

On high-order multilevel optimization strategies

Henri Calandra, Serge Gratton, Elisa Riccietti, Xavier Vasseur

▶ **To cite this version:**

Henri Calandra, Serge Gratton, Elisa Riccietti, Xavier Vasseur. On high-order multilevel optimization strategies. SIAM Journal on Optimization, 2020, 31 (1), pp.307-330. 10.1137/19M1255355 . hal-02943229

HAL Id: hal-02943229
https://hal.science/hal-02943229

Submitted on 18 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On high-order multilevel optimization strategies*

Henri Calandra,†Serge Gratton*, Elisa Riccietti*, Xavier Vasseur‡

April 10, 2019

Abstract

We propose a new family of multilevel methods for unconstrained minimization. The resulting strategies are multilevel extensions of high-order optimization methods based on q-order Taylor models (with $q \geq 1$) that have been recently proposed in the literature. The use of high-order models, while decreasing the worst-case complexity bound, makes these methods computationally more expensive. Hence, to counteract this effect, we propose a multilevel strategy that exploits a hierarchy of problems of decreasing dimension, still approximating the original one, to reduce the global cost of the step computation. A theoretical analysis of the family of methods is proposed. Specifically, local and global convergence results are proved and a complexity bound to reach first order stationary points is also derived. A multilevel version of the well known adaptive method based on cubic regularization (ARC, corresponding to $q = 2$ in our setting) has been implemented. Numerical experiments clearly highlight the relevance of the new multilevel approach leading to considerable computational savings in terms of floating point operations compared to the classical one-level strategy.

arXiv:1904.04692v1 [math.NA] 9 Apr 2019

1 Introduction

We propose a new family of high-order multilevel optimization methods for unconstrained minimization. Exploiting ideas stemming from multilevel methods allows us to reduce the cost of the step computation, which represents the major cost per iteration of the standard single level procedures. We have been mainly inspired by two driving ideas: the use of high-order models in optimization as introduced in [2], and the multilevel recursive strategy proposed in [13].

When solving unconstrained minimization problems, quadratic models are widely used. These are usually regularized by a quadratic term. For example, trust-region methods have been widely studied and used to globalize Newton-like iterations [12, 22]. Lately in the literature, a different option has received a growing attention: the use of a cubic overestimator of the objective function as a regularization technique for the computation of the step from one iterate to the next, giving rise to quadratic models with cubic regularization. This idea first appeared in [14] and then was reconsidered in [21], where the authors proved that the method has a better worst-case complexity bound compared to standard trust-region methods. Later, in [8, 10], an adaptive variant of the method has been proposed, based on a dynamical choice of the regularization parameters and on an approximate solution of the subproblems. The resulting method is known as adaptive method based on cubic regularization (ARC) and is shown to preserve the attractive global complexity bound established in [21]. In recent years the method has attracted further interest, see for example [9, 23, 26].

In recent publications also methods of higher order start to gain interest, see for example [2, 24]. In [2] in particular, it has been observed that the good complexity bound of ARC can be made even lower, if one is willing to use higher-order derivatives. In specific applications this computation is indeed feasible, for example when considering partially separable functions [11]. The authors in [2] present a family of methods that generalizes ARC, and that uses high-order regularized models. Specifically, they are based on models of order $q \geq 1$, regularized by a term of order $q + 1$. ARC belongs to this family and corresponds to the choice $q = 2$. The authors in [2] propose a unifying framework to describe the theoretical properties of the methods in this class. It is proved that the method based on the q-th order model requires at most $O\left(\epsilon^{-\frac{q+1}{q}}\right)$ function evaluations to find a first-order critical point, where ϵ denotes the absolute accuracy level.

However, the use of higher-models come along with higher computational costs. The main cost per iteration of the methods described in [2] is represented by the step computation through the model minimization. This cost is proportional to the dimension of the problem, it can therefore be significant for large-scale problems. For second-order models this issue has been faced for example in [13], where the authors exploit ideas coming from multigrid [15] to reduce the cost of the minimization. Indeed, the idea of making use of more grids to solve a large-scale problem has been extended also to optimization, see for example [13, 16, 17, 18, 19, 20, 25]. These methods share with classical multigrid methods the idea of exploiting a hierarchy of problems (in this case a sequence of nonlinear functions)

*INPT-IRIT, University of Toulouse and ENSEEIHT, 2 Rue Camichel, BP 7122, F-31071 Toulouse Cedex 7, France (serge.gratton@enseeiht.fr, elisa.riccietti@enseeiht.fr)

*This work was funded by TOTAL.

†TOTAL, Centre Scientifique et Technique Jean Féger, avenue de Larribau F-64000 Pau, France (henri.calandra@total.com)

‡ISAE-SUPAERO, University of Toulouse, 10, avenue Edouard Belin, BP 54032, F-31055 Toulouse Cedex 4, France (xavier.vasseur@isae-supaero.fr).

defined on lower dimensional spaces, approximating the original objective function f. The simplified expressions of the objective function are used to build models that are cheaper to minimize, and are used to define the step. Specifically, in [13], the authors present an extension of classical multigrid methods for nonlinear optimization problems, [4, 5] or [6, Ch. 3], to a class of multilevel trust-region based optimization algorithms.

Our contributions Inspired by the ideas presented in [2, 13], we propose a family of multilevel optimization methods using high-order regularized models that generalizes the methods proposed in both papers. The aim is to decrease the computational cost of the methods in [2], extending the ideas in [13] to higher-order models. We also develop a theoretical analysis for the resulting family of methods. The main theoretical results are provided in Theorems 1, 2 and 3, respectively. In these theorems we successively prove the global convergence property of the methods, then evaluate a worst-case complexity bound to reach a first-order critical point and finally provide local convergence rates. The global convergence analysis generalizes the results in [2] and appears as much simpler than that in [13]. Moreover we establish local convergence results towards second-order stationary points, that are not present neither in [2] nor in [13]. These results not only generalize those in [27], that are valid only for $q = 2$, but also apply to the one level methods in [2]. From a practical point of view, we implemented the method of the family corresponding to $q = 2$. This represents a multilevel version of ARC method.

To the best of our knowledge, this is the first time that multilevel optimization strategies, based on models of generic order $q \geq 1$, are proposed, and that a unifying framework is introduced to study their convergence. Moreover, multilevel versions of ARC have never been analysed or tested numerically before.

The manuscript is organized as follows. In Section 2, we briefly introduce the family of optimization methods using high-order regularized models considered in [2]. Section 3 and Section 4 represent our main contribution. We introduce in Section 3 the multilevel extensions of the methods presented in Section 2, and we provide a theoretical analysis in Section 4. Specifically, we focus on global convergence in Section 4.1, worst-case complexity in Section 4.2 and local convergence in Section 4.3. In Section 5 we then present results related to numerical experiments performed with the multilevel method corresponding to $q = 2$. Finally, conclusions are drawn in Section 6.

2 High-order iterative optimization methods

Let $q \geq 1$ be an integer. Let us consider a minimization problem of the form:

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

with $f : \mathbb{R}^n \to \mathbb{R}$ a bounded below and q-times continuously differentiable function, called the objective function.

Classical iterative optimization methods for unconstrained minimization are based on the use of a model to approximate the objective function at each iteration. In this section, we describe the iterative optimization methods using high-order models presented in [2].

2.1 Model definition and step acceptance

At each iteration k, given the current iterate x_k, the objective function is approximated by the Taylor series $T_{q,k}$ of $f(x_k + s)$ (with $s \in \mathbb{R}^n$) truncated at order q. The Taylor model of order q denoted as $m_{q,k}$ is then defined as:

$$m_{q,k}(x_k, s) = T_{q,k}(x_k, s). \tag{2}$$

A step s_k is then found minimizing (possibly approximately) the regularized model

$$T_{q,k}(x_k, s) + \frac{\lambda_k}{q+1}\|s\|^{q+1}, \tag{3}$$

where λ_k is a positive value called regularization parameter. The step s_k is used to define a trial point i.e. $x_{k+1} = x_k + s_k$. At each iteration, it has to be decided whether to accept the step or not. This decision is based on the accordance between the decrease in the function and in the model. More precisely, at each iteration both the decrease achieved in the model, that we call *predicted reduction*, $pred = m_{q,k}(x_k) - m_{q,k}(x_k, s_k)$, and that achieved in the objective function, that we call *actual reduction*, $ared = f(x_k) - f(x_k + s_k)$, are computed. The step acceptance is then based on the ratio:

$$\rho_k = \frac{ared}{pred} = \frac{f(x_k) - f(x_k + s_k)}{m_{q,k}(x_k) - m_{q,k}(x_k, s_k)}. \tag{4}$$

If the model is accurate, ρ_k will be close to one. Then, the step s_k is accepted if ρ_k is larger than or equal to a chosen threshold $\eta_1 \in (0, 1)$ and is rejected otherwise. In the first case, the step is said to be *successful*, and otherwise the step is *unsuccessful*.

After the step acceptance, the regularization parameter is updated for the next iteration. The update is still based on the ratio (4). If the step is successful, the regularization parameter is decreased,

otherwise it is increased. The whole procedure is stopped when a minimizer of f is reached. Usually, the stopping criterion is based on the norm of the gradient, i.e. given an absolute accuracy level $\epsilon > 0$ the iterations are stopped as soon as $\|\nabla_x f(x_k)\| < \epsilon$. The whole procedure is sketched in Algorithm 1.

