitz continuous. However the smoothness and Lipschitz continuity moduli may be unknown and not be valid in a global sense. They have to be estimated in an online fashion to prevent unwanted divergent behaviors.

The next lemma shows that the algorithm is well defined and the sequence of objective values is nonincreasing (the proof is given in Section 4). The next theorem is our main result and the proof is given in Section 5.

Lemma 2.1. *For each k, the while loop stops after a finite number of iterations and we have*

$$g\left(F(x_{k+1})\right) \le g(F(x_k) + \nabla F(x_k)(x_{k+1} - x_k)) + \frac{\mu_k}{2}\|x_{k+1} - x_k\|^2,$$

and $\{g(F(x_k))\}_{k\in\mathbb{N}}$ is a nonincreasing sequence.

Theorem 2.2. *Under Assumption 1, we have the alternatives when $k \to +\infty$.*

- $\|x_k\| \to +\infty$.
- *x_k converges to a critical point of Problem (2), the sequence $\|x_{k+1} - x_k\|$ is summable, $\{\mu_k\}_{k\in\mathbb{N}}$ is bounded.*

Remark 2. In the alternatives of Theorem 2.2, the unbounded case is due to a lack of coercivity rather than a bad adjustment of the local model through μ_k. Indeed, if we suppose that x_0 is chosen such that the set $D \cap \{x \in \mathbb{R}^n;\ g(F(x)) \le g(F(x_0))\}$ is compact, Lemma 2.1 ensures that the divergent option cannot hold and the sequence converges. This phenomenon was guessed in [3] and also appeared in [8]. Accounting for the dynamical feature of μ_k in our analysis is a contribution of this work.

3. Notations and preliminary results

3.1. Notations

The symbol ∂ refers to the limiting subdifferential. The notion of a critical point is that of a limiting critical point: zero is in the limiting subdifferential, a necessary condition of optimality (nonsmooth Fermat's rule). We refer, for instance, the reader to [23, Chapter 8] for further details on the subject.

An o-minimal structure on the field of real numbers is a structured collection of definable subsets of finite dimensional Euclidean spaces. It is required to satisfy some of the properties of semi-algebraic sets. Semi-algebraic sets form an o-minimal structure but there are many extensions. An introduction to the subject can be found in [14] and a survey of relevant results is available in [24]. In Assumption 1, we have fixed an o-minimal structure. Definable sets are subsets of Euclidean spaces which belong to it and a definable function is a function which graph is definable.

The normal cone to D at $x \in D$ is denoted by $N_D(x)$ and the indicator function of D is denoted by i_D (whose value is constantly 0 on D, $+\infty$ otherwise). $\|\cdot\|$ denotes the Euclidean norm (which is semi-algebraic). Being given a function $f\colon \mathbb{R}^p \to \mathbb{R}$, real numbers a and b, we set $[a < f < b] = \{x \in \mathbb{R}^n :\ a < f(x) < b\}$.