Algorithm 1 AR**q**$(x_0, \lambda_0, \epsilon)$ (Adaptive Regularization method of order q)

Given $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_2 \leq \gamma_1 < 1 < \gamma_3$, $\lambda_{\min} > 0$.
Input: $x_0 \in \mathbb{R}^n$, $\lambda_0 > \lambda_{\min}$, $\epsilon > 0$.
$k = 0$
while $\|\nabla_x f(x_k)\| > \epsilon$ **do**
  • **Initialization:** Define the model $m_{q,k}$ as in (2).
  • **Model minimization:** Find a step s_k that sufficiently reduces the model.
  • **Acceptance of the trial point:** Compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{m_{q,k}(x_k) - m_{q,k}(x_k, s_k)}$.
  if $\rho_k \geq \eta_1$ **then**
    $x_{k+1} = x_k + s_k$
  else
    $x_{k+1} = x_k$.
  end if
  • **Regularization parameter update:**
  if $\rho_k \geq \eta_1$ **then**

$$\lambda_{k+1} = \begin{cases} \max\{\lambda_{\min}, \gamma_2 \lambda_k\}, & \text{if } \rho_k \geq \eta_2, \\ \max\{\lambda_{\min}, \gamma_1 \lambda_k\}, & \text{if } \rho_k < \eta_2, \end{cases}$$

  else
    $\lambda_{k+1} = \gamma_3 \lambda_k$.
  end if
  $k = k + 1$
end while

2.2 Minimization of the model

The main computational work per iteration in this kind of methods is represented by the minimization of the regularized model (3). This is the most expensive task, and the cost naturally depends on the dimension of the problem. However, from the convergence theory of such methods, it is well known that it is not necessary to minimize the model exactly to get a globally convergent method.

A well-known possibility is to minimize the model until the Cauchy decrease is achieved, i.e. until a fraction of the decrease provided by the Cauchy step (the step that minimizes the model in the direction of the negative gradient) is obtained. In [2] the authors consider a different stopping criterion for the inner iterations, the one originally proposed in [8, 10], which has the advantage of allowing for simpler convergence proofs. The inner iterations are stopped as soon as the norm of the gradient of the regularized model becomes lower or equal than a multiple of the power q of the norm of the step s_k:

$$\|\nabla_s m_{q,k}(x_k, s_k) + \lambda_k \|s_k\|^{q-1} s_k\| \leq \theta \|s_k\|^q, \tag{5}$$

for a chosen constant $\theta > 0$.

For very large-scale problems however, even an approximate minimization of (3) may be really costly. Then, in the next section we propose multilevel variants of the procedures, that rely on simplified models of the objective function, cheaper to optimize, allowing to reduce the global cost of the optimization procedure.

3 Multilevel optimization methods

We describe the multilevel extension of the family of methods presented in Section 2. The procedures are inspired by the multilevel trust-region approach presented in [13], where only second-order models with quadratic regularization have been considered. Here, we generalize this approach by allowing also higher-order models, i.e. $q > 2$.

3.1 Preliminaries and notations

In standard optimization methods the minimization of (3) represents the major cost per iteration, which crucially depends on the dimension n of the problem. When n is large, the solution cost is therefore often significant. We want to reduce this cost by exploiting the knowledge of alternative simplified expressions of the objective function. More specifically, we assume that we know a collection of functions $\{f_l\}_{l=1}^{l_{\max}}$ such that each f_l is a q-times continuously differentiable function from $\mathbb{R}^{n_l} \to \mathbb{R}$ and $f^{l_{\max}}(x) = f(x)$ for all $x \in \mathbb{R}^n$. We will also assume that, for each $l = 2, \ldots, l_{\max}$, f_l is more costly to minimize than f_{l-1}. This is the typical scenario when the problem arises from the discretization of an infinite dimensional problem and f_l represent increasingly finer discretizations. In this case,

$n_l \geq n_{l-1}$ for all l, but of course this is not the only possible application. As we do not assume the hierarchy to come from a discretization process, we do not use the terminology typically used in the field of multigrid methods. We will then use 'levels' rather than 'grids'.

The methods we propose are recursive procedures, so it suffices to describe the two-level case. Then, for sake of simplicity, from now on, we will assume that we have just two approximations to our objective f at disposal. This amounts to consider $l_{\max} = 2$.

For ease of notation, we will denote by $f^h : \mathcal{D}^h \subseteq \mathbb{R}^{n_h} \to \mathbb{R}$ the approximation at the highest level ($f^h(x) = f^{l_{\max}}(x)$ in the notation previously used) and by $f^H : \mathcal{D}^H \subseteq \mathbb{R}^{n_H} \to \mathbb{R}$ the other approximation available, that is cheaper to optimize. The quantities on the highest level will be denoted by a superscript h, whereas the quantities on the lower level will be denoted by a superscript H. Let x_k^h denote the k-th iteration at the highest level. In the following $\|\cdot\|$ will denote the Euclidean norm. We will use the same notation for all the spaces we will consider, the space on which the norm is defined will be clear by the context.

Remark 1. *To deal with high order derivatives, and to properly handle the concept of coherence between lower and higher level model, we will need to use a tensor notation, that we introduce here for convenience of the reader, see [3]. We first consider a tensor of order three, and then extend the definition to a tensor of order $i \in \mathbb{N}$.*

Definition 1. *Let $T \in \mathbb{R}^{n\times n\times n}$, and $u, v, w \in \mathbb{R}^n$. Then $T(u, v, w) \in \mathbb{R}$, $T(u, v) \in \mathbb{R}^n$ and*

$$T(u,v,w) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} T(i,j,k)u(i)v(j)w(k),$$
$$T(v,w)(i) = \sum_{j=1}^{n}\sum_{k=1}^{n} T(i,j,k)v(j)w(k), \quad i = 1,\dots,n.$$

Definition 2. *Let $i \in \mathbb{N}$ and $T \in \mathbb{R}^{n^i}$, and $u_1, \dots, u_i \in \mathbb{R}^n$. Then $T(u_1, \dots, u_i) \in \mathbb{R}$, $T(u_1, \dots, u_{i-1}) \in \mathbb{R}^n$ and*

$$T(u_1,\dots,u_i) = \sum_{j_1=1}^{n}\cdots\sum_{j_i=1}^{n} T(j_1,\dots,j_i)u_1(j_1)\dots u_i(j_i),$$
$$T(u_1,\dots,u_{i-1})(j_1) = \sum_{j_2=1}^{n}\cdots\sum_{j_i=1}^{n} T(j_1,\dots,j_i)u_i(j_2),\dots u_{i-1}(j_i), \quad j_1 = 1,\dots,n.$$

3.2 Construction of the lower level model

The main idea is to use f^H to construct, in the neighbourhood of the current iterate, an alternative model $m_{q,k}^H$ to the Taylor model $m_{q,k}^h$ in (2) for $f^h = f$ [13]. The alternative model $m_{q,k}^H$ should be cheaper to optimize than $m_{q,k}^h$, and will be used, whenever suitable, to define the step. Of course, for f^H to be useful at all in minimizing f^h, there should be some relation between the variables of these two functions. We henceforth assume the following.

Assumption 1. *Let us assume that there exist two full-rank linear operators $R : \mathbb{R}^{n_h} \to \mathbb{R}^{n_H}$ and $P : \mathbb{R}^{n_H} \to \mathbb{R}^{n_h}$ such that $P = \alpha R^T$, for a fixed scalar $\alpha > 0$. Let us assume also that it exists $\kappa_R > 0$ such that $\max\{\|R\|, \|P\|\} \leq \kappa_R$, where $\|\cdot\|$ denotes the matrix norm induced by the Euclidean norm at the fine level.*

In the following, we can assume $\alpha = 1$, without loss of generality, as the problem can be easily scaled to handle the case $\alpha \neq 1$.

At each iteration k at highest level we set $x_{0,k}^H = R\, x_k^h$, i.e. the initial iterate at the lower level is set as the projection of the current iterate, and we define the lower level model $m_{q,k}^H$ as a modification of the coarse function f^H. Given q, f^H is modified adding q correction terms, to enforce the following relation:

$$\nabla_s^i m_{q,k}^H(x_{0,k}^H, \underbrace{s^H,\dots,s^H}_{i \text{ times}}) = [\mathcal{R}(\nabla_x^i f^h(x_k^h))](\underbrace{s^H,\dots,s^H}_{i \text{ times}}), \quad i = 1,\dots,q, \tag{6}$$

where $\mathcal{R}(\nabla_x^i f^h(x_k^h))$ is such that for all $i = 1,\dots,q$ and $s_1^H,\dots,s_i^H \in \mathbb{R}^{n_H}$

$$[\mathcal{R}(\nabla_x^i f^h(x_k^h))](s_1^H,\dots,s_i^H) := \nabla_x^i f^h(x_k^h, Ps_1^H,\dots,Ps_i^H), \tag{7}$$
$$\langle[\mathcal{R}(\nabla_x^i f^h(x_k^h))](s_1^H,\dots,s_{i-1}^H), s_i^H\rangle := \langle\nabla_x^i f^h(x_k^h, Ps_1^H,\dots,Ps_{i-1}^H), Ps_i^H\rangle, \tag{8}$$

where $\nabla_x^i f^h, \nabla_s^i m_{q,k}^H$ denote the i-th order tensor of f^h and $m_{q,k}^H$ respectively, $\langle\cdot,\cdot\rangle$ denotes the scalar product, and for generic $g : \mathbb{R}^n \to \mathbb{R}$ and $s_1,\dots,s_i \in \mathbb{R}^n$, $\nabla^i g(x, s_1,\dots,s_i)$ is the same as $\nabla^i g(x)(s_1,\dots,s_i)$, which is given in Definition 2.

For instance, if $q = 2$, relation (6) simply becomes:

$$\nabla_s m_{q,k}^H(x_{0,k}^H)^T s^H = (R\,\nabla_x f^h(x_k^h))^T s^H, \quad (s^H)^T\nabla_x^2 m_{q,k}^H(x_{0,k}^H)s^H = (s^H)^T R\,\nabla_x^2 f^h(x_k^h)\,Ps^H.$$

Relation (6) crucially ensures that the behaviours of f^h and $m^H_{q,k}$ are coherent up to order q in a neighbourhood of x^h_k and $x^H_{0,k}$. To achieve (6), we define $m^H_{q,k}$ as

$$m^H_{q,k}(x^H_{0,k}, s^H) = f^H(x^H_{0,k} + s^H) + \sum_{i=1}^{q} \frac{1}{i!}[\mathcal{R}(\nabla^i_x f^h(x^h_k)) - \nabla^i_x f^H(x^H_{0,k})](\underbrace{s^H, \ldots, s^H}_{i \text{ times}}), \tag{9}$$

with $\mathcal{R}(\nabla^i_x f^h(x^h_k))$ defined in (7)-(8). When $q = 2$ this is simply:

$$\begin{aligned} m^H_{2,k}(x^H_{0,k}, s^H) =& f^H(x^H_{0,k} + s^H) + (R\nabla_x f^h(x^h_k) - \nabla_x f^H(x^H_{0,k}))^T s^H \\ &+ \frac{1}{2}(s^H)^T (R\nabla^2_x f^h(x^h_k)P - \nabla^2_x f^H(x^H_{0,k}))s^H. \end{aligned}$$

3.3 Step computation and step acceptance

At each generic iteration k of our method, a step s^h_k has to be computed to define the new iterate. Then, one has the choice between the Taylor model (2) and a lower level model (9).

Obviously, it is not always possible to use the lower level model. For example, it may happen that $\nabla_x f^h(x^h_k)$ lies in the nullspace of R and thus that $R\nabla_x f^h(x^h_k)$ is zero while $\nabla_x f^h(x^h_k)$ is not. In this case, the current iterate appears to be first-order critical for $m^H_{q,k}$ while it is not for f^h. Using the model $m^H_{q,k}$ is hence potentially useful only if $\|\nabla_s m^H_{q,k}(x^H_{0,k})\| = \|R\nabla_x f^h(x^h_k)\|$ is large enough compared to $\|\nabla_x f^h(x^h_k)\|$ [13]. We therefore restrict the use of the model $m^H_{q,k}$ to iterations where

$$\|R\nabla_x f^h(x^h_k)\| \geq \kappa_H \|\nabla_x f^h(x^h_k)\| \quad \text{and} \quad \|R\nabla_x f^h(x^h_k)\| > \epsilon_H, \tag{10}$$

for some constant $\kappa_H \in (0, \min\{1, \|R\|\})$ and where $\epsilon_H \in (0, 1)$ is a measure of the first-order criticality for $m^H_{q,k}$ that is judged sufficient at level H [13]. Note that, given $\nabla_x f^h(x^h_k)$ and R, this condition is easy to check before even attempting to compute a step at a lower level.

If the Taylor model is chosen, then we just compute a step as in standard methods, minimizing (possibly approximately) the corresponding regularized model (3). If the lower level model is chosen, we then minimize the following regularized model:

$$m^H_{q,k}(x^H_{0,k}, s^H) + \frac{\lambda_k}{q+1}\|s^H\|^{q+1} \tag{11}$$

(possibly approximately) and obtain a point $x^H_{*,k}$ such that (if the minimization is successful) the value of the regularized model has been reduced, and a step $s^H_k = x^H_{*,k} - x^H_{0,k}$ (note that the iteration indices always refer to the highest level, we are not indexing the iterations on the lower level for the minimization of the lower level model). This step has to be prolongated back on the fine level, i.e. we define $s^h_k = Ps^H_k$.

Then, $m^h_{q,k}$ will be defined as:

$$m^h_{q,k}(x^h_k, s^h_k) = \begin{cases} T^h_{q,k}(x^h_k, s^h_k) & \text{(Taylor model)}, \\ m^H_{q,k}(Rx^h_k, s^H_k),\ s^h_k = Ps^H_k & \text{(lower level model)}. \end{cases} \tag{12}$$

In both cases, after the step is found, we have to decide whether to accept it or not. The step acceptance is based on the ratio:

$$\rho_k = \frac{f^h(x^h_k) - f^h(x^h_k + s^h_k)}{m^h_{q,k}(x^h_k) - m^h_{q,k}(x^h_k, s^h_k)},$$

where we remind that, from (12), the denominator is defined as:

$$m^h_{q,k}(x^h_k) - m^h_{q,k}(x^h_k, s^h_k) = \begin{cases} T^h_{q,k}(x^h_k) - T^h_{q,k}(x^h_k + s^h_k), & \text{(Taylor model)}, \\ m^H_{q,k}(Rx^h_k) - m^H_{q,k}(Rx^h_k, s^H_k), & \text{(lower level model)}. \end{cases} \tag{13}$$

As in the standard form of the methods, the step is accepted if it provides a sufficient decrease in the function, i.e. if given $\eta_1 > 0$, $\rho_k \geq \eta_1$. The regularization parameter is also updated as in Algorithm 1. We sketch the whole procedure in Algorithm 2.

Some comments are necessary to explain Step 6 in Algorithm 2. The generic framework sketched in Algorithm 2 comprises different possible methods. Specifically, one of the flexible features (inherited by the method in [13]) is that, to ensure convergence, the minimization at lower levels can be stopped after the first successful iteration, as we will see in the next section. This therefore opens the possibility to consider both fixed form recursion patterns and free form ones. A free form pattern is obtained when Algorithm 2 is run carrying the minimization at each level out, until the norm of the gradient becomes small enough. The actual recursion pattern is then uniquely determined by the progress of minimization at each level and may be difficult to forecast. By contrast, the fixed form recursion patterns are obtained by specifying a maximum number of successful iterations at each level, a technique directly inspired from the definitions of V- and W-cycles in multigrid algorithms [15].

Algorithm 2 MAR**q**$(l, f^l, x_0^l, \lambda_0^l, \epsilon^l)$ (Multilevel Adaptive Regularization method of order q)

Input: $l \in \mathbb{N}$ (index of the current level, $1 \leq l \leq l_{\max}$, $l_{\max}$ being the highest level), $f^l : \mathbb{R}^{n_l} \to \mathbb{R}$ function to be optimized ($f^{l_{\max}} = f$), $x_0^l \in \mathbb{R}^{n_l}$, $\lambda_0^l > \lambda_{\min}$, $\epsilon^l > 0$.
Given $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_2 \leq \gamma_1 < 1 < \gamma_3$, $\lambda_{\min} > 0$.
R_l denotes the restriction operator from level l to $l-1$, P_l the prolongation operator from level $l-1$ to l.
$k = 0$
while $\|\nabla_x f^l(x_k^l)\| > \epsilon^l$ **do**
• **Model choice:** If $l > 1$ compute $R_l \nabla_x f^l(x_k^l)$ and check (10). If $l = 1$ or (10) fails, go to Step 7. Otherwise, choose to go to Step 7 or to Step 8.
• **Taylor step computation:** Define $m_{q,k}^l(x_k^l, s^l) = T_{q,k}^l(x_k^l, s^l)$, the Taylor series of $f^l(x_k^l + s^l)$ truncated at order q. Find a step s_k^l that sufficiently reduces $m_{q,k}^l(x_k^l, s^l) + \frac{\lambda_k^l}{q+1}\|s^l\|^{q+1}$. Go to Step 9.
• **Recursive step computation:** Define

$$m_{q,k}^{l-1}(R_l\ x_k^l, s^{l-1}) = f^{l-1}(R_l\ x_k^l, s^{l-1}) + \sum_{i=1}^{q} \frac{1}{i!}[\mathcal{R}(\nabla_x^i f^l(x_k^l)) - \nabla_x^i f^{l-1}(R_l\ x_k^l)](\underbrace{s^{l-1}, \ldots, s^{l-1}}_{i \text{ times}}).$$

Choose ϵ^{l-1} and call MAR**q**$(l-1, m_{q,k}^{l-1}, R_l\ x_k^l, \lambda_k^l, \epsilon^{l-1})$ yielding an approximate solution $x_{*,k}^{l-1}$ of the minimization of $m_{q,k}^{l-1}$. Define $s_k^l = P_l\ (x_{*,k}^{l-1} - R_l\ x_k^l)$ and $m_{q,k}^l(x_k^l, s^l) = m_{q,k}^{l-1}(R_l\ x_k^l, s^{l-1})$ for all $s^l = P s^{l-1}$.
• **Acceptance of the trial point:** Compute $\rho_k^l = \dfrac{f^l(x_k^l) - f^l(x_k^l + s_k^l)}{m_{q,k}^l(x_k^l) - m_{q,k}^l(x_k^l, s_k^l)}$.
if $\rho_k^l \geq \eta_1$ **then**
  $x_{k+1}^l = x_k^l + s_k^l$
else
  $x_{k+1}^l = x_k^l$.
end if
• **Regularization parameter update:**
if $\rho_k^l \geq \eta_1$ **then**

$$\lambda_{k+1}^l = \begin{cases} \max\{\lambda_{\min}, \gamma_2 \lambda_k^l\}, & \text{if } \rho_k^l \geq \eta_2, \\ \max\{\lambda_{\min}, \gamma_1 \lambda_k^l\}, & \text{if } \rho_k^l < \eta_2 \end{cases}$$

else
  $\lambda_{k+1}^l = \gamma_3 \lambda_k^l$.
end if
$k = k + 1$
end while

4 Convergence theory

In this section, we provide a theoretical analysis of the proposed family of multilevel methods. Inspired by the convergence theory reported in [2], we prove global convergence of the proposed methods to first-order critical points and we provide a worst-case complexity bound to reach such a point, generalizing the theory proposed in [2, 13]. At the same time the proposed analysis also appears as simpler than that in [13], since the regularization parameter λ_k is directly updated, rather than the trust-region radius, and since we use the stopping criterion (5) as in [2]. The use of this criterion allows for simpler convergence proofs and enables us to concentrate on the multilevel algorithm, that is the main contribution of the paper. Moreover, we also propose local convergence results, which also apply to the methods in [2], and that extend those in [27] to higher-order models.

Note that, as the methods are recursive, we can restrict the analysis to the two-level case. For the analysis we need the following regularity assumptions as in [2].

Assumption 2. *Let f^h and f^H be q-times continuously differentiable and bounded below functions. Let us assume that the q-th derivative tensors of f^h and f^H are Lipschitz continuous, i.e. that there exist constants L_h, L_H such that*

$$\|\nabla^q f^h(x) - \nabla^q f^h(y)\|_T \leq (q-1)!\, L_h \|x-y\| \quad \text{for all} \quad x, y \in \mathcal{D}^h,$$
$$\|\nabla^q f^H(x) - \nabla^q f^H(y)\|_T \leq (q-1)!\, L_H \|x-y\| \quad \text{for all} \quad x, y \in \mathcal{D}^H,$$

where $\|\cdot\|_T$ is the tensor norm recursively induced by the Euclidean norm on the space of q-th order tensors, which for a tensor H of order q is given by

$$\|H\|_T \stackrel{\text{def}}{=} \max_{\|u_1\|=\cdots=\|u_q\|=1} |H(u_1, \ldots, u_q)|$$

where the action of H on $(u_1, \ldots, u_q)$ is given in Definition (2).

We remind three useful relations, following from Taylor's theorem, see for example relations (2.3) and (2.4) in [2].

Lemma 1. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a q-times continuously differentiable function with Lipschitz continuous q-th order tensor, with L the corresponding Lipschitz constant. Given its Taylor series $T_q(x, s)$ truncated at order q, it holds:*

$$g(x+s) = T_q(x,s) + \frac{1}{(q-1)!}\int_0^1 (1-\xi)^{q-1}[\nabla^q g(x+\xi s) - \nabla^q g(x)](\overbrace{s, \ldots, s}^{q \text{ times}})\, d\xi, \tag{14}$$

$$|g(x+s) - T_q(x,s)| \leq \frac{L}{q}\|s\|^{q+1}, \tag{15}$$

$$\|\nabla g(x+s) - \nabla_s T_q(x,s)\| \leq L\|s\|^q. \tag{16}$$

4.1 Global convergence

In this section we prove the global convergence property of the method. Our analysis proceeds in three steps. First, we bound the quantity $|1-\rho_k|$ to prove that λ_k must be bounded above. Then, we relate the norm of the step and the norm of the gradient. Finally, we use these two ingredients to conclude proving that the norm of the gradient goes to zero.

4.1.1 Upper bound for the regularization parameter λ_k

At iteration k we either minimize (decrease) the regularized Taylor model (3), or the regularized lower level model (11). Consequently, it respectively holds:

$$T^h_{q,k}(x^h_k) - T^h_{q,k}(x^h_k, s^h_k) \geq \frac{\lambda_k}{q+1}\|s^h_k\|^{q+1}, \tag{17a}$$

$$m^H_{q,k}(x^H_{0,k}) - m^H_{q,k}(x^H_{0,k}, s^H_k) \geq \frac{\lambda_k}{q+1}\|s^H_k\|^{q+1}. \tag{17b}$$

In both cases, the minimization process is stopped as soon as the stopping condition

$$\|\nabla_s T^h_{q,k}(x^h_k, s^h_k) + \lambda_k\|s^h_k\|^{q-1}s^h_k\| \leq \theta\|s^h_k\|^q, \quad \text{or}$$
$$\|\nabla_s m^H_{q,k}(x^H_{0,k}, s^H_k) + \lambda_k\|s^H_k\|^{q-1}s^H_k\| \leq \theta\|s^H_k\|^q, \tag{18}$$

for $\theta > 0$ is satisfied, respectively. In both cases we are sure that it will exist a point that satisfies (18), as when the level is selected, a standard one-level optimization method is used, and the analysis in [2] applies.

Let us consider the quantity

$$|1-\rho_k| = \left|1 - \frac{f^h(x^h_k) - f^h(x^h_k + s^h_k)}{m^h_{q,k}(x^h_k) - m^h_{q,k}(x^h_k, s^h_k)}\right|, \tag{19}$$

with the denominator defined in (13). If at step k the Taylor model is chosen, from relation (15) applied to f^h and (17a) we obtain the inequality:

$$|1-\rho_k| = \left| \frac{f^h(x_k^h + s_k^h) - T_{q,k}^h(x_k^h, s_k^h)}{m_{q,k}^h(x_k^h) - m_{q,k}^h(x_k^h, s_k^h)} \right| \leq \frac{L_h(q+1)}{\lambda_k q}.$$

If the lower level model is used, we have

$$|1-\rho_k| = \left| \frac{m_{q,k}^H(x_{0,k}^H) - m_{q,k}^H(x_{0,k}^H, s_k^H) - (f^h(x_k^h) - f^h(x_k^h + s_k^h))}{m_{q,k}^H(x_{0,k}^H) - m_{q,k}^H(x_{0,k}^H, s_k^H)} \right|.$$

Let us consider the numerator in this expression. From relations (9) and (14) applied to f^H, using its Taylor series $T_{q,k}^H$, it follows

$$\begin{aligned}
& m_{q,k}^H(x_{0,k}^H) - m_{q,k}^H(x_{0,k}^H, s_k^H) \overset{(9)}{=} \\
& T_{q,k}^H(x_{0,k}^H, s_k^H) - f^H(x_{0,k}^H + s_k^H) - \sum_{i=1}^{q} \frac{1}{i!} \left[\mathcal{R}(\nabla_x^i f^h(x_k^h))\right] (\overbrace{s_k^H, \ldots, s_k^H}^{i \text{ times}}), \\
\overset{(14)}{=} & -\frac{1}{(q-1)!} \int_0^1 (1-\xi)^{q-1} \left[\nabla^q f^H(x_{0,k}^H + \xi s_k^H) - \nabla^q f^H(x_{0,k}^H)\right] (\overbrace{s_k^H, \ldots, s_k^H}^{q \text{ times}}) d\xi \\
& - \sum_{i=1}^{q} \frac{1}{i!} \left[\mathcal{R}(\nabla_x^i f^h(x_k^h))\right] (\overbrace{s_k^H, \ldots, s_k^H}^{i \text{ times}}).
\end{aligned} \tag{20}$$

Similarly the relation (14) applied to f^h yields

$$\begin{aligned}
f^h(x_k^h) - f^h(x_k^h + s_k^h) &= f^h(x_k^h) - T_{q,k}^h(x_k^h, s_k^h) \\
&- \frac{1}{(q-1)!} \int_0^1 (1-\xi)^{q-1} [\nabla^q f^h(x_k^h + \xi s_k^h) - \nabla^q f^h(x_k^h)] (\overbrace{s_k^h, \ldots, s_k^h}^{q \text{ times}}) d\xi.
\end{aligned} \tag{21}$$

From relation (6) we can rewrite $f^h(x_k^h) - T_{q,k}^h(x_k^h, s_k^h)$ as:

$$\begin{aligned}
f^h(x_k^h) - T_{q,k}^h(x_k^h, s_k^h) &= -\sum_{i=1}^{q} \frac{1}{i!} (\nabla_x^i f^h)(x_k^h, \overbrace{Ps_k^H, \ldots, Ps_k^H}^{i \text{ times}}) \\
&= -\sum_{i=1}^{q} \frac{1}{i!} \left[\mathcal{R}(\nabla_x^i f^h(x_k^h))\right] (\overbrace{s_k^H, \ldots, s_k^H}^{i \text{ times}}).
\end{aligned}$$

Then, subtracting (21) from (20), we obtain

$$\begin{aligned}
& m_{q,k}^H(x_{0,k}^H) - m_{q,k}^H(x_{0,k}^H, s_k^H) - (f^h(x_k^h) - f^h(x_k^h + s_k^h)) = \\
& - \frac{1}{(q-1)!} \int_0^1 (1-\xi)^{q-1} [\nabla^q f^H(x_{0,k}^H + \xi s_k^H) - \nabla^q f^H(x_{0,k}^H)] (\overbrace{s_k^H, \ldots, s_k^H}^{q \text{ times}}) \, d\xi \\
& + \frac{1}{(q-1)!} \int_0^1 (1-\xi)^{q-1} [\nabla^q f^h(x_k^h + \xi s_k^h) - \nabla^q f^h(x_{0,k}^h)] (\overbrace{s_k^h, \ldots, s_k^h}^{q \text{ times}}) \, d\xi.
\end{aligned}$$

Using Assumption 2, we obtain:

$$\begin{aligned}
& |m_{q,k}^H(x_{0,k}^H) - m_{q,k}^H(x_{0,k}^H, s_k^H) - (f^h(x_k^h) - f^h(x_k^h + s_k^h))| \\
& \leq \frac{1}{(q-1)!} \int_0^1 (1-\xi)^{q-1} |[\nabla^q f^H(x_{0,k}^H + \xi s_k^H) - \nabla^q f^H(x_{0,k}^H)] (\overbrace{s_k^H, \ldots, s_k^H}^{q \text{ times}})| \, d\xi \\
& + \frac{1}{(q-1)!} \int_0^1 (1-\xi)^{q-1} |[\nabla^q f^h(x_k^h + \xi s_k^h) - \nabla^q f^h(x_k^h)] (\overbrace{s_k^h, \ldots, s_k^h}^{q \text{ times}})| \, d\xi \\
& \leq \frac{1}{q!} \|s_k^H\|^q \max_{\xi \in [0,1]} \|\nabla^q f^H(x_k^H + \xi s_k^H) - \nabla^q f^H(x_k^H)\|_T \\
& + \frac{1}{q!} \|s_k^h\|^q \max_{\xi \in [0,1]} \|\nabla^q f^h(x_k^h + \xi s_k^h) - \nabla^q f^h(x_k^h)\|_T \leq \frac{1}{q} \left(L_H + L_h \kappa_R^{q+1}\right) \|s_k^H\|^{q+1}.
\end{aligned}$$

From relation (17b) we finally obtain:

$$|1-\rho_k| \leq \frac{(q+1)\left(L_H + L_h \kappa_R^{q+1}\right)}{q \lambda_k}.$$

Then, in both cases (when either a Taylor model or a lower level model is used), it exists a strictly positive constant K such that the following relation holds:

$$|1-\rho_k| \leq \frac{K}{\lambda_k}, \qquad K = \begin{cases} \dfrac{(q+1)L_h}{q} & \text{(Taylor model)}, \\ \dfrac{(q+1)\left(L_H + L_h \kappa_R^{q+1}\right)}{q} & \text{(lower level model)}. \end{cases} \tag{22}$$

Using this last relation and the updating rule of the regularization parameter, we deduce that λ_k must be bounded above. Indeed, in case of unsuccessful iterations, λ_k is increased. If λ_k is increased, the ratio appearing in the right hand side of (22) is progressively decreased, until it becomes smaller than $1-\eta_1$. In this case, $\rho_k > \eta_1$, so a successful step is taken and λ_k is decreased. Hence λ_k cannot be greater than

$$\lambda_{\max} = \frac{K}{1-\eta_1}. \tag{23}$$

4.1.2 Relating the steplength to the norm of the gradient

Our next step is to show that the steplength cannot be arbitrarily small, compared to the norm of the gradient of the objective function. If the Taylor model is used, from [2, Lemma 2.3] it follows:

$$\|\nabla_x f^h(x_k^h + s_k^h)\| \leq (L_h + \theta + \lambda_k)\|s_k^h\|^q := K_1 \|s_k^h\|^q. \tag{24}$$

If the lower level model is chosen, we have:

$$\begin{aligned} \|R\nabla_x f^h(x_k^h + s_k^h)\| \leq & \left\| R\left[\nabla_x f^h(x_k^h + s_k^h) - \nabla_s T_{q,k}^h(x_k^h, s_k^h)\right] \right\| \\ & + \|R\nabla_s T_{q,k}^h(x_k^h, s_k^h) - \nabla_s m_{q,k}^H(x_{0,k}^H, s_k^H)\| \\ & + \|\nabla_s m_{q,k}^H(x_{0,k}^H, s_k^H) + \lambda_k \|s_k^H\|^{q-1} s_k^H\| + \lambda_k \|s_k^H\|^q. \end{aligned}$$

By (16), the first term can be bounded by $\kappa_R L_h \|s_k^h\|^q$. Considering that $s_k^h = P s_k^H$ and $\|P\| \leq \kappa_R$, we obtain the upper bound $\kappa_R^2 L_h \|s_k^H\|^q$. Regarding the second term, taking into account that from relations $s_k^h = P s_k^H$, $R = P^T$, and (8), for all $p^H \in \mathbb{R}^{n_H}$ it holds:

$$\begin{aligned} \langle [\mathcal{R}(\nabla_x^i f^h(x_k^h))](\underbrace{s_k^H, \ldots, s_k^H}_{i-1 \text{ times}}), p^H \rangle &= \langle \nabla_x^i f^h(x_k^h, \underbrace{P s_k^H, \ldots, P s_k^H}_{i-1 \text{ times}}), P p^H \rangle \\ &= \langle R[\nabla_x^i f^h(x_k^h, \underbrace{P s_k^H, \ldots, P s_k^H}_{i-1 \text{ times}})], p^H \rangle, \end{aligned}$$

we can write

$$\begin{aligned} R\nabla_s T_{q,k}^h(x_k^h, P s_k^H) &= \sum_{i=1}^{q} \frac{1}{(i-1)!} R\nabla_x^i f^h(x_k^h)(\underbrace{P s_k^H, \ldots, P s_k^H}_{i-1 \text{ times}}) \\ &= \sum_{i=1}^{q} \frac{1}{(i-1)!} \left[\mathcal{R}(\nabla_x^i f^h(x_k^h))\right] (\underbrace{s_k^H, \ldots, s_k^H}_{i-1 \text{ times}}). \end{aligned}$$

Then, from

$$\begin{aligned} \nabla_s m_{q,k}^H(x_{0,k}^H, s_k^H) = & \nabla_x f^H(x_{0,k}^H + s_k^H) \\ & + \sum_{i=1}^{q} \frac{1}{(i-1)!} \left[\mathcal{R}(\nabla_x^i f^h(x_k^h)) - \nabla_x^i f^H(x_{0,k}^H)\right] (\underbrace{s_k^H, \ldots, s_k^H}_{i-1 \text{ times}}), \end{aligned} \tag{25}$$

we obtain

$$\|R\nabla_s T_{q,k}^h(x_k^h, s_k^h) - \nabla_s m_{q,k}^H(x_{0,k}^H, s_k^H)\| = \left\| \nabla_x f^H(x_{0,k}^H + s_k^H) - \nabla_s T_{q,k}^H(x_{0,k}^H, s_k^H) \right\|,$$

which represents the Taylor remainder for the approximation of $\nabla_x f^H$ by $\nabla_s T_{q,k}^H$. Therefore, by relation (16), this quantity can be bounded above by $L_H \|s_k^H\|^q$. The third term, from (18), is less than $\theta \|s_k^H\|^q$. Then, since $\lambda_k \leq \lambda_{\max}$, we finally obtain

$$\|R\nabla_x f^h(x_k^h + s_k^h)\| \leq \left(\kappa_R^2 L_h + L_H + \theta + \lambda_{\max}\right) \|s_k^H\|^q := K_2 \|s_k^H\|^q. \tag{26}$$

4.1.3 Proof of global convergence

Let us consider the sequence of successful iterations ($\rho_k \geq \eta_1$). They are divided into two groups, $K_{s,f}$ the successful iterations at which the fine model has been employed and $K_{s,l}$ the ones at which the lower level model has been employed. Let us define k_1 the index of the first successful iteration. We

remind that at successful iterations $\rho_k \geq \eta_1$. Due to the updating rule of the regularization parameter in Algorithm 2 we have $\lambda_k \geq \lambda_{\min}$. Hence from relations (13), (17), (24) and (26), (10) it follows that:

$$
\begin{aligned}
f^h(x^h_{k_1}) - \liminf_{k\to\infty} f^h(x^h_k) &\geq \sum_{k succ} f^h(x^h_k) - f^h(x^h_k + s^h_k) \\
&\overset{(13)}{\geq} \eta_1 \sum_{K_{s,l}} (m^H_{q,k}(x^H_{0,k}) - m^H_{q,k}(x^H_{0,k}, s^H_k)) + \eta_1 \sum_{K_{s,f}} (T^h_{q,k}(x^h_k) - T^h_{q,k}(x^h_k, s^h_k)) \\
&\overset{(17)}{\geq} \frac{\eta_1 \lambda_k}{q+1} \left(\sum_{K_{s,l}} \|s^H_k\|^{q+1} + \sum_{K_{s,f}} \|s^h_k\|^{q+1} \right) \\
&\overset{(24)+(26)}{\geq} \frac{\eta_1 \lambda_{\min}}{q+1} \left(\frac{1}{K_2^{\frac{q+1}{q}}} \sum_{K_{s,l}} \|R\nabla_x f^h(x^h_k + s^h_k)\|^{\frac{q+1}{q}} + \frac{1}{K_1^{\frac{q+1}{q}}} \sum_{K_{s,f}} \|\nabla_x f^h(x^h_k + s^h_k)\|^{\frac{q+1}{q}} \right) \\
&\overset{(10)}{\geq} \frac{\eta_1 \lambda_{\min}}{q+1} \left(\frac{1}{K_2^{\frac{q+1}{q}}} \sum_{K_{s,l}} \kappa_H^{\frac{q+1}{q}} \|\nabla_x f^h(x^h_k + s^h_k)\|^{\frac{q+1}{q}} + \frac{1}{K_1^{\frac{q+1}{q}}} \sum_{K_{s,f}} \|\nabla_x f^h(x^h_k + s^h_k)\|^{\frac{q+1}{q}} \right). \qquad (27)
\end{aligned}
$$

Hence we conclude that $\sum_{K_{s,f} \cup K_{s,l}} \|\nabla_x f^h(x^h_k + s^h_k)\|$ is a bounded series and therefore has a convergent subsequence. Then, $\|\nabla_x f^h(x^h_k + s^h_k)\|$ converges to zero on the subsequence of successful iterations.

We can then state the global convergence property towards first-order critical points in the following theorem.

Theorem 1. *Let Assumptions 1 and 2 hold. Let $\{x^h_k\}$ be the sequence of fine level iterates generated by Algorithm 2. Then, $\{\|\nabla_x f^h(x^h_k)\|\}$ converges to zero on the subsequence of successful iterations.*

4.2 Worst-case complexity

We now want to evaluate the worst-case complexity of our methods, to reach a first order stationary point. We assume then that the procedure is stopped as soon as $\|\nabla_x f^h(x^h_k)\| \leq \epsilon$ for $\epsilon > 0$. The proof is similar to that of Theorem 2.5 in [2].

To evaluate the complexity of the proposed methods, we have to bound the number of successful and unsuccessful iterations performed before the stopping condition is met. Let us then define k_f the index of the last iterate for which $\|\nabla_x f^h(x^h_k)\| > \epsilon$, $K_s = \{0 < j \leq k_f \,|\, \rho_j \geq \eta_1\}$ the set of successful iterations before iteration k_f, and K_u its complementary in $\{1, \ldots, k_f\}$. We can use the same reasoning as that used to derive (27), but considering in the sum just the successful iterates in K_s. Remind that before termination $\|\nabla_x f^h(x^h_k)\| > \epsilon$ and, in case the lower level model is used, $\|R\nabla f^H(x^h_k)\| > \kappa_H \|\nabla f^H(x^h_k)\| > \kappa_H \epsilon$ (otherwise at that iteration the Taylor model would have been used). It then follows:

$$
\begin{aligned}
f^h(x^h_{k_1}) - \liminf_{k\to\infty} f^h(x^h_k) \geq f^h(x^h_{k_1}) - f^h(x^h_{k_f+1}) &= \sum_{j \in K_s} f^h(x^h_k) - f^h(x^h_k + s^h_k) \\
&\geq \frac{\eta_1 \lambda_{\min}}{q+1} \min\left\{ \frac{\kappa_H}{K_2}, \frac{1}{K_1} \right\}^{\frac{q+1}{q}} |K_s| \epsilon^{\frac{q+1}{q}},
\end{aligned}
$$

from which we get the desired bound on the total number of successful iterations. We can then bound the cardinality of K_u, with respect to the cardinality of K_s. From the updating rule of the regularization parameter, it holds:

$$
\gamma_1 \lambda_k \leq \lambda_{k+1},\ k \in K_s \qquad \gamma_3 \lambda_k = \lambda_{k+1},\ k \in K_u.
$$

Then, proceeding inductively, we conclude that:

$$
\lambda_0 \gamma_1^{|K_s|} \gamma_3^{|K_u|} \leq \lambda_{k_f} \leq \lambda_{\max}.
$$

Then,

$$
|K_s| \log \gamma_1 + |K_u| \log \gamma_3 \leq \log \frac{\lambda_{\max}}{\lambda_0},
$$

and, given that $\gamma_1 < 1$, we obtain:

$$
|K_u| \leq \frac{1}{\log \gamma_3} \log \frac{\lambda_{\max}}{\lambda_0} + |K_s| \frac{|\log \gamma_1|}{\log \gamma_3}.
$$

We can then state the following result.

Theorem 2. *Let Assumptions 1 and 2. Let f_{low} denote a lower bound on f and let k_1 denote the index of the first successful iteration in Algorithm 2. Then, given an absolute accuracy level $\epsilon > 0$, Algorithm 2 needs at most*

$$
K_3 \frac{(f(x_{k_1}) - f_{low})}{\epsilon^{\frac{q+1}{q}}} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_3} \right) + \frac{1}{\log \gamma_3} \log \left(\frac{\lambda_{\max}}{\lambda_0} \right)
$$

iterations in total to produce an iterate x_k^h *such that* $\|\nabla_x f(x_k)\| \leq \epsilon$*, where*

$$K_3 := \frac{q+1}{\eta_1 \lambda_{\min}} \max\left\{K_1, \frac{K_2}{\kappa_H}\right\}^{q+1/q},$$

with K_1 *and* K_2 *defined in* (24), (26), $\gamma_1, \gamma_3, \lambda_0, \lambda_{\min}$ *defined in Algorithm 2 and* $\lambda_{\max}$ *defined in* (23)*.*

Theorem 2 reveals that the use of lower level steps does not deteriorate the complexity of the method, and that the complexity bound $O(\epsilon^{-\frac{q+1}{q}})$ is preserved. This is a very satisfactory result, because each iteration of the multilevel methods will be less expensive than one iteration of the corresponding one-level method, thanks to the use of the cheaper lower level models. Consequently, if the number of iterations in the multilevel strategy is not increased, we can expect global computational savings.

4.3 Local convergence

In this section we study the local convergence of the proposed methods towards second-order stationary points. We assume $q \geq 2$ in this section, otherwise the problem is not well-defined. Thanks to the use of high order models, our methods are expected to attain a fast local convergence rate, especially for growing q. The results reported here are inspired by [27] and extend the analysis proposed therein.

We denote by $\mathcal{X}$ the set of second-order critical points of f, i.e. of points x^* satisfying the second-order necessary conditions:

$$\nabla_x f(x^*) = 0, \quad \nabla_x^2 f(x^*) \succeq 0,$$

i.e. $\nabla_x^2 f(x^*)$ is a symmetric positive semidefinite matrix. We denote by $\mathcal{B}(x,\rho) = \{y \, s.t. \, \|y - x\| \leq \rho\}$ and for all $x \in \mathbb{R}^n$, $\mathcal{L}(f(x)) = \{y \in \mathbb{R}^n \mid f(y) \leq f(x)\}$ for $f : \mathbb{R}^n \to \mathbb{R}$.

Remark 2. *From the assumption that* f *is a* q *times continuously differentiable function, it follows that its* i*-th derivative tensor is locally Lipschitz continuous for all* $i \leq q-1$*.*

Following [27], we first prove an intermediate lemma that allows us to relate, at generic iteration k, the norm of the step and the distance of the current iterate from the space of second-order stationary points. This lemma holds without need of assuming a stringent non-degeneracy condition, but rather under a local error bound condition, which is a much weaker requirement as it can be satisfied also when f has non isolated second-order critical points.

Assumption 3. *There exist strictly positive scalars* $\kappa_{EB}, \rho > 0$ *such that*

$$\operatorname{dist}(x, \mathcal{X}) \leq \kappa_{EB} \|\nabla_x f(x)\|, \quad \forall x \in \mathcal{N}(\mathcal{X}, \rho), \tag{28}$$

where $\mathcal{X}$ *is the set of second-order critical points of* f*,* $\operatorname{dist}(x, \mathcal{X})$ *denotes the distance of* x *to* $\mathcal{X}$ *and* $\mathcal{N}(\mathcal{X}, \rho) = \{x \mid \operatorname{dist}(x, \mathcal{X}) \leq \rho\}$*.*

This condition has been proposed for the first time in [27]. It is different from other error bound conditions in the literature as, in contrast to them, $\mathcal{X}$ is not the set of first-order critical points, but of second-order-critical points. In addition to being useful for proving convergence, it is also interesting on its own, as it is shown to be equivalent to a quadratic growth condition ([27, Theorem 1]) under mild assumptions on f.

Lemma 2. *Let Assumptions 1 and 2 hold. Let* $\{x_k^h\}$ *be the sequence generated by Algorithm 2 and* x_k^* *be a projection point of* x_k^h *onto* $\mathcal{X}$*. Assume that it exists a strictly positive constant* $\underline{\rho}$ *such that* $\{x_k^h\} \in \mathcal{B}(x_k^*, \underline{\rho})$ *and that* $\nabla_x^2 f$ *is Lipschitz continuous in* $\mathcal{B}(x_k^*, \underline{\rho})$ *with Lipschitz constant* L_2*. Then, it holds:*

$$\|s_k^h\| \leq C \operatorname{dist}(x_k^h, \mathcal{X}), \tag{29}$$

with

$$C = \begin{cases} C_f = \dfrac{1}{2\lambda_{\max}} \left[L_2 + \sqrt{L_2^2 + 4L_2\lambda_{\max}}\right], & \textit{(Taylor model)}, \\ \kappa_R C_c = \dfrac{\kappa_R}{2\lambda_{\max}} \left[\kappa_R L_2 + \sqrt{\kappa_R^2 L_2^2 + 4L_2\kappa_R\lambda_{\max}}\right], & \textit{(lower level model)}, \end{cases}$$

with $\lambda_{\max}$ *and* κ_R *defined respectively in* (23) *and Assumption 1.*

The proof of the lemma is reported in the appendix. This lemma can be used to prove that if it exists an accumulation point of $\{x_k^h\}$ that belongs to $\mathcal{X}$, then the full sequence converges to that point and that the rate of convergence depends on q. First, we can prove that the set of accumulation points is not empty.

Lemma 3. *Let Assumptions 1 and 2 hold. Let* $\{x_k^h\}$ *be the sequence of fine level iterates generated by Algorithm 2. If* $\mathcal{L}(f(x_k^h))$ *is bounded for some* $k \geq 0$*, then the sequence has an accumulation point that is a first-order stationary point.*

Proof. As $\{f(x_k^h)\}$ is a decreasing sequence, and $\mathcal{L}(f(x_k^h))$ is bounded for some $k \geq 0$, $\{x_k^h\}$ is a bounded sequence and it has an accumulation point. From Theorem 1, all the accumulation points are first-order stationary points. □

Theorem 3. *Let Assumptions 1 and 2 hold. Let $\{x_k^h\}$ be the sequence of fine level iterates generated by Algorithm 2. Assume that $\mathcal{L}(f(x_k^h))$ is bounded for some $k \geq 0$ and that it exists an accumulation point x^* such that $x^* \in \mathcal{X}$. Then, the whole sequence $\{x_k^h\}$ converges to x^* and it exist strictly positive constants $c \in \mathbb{R}$ and $\bar{k} \in \mathbb{N}$ such that:*

$$\frac{\|x_{k+1}^h - x^*\|}{\|x_k^h - x^*\|^q} \leq c, \quad \forall k \geq \bar{k}. \tag{30}$$

Proof. As x^* is an accumulation point, we have that $\lim_{k\to\infty} \text{dist}(x_k^h, \mathcal{X}) = 0$. Then, it exist ρ and k_1 such that $x_k^h \in \mathcal{N}(\mathcal{X}, \rho)$ for all $k \geq k_1$. Therefore, from Assumption 3 it holds

$$\text{dist}(x_k^h, \mathcal{X}) \leq \kappa_{EB} \, \|\nabla_x f^h(x_k^h)\|, \quad \forall k \geq k_1. \tag{31}$$

Moreover, from Remark 2, $\nabla_x^2 f$ is locally Lipschitz continuous, so Lemma 2 applies to all $k \geq k_1$.

Let us first consider the case in which the Taylor model is employed. It follows from (31), (24) and (29) that for all $k \geq k_1$

$$\text{dist}(x_{k+1}^h, \mathcal{X}) \leq \kappa_{EB} \, \|\nabla_x f^h(x_{k+1}^h)\| \leq \kappa_{EB} \, K_1 \, \|s_k^h\|^q \leq \kappa_{EB} \, K_1 \, C_f^q \, \text{dist}^q(x_k^h, \mathcal{X}).$$

If the lower level model is employed, from (31), (10), (26) and (36) it follows that for all $k \geq k_1$

$$\begin{aligned}\text{dist}(x_{k+1}^h, \mathcal{X}) \leq & \kappa_{EB} \, \|\nabla_x f^h(x_{k+1}^h)\| \leq \kappa_{EB} \, \kappa_H \, \|R \nabla_x f^h(x_{k+1}^h)\| \\ \leq & \kappa_{EB} \, \kappa_H \, K_2 \, \|s_k^H\|^q \leq \kappa_{EB} \, \kappa_H \, K_2 \, C_c^q \, \text{dist}^q(x_k^h, \mathcal{X}).\end{aligned}$$

Then in both cases, it exists $\bar{C}$ such that

$$\text{dist}(x_{k+1}^h, \mathcal{X}) \leq \bar{C} \, \text{dist}^q(x_k^h, \mathcal{X}), \quad \forall k \geq k_1,$$

where

$$\bar{C} = \begin{cases} \kappa_{EB} K_1 \, C_f^q & \text{(Taylor model)}, \\ \kappa_{EB} \kappa_H \, K_2 \, C_c^q & \text{(lower level model)}. \end{cases}$$

With this result, we can prove the convergence of $\{x_k^h\}$ with standard arguments. We repeat for example the arguments of the proof of [27, Theorem2] for convenience. Let $\eta > 0$ be an arbitrary value. As $\lim_{k\to\infty} \text{dist}(x_k^h, \mathcal{X}) = 0$, it exists $k_2 \geq 0$ such that

$$\text{dist}(x_k^h, \mathcal{X}) \leq \min\Big\{\frac{1}{2\bar{C}}, \frac{\eta}{2C}\Big\}, \quad \forall k \geq k_2.$$

Then,

$$\text{dist}(x_{k+1}^h, \mathcal{X}) \leq \bar{C} \text{dist}^q(x_k^h, \mathcal{X}) \leq \frac{1}{2} \text{dist}(x_k^h, \mathcal{X}), \quad \forall k \geq \bar{k} = \max\{k_1, k_2\}.$$

From (29), it then holds for all $k \geq \bar{k}$ and $j \geq 0$:

$$\begin{aligned}\|x_{k+j}^h - x_k^h\| \leq & \sum_{i=k}^{\infty} \|x_{i+1}^h - x_i^h\| \leq \sum_{i=k}^{\infty} C \text{dist}(x_i^h, \mathcal{X}) \\ \leq & \, C \text{dist}(x_k^h, \mathcal{X}) \sum_{i=0}^{\infty} \frac{1}{2^i} \leq 2C \text{dist}(x_k^h, \mathcal{X}) \leq \eta,\end{aligned}$$

i.e. that $\{x_k^h\}_{k \geq \bar{k}}$ is a Cauchy sequence and so the whole sequence is convergent. Finally we establish the q-th order rate of convergence of the sequence. For any $k \geq \bar{k}$,

$$\|x^* - x_{k+1}^h\| = \lim_{j\to\infty} \|x_{k+j+1}^h - x_{k+1}^h\| \leq 2C \text{dist}(x_{k+1}^h, \mathcal{X}) \leq 2C\bar{C} \text{dist}^q(x_k^h, \mathcal{X}). \tag{32}$$

Combining this with $\text{dist}(x_k^h, \mathcal{X}) \leq \|x_k^h - x^*\|$, and setting $c = 2C\bar{C}$ we obtain the thesis (30).

Therefore $\{x_k^h\}$ converges at least with order q to x^*. □

5 Numerical results

In this section, we report on the practical performance of a method in the family.

We have implemented the method corresponding to $q = 2$ in Algorithm 2 in Julia [1] (version 0.6.1). This is a multilevel extension of the method AR2 in Algorithm 1, which is better known as ARC [8, 10]. We will therefore denote the implemented multilevel method as MARC (multilevel adaptive method based on cubic regularization), rather than MAR2.

We consider the following two-dimensional nonlinear problem in the unit square domain S_2:

$$\begin{cases} -\Delta u(x,y) + e^{u(x,y)} = g(x,y) & \text{in } S_2, \\ u(x,y) = 0 & \text{on } \partial S_2, \end{cases}$$

where g is obtained such that the analytical solution to this problem is given by

$$u(x,y) = \sin(2\pi x(1-x)) \sin(2\pi y(1-y)).$$

		$n_h = 4096$		
u_0	Method	it_T/it_f	RMSE	save
$\bar{u}_1$	ARC	6/6	10^{-4}	
	MARC	9/4	10^{-4}	1.7-2.0-2.3
$\bar{u}_2$	ARC	17/17	10^{-4}	
	MARC	10/3	10^{-4}	1.9-5.8-8.3

Table 1: Solution of the minimization problem (33) with the one level ARC method and a four level ARC (MARC) (case of $n_h = 4096$) with $\bar{u}_1 = 1\,\mathrm{rand}(n_h, 1)$, $\bar{u}_2 = 3\,\mathrm{rand}(n_h, 1)$. it_T denotes the average number of iterations over ten simulations, it_f the average number of iterations in which the fine level model has been used, RMSE the root-mean square error with respect to the true solution and save the ratio between the total number of floating point operations required for the Cholesky factorizations in ARC and MARC, respectively.

The negative Laplacian operator is discretized using finite difference, giving a symmetric positive definite matrix A, that also takes into account the boundary conditions. The discretized version of the problem is then a system of the form $Au + e^u = g$, where u, g, e^u are vectors in $\mathbb{R}^{n_h}$, in which the columns of matrices U, G, E are stacked, with $U_{i,j} = u(x_i, y_j)$, $G_{i,j} = g(x_i, y_j)$, $E_{i,j} = e^{u(x_i,y_j)}$, for x_i, y_j grid points, $i, j = 1, \ldots, \sqrt{n_h}$.

The MARC algorithm is then used on the nonlinear minimization problem

$$\min_{u \in \mathbb{R}^{n_h}} \frac{1}{2} u^T A u + \|e^{u/2}\|^2 - g^T u, \tag{33}$$

which is equivalent to the system $Au + e^u = g$. The coarse approximations to the objective function arise from a coarser discretization of the problem. Each coarse two-dimensional grid has a dimension that is four times lower than the dimension of the grid on the corresponding upper level.

The prolongation operators P_ℓ from level $\ell - 1$ to ℓ are based on the nine-point interpolation scheme defined by the stencil $\begin{pmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$ and the full weighting operators defined by $R_\ell = \frac{1}{4} P_\ell^T$ are used as restriction operators [6].

We compare the one-level ARC with MARC. Parameters common to both methods are set as: $\epsilon^{l_{\max}} = 10^{-7}$, $\gamma_1 = 0.85$, $\gamma_2 = 0.5$, $\gamma_3 = 2$, $\lambda_0 = 0.05$ $\eta_1 = 0.1$, $\eta_2 = 0.75$. For MARC we set $\kappa = 0.1$ and $\epsilon^\ell = \epsilon^{\ell_{\max}}$ for all ℓ.

At each iteration we find an approximate minimizer of the cubic models as described in [8, §6.2]. This requires a sequence of Cholesky factorizations, which represents the dominant cost per nonlinear iteration. We measure the performance of the methods in terms of total number of floating point iterations required for these factorizations.

We study the effect of the multilevel strategy on the convergence of the method for problems of fixed dimension n_h. We then consider the solution of problem (33) using two different discretizations with $n_h = 4096$ (Table 1) and $n_h = 16384$ (Table 2), respectively. We allow 4 levels in MARC. We report the results of the average of ten simulations with different random initial guesses of the form $u_0 = a\,\mathrm{rand}(n_h, 1)$, for different values of a. In each simulation the random starting guess is the same for the two considered methods. All the quantities reported in Tables 1 and 2 are the average of the values obtained over the ten simulations. it_T denotes the number of total iterations, it_f denotes the number of iterations in which the Taylor model has been used, RMSE is the root-mean square error with respect to the true solution and save is the ratio between the total number of floating point operations required for the Cholesky factorizations by ARC and MARC, respectively. For the save quantity, we report three values: the minimum, the average and the maximum value obtained over the ten simulations.

The results reported in Tables 1 and 2 confirm the relevance of MARC as compared to ARC. The numerical experiments highlight the different convergence properties of both algorithms. The use of MARC is especially convenient when the initial guess is not so close to the true solution. Indeed, the performance of ARC deteriorates as the distance of the initial guess from the true solution increases, while MARC seems to be much less sensible to this choice. For the problem of smaller dimension, ARC still manages to find a solution for further initial guesses, even if this requires an higher number of iterations, while for the problem of larger dimension the method fails to find a solution in feasible time. The new multilevel approach is found to lead to considerable computational savings in terms of floating point operations compared to the classical one-level strategy.

6 Conclusions

We have introduced a family of multilevel methods of order $q \geq 1$ for unconstrained minimization. These methods represent an extension of the higher-order methods presented in [2] and of the multilevel trust-region method proposed in [13]. We have proposed a unifying framework to analyse these methods, which is useful to prove their convergence properties and evaluate their worst-case complexity to reach first-order stationary points. As expected, we show that the local rate of convergence and the complexity bound depend on q and high values of q allow both fast local convergence and lower complexity bounds.

		$n_h = 16384$		
u_0	Method	it_T/it_f	RMSE	save
$\bar{u}_1$	ARC	6/6	10^{-5}	
	MARC	12/3	10^{-5}	1.5-2.0-2.5
$\bar{u}_3$	ARC	FAIL	FAIL	
	MARC	18/5	10^{-5}	-

Table 2: Solution of the minimization problem (33) with the one level ARC method and a four level ARC (MARC) (case of $n_h = 16384$) with $\bar{u}_1 = 1\,\mathrm{rand}(n_h, 1)$, $\bar{u}_3 = 6\,\mathrm{rand}(n_h, 1)$. it_T denotes the average number of iterations over ten simulations, it_f the average number of iterations in which the fine level model has been used, RMSE the root-mean square error with respect to the true solution and save the ratio between the total number of floating point operations required for the Cholesky factorizations in ARC and MARC, respectively.

We believe this represents a contribution in the optimization field, as the use of multilevel ideas allows to reduce the major cost per iteration of the high-order methods. This gives a first answer to the question posed in [2] about whether the approach presented there can have practical implications, in applications for which computing q derivatives is feasible.

We have implemented the multilevel method corresponding to $q = 2$ and presented numerical results that show the considerable benefits of the multilevel strategy in terms of savings in floating point operations. Additional numerical results can be found in [7], where the authors apply the multilevel method in the family corresponding to $q = 1$ to problems arising in the training of artificial neural networks for the approximate solution of partial differential equations. This case is particularly interesting as it allows to show the efficiency of multilevel methods even for problems without an underlying geometrical structure.

References

[1] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah. Julia: A fresh approach to numerical computing. *SIAM Rev.*, 59(1):65–98, 2017.

[2] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Math. Program.*, 163(1):359–368, 2017.

[3] A. Bouaricha. Tensor methods for large, sparse unconstrained optimization. *SIAM J. Opt.*, 7(3):732–756, 1997.

[4] A. Brandt. A multi-level adaptative solution to boundary-value problems. *Math. Comp.*, 31:333–390, 1977.

[5] A. Brandt and O. E. Livne. *Multigrid Techniques: 1984 Guide with Applications to Fluid Dynamics*. SIAM, Philadelphia, 2011. Revised Edition.

[6] W. Briggs, V. Henson, and S. McCormick. *A Multigrid Tutorial*. SIAM, Philadelphia, second edition, 2000.

[7] H. Calandra, S. Gratton, E. Riccietti, and X. Vasseur. On the approximation of the solution of partial differential equations by artificial neural networks trained by a multilevel Levenberg-Marquardt method. *Technical report*, 2019.

[8] C. Cartis, N. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Math. Program., Series A*, 127(2):245–295, 2011.

[9] C. Cartis, N.I.M. Gould, and P. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *SIAM J. Opt.*, 20(6):2833–2852, 2010.

[10] C. Cartis, N.I.M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Math. Program.*, 130(2):295–319, Dec 2011.

[11] X. Chen, P. Toint, and H. Wang. Complexity of partially separable convexly constrained optimization with non-lipschitzian singularities. *SIAM Journal on Optimization*, 29(1):874–903, 2019.

[12] A. R. Conn, N. Gould, and Ph. L. Toint. *Trust region methods*. SIAM, 2000.

[13] S. Gratton, A. Sartenaer, and Ph L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM J. Opt.*, 19(1):414–444, 2008.

[14] A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical report, Technical report, University of Cambridge, 1981.

[15] W. Hackbusch. *Multi-grid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Heidelberg, 1985.

[16] M. Kočvara and S. Mohammed. A first-order multigrid method for bound-constrained convex optimization. *Optimization Methods and Software*, 31(3):622–644, 2016.

[17] R.M. Lewis and S.G. Nash. Model problems for the multigrid optimization of systems governed by differential equations. *SIAM J. Sci. Comput.*, 26(6):1811–1837, 2005.

[18] R.M. Lewis and S.G. Nash. Using inexact gradients in a multilevel optimization algorithm. *Computational Optimization and Applications*, 56(1):39–61, 2013.

[19] S.G. Nash. A multigrid approach to discretized optimization problems. *Optimization Methods and Software*, 14(1-2):99–116, 2000.

[20] S.G. Nash. Properties of a class of multilevel optimization algorithms for equality constrained problems. *Optimization Methods and Software*, 29(1):137–159, 2014.

[21] Y. Nesterov and B.T. Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, pages 177–205, 2006.

[22] J. Nocedal and S. Wright. *Numerical Optimization*. Springer-Verlag, New York, 2006.

[23] P. L. Toint. Nonlinear stepsize control, trust regions and regularizations for unconstrained optimization. *Optimization Methods and Software*, 28(1):82–95, 2013.

[24] S. Wang and S. Liu. A tensor trust-region model for nonlinear system. *Journal of Inequalities and Applications*, 2018(1):343, 2018.

[25] Z. Wen and D. Goldfarb. A line search multigrid method for large-scale nonlinear optimization. *SIAM J. Opt.*, 20(3):1478–1503, 2009.

[26] Y. Yuan. Recent advances in trust region algorithms. *Math. Program.*, 151(1):249–281, 2015.

[27] M.C. Yue, Z. Zhou, and A.M.C. So. On the quadratic convergence of the cubic regularization method under a local error bound condition. *eprint arXiv 1801.09387*, 2018.

A Proof of Lemma 4.5

In this appendix we report the proof of Lemma 2. We restate it here for convenience of the reader.

Lemma (Lemma 4.5)**.** *Let Assumptions 1 and 2 hold. Let $\{x_k^h\}$ be the sequence generated by Algorithm 2 and x_k^* be a projection point of x_k^h onto $\mathcal{X}$. Assume that it exists a strictly positive constant $\underline{\rho}$ such that $\{x_k^h\} \in \mathcal{B}(x_k^*, \underline{\rho})$ and that $\nabla_x^2 f$ is Lipschitz continuous in $\mathcal{B}(x_k^*, \underline{\rho})$ with Lipschitz constant L_2. Then, it holds:*

$$\|s_k^h\| \leq C \operatorname{dist}(x_k^h, \mathcal{X}),$$

with

$$C = \begin{cases} C_f = \dfrac{1}{2\lambda_{\max}} \left[L_2 + \sqrt{L_2^2 + 4L_2\lambda_{\max}} \right], & \textit{(Taylor model)}, \\ \kappa_R C_c = \dfrac{\kappa_R}{2\lambda_{\max}} \left[\kappa_R L_2 + \sqrt{\kappa_R^2 L_2^2 + 4L_2\kappa_R\lambda_{\max}} \right], & \textit{(lower level model)}, \end{cases}$$

with $\lambda_{\max}$ and κ_R defined respectively in (23) *and Assumption 1.*

Proof. The proof is divided into two parts. We first consider the case in which s_k^h has been obtained from the approximate minimization of the Taylor model, and then the case in which it has been obtained as prolongation of the step obtained from the approximate minimization of the coarse model.

Le us then assume that the Taylor model has been employed. Reminding that $\nabla_x f^h(x^*) = 0$ for each $x^* \in \mathcal{X}$, and definition (2) we obtain:

$$\begin{aligned} \nabla_s \left(m_{q,k}^h(x_k^h, s_k^h) + \frac{\lambda_k}{q+1}\|s_k^h\|^{q+1} \right) = &- \nabla_x f^h(x_k^*) + \nabla_x f^h(x_k^h) + \nabla_x^2 f^h(x_k^h) s_k^h \\ &+ H(s_k^h) + \lambda_k \|s_k^h\|^{q-1} s_k^h, \end{aligned} \tag{34}$$

with

$$H(s_k^h) = \sum_{i=3}^{q} \frac{1}{(i-1)!} \nabla_x^i f^h(x_k^h, \underbrace{s_k^h, \ldots, s_k^h}_{i-1 \text{ times}}).$$

Some algebraic manipulations (adding $\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^*)$ to both sides of (34) and expressing $(x_{k+1}^h - x_k^*) = s_k^h + (x_k^h - x_k^*)$) lead to:

$$\begin{aligned} &\left(\nabla_x^2 f^h(x_k^*) + \lambda_k \|s_k^h\|^{q-1}\right)(x_{k+1}^h - x_k^*) = \\ &\nabla_s \left(m_{q,k}^h(x_k^h, s_k^h) + \frac{\lambda_k}{q+1}\|s_k^h\|^{q+1} \right) + \nabla_x f^h(x_k^*) - \nabla_x f^h(x_k^h) - \nabla_x^2 f^h(x_k^*)(x_k^* - x_k^h) \\ &- H(s_k^h) + (\nabla_x^2 f^h(x_k^*) - \nabla_x^2 f^h(x_k^h)) s_k^h - \lambda_k \|s_k^h\|^{q-1}(x_k^* - x_k^h). \end{aligned}$$

Using the fact that $\nabla_x^2 f^h(x_k^*) \succeq 0$, the stopping criterion (18), and the triangle inequality, it follows

$$\lambda_k \|s_k^h\|^{q-1} \|x_{k+1}^h - x_k^*\| \leq \theta \|s_k^h\|^q + \|\nabla_x f^h(x_k^*) - \nabla_x f^h(x_k^h) - \nabla_x^2 f^h(x_k^*)(x_k^* - x_k^h)\| + \\ \|H(s_k^h)\| + \|\nabla_x^2 f^h(x_k^*) - \nabla_x^2 f^h(x_k^h)\| \|s_k^h\| + \lambda_k \|s_k^h\|^{q-1} \|x_k^* - x_k^h\|.$$

Using the Lipschitz continuity of $\nabla_x^2 f$ in $\mathcal{B}(x_k^*, \underline{\rho})$, the relation (16) with $q = 2$ and the triangle inequality $\|x_{k+1}^h - x_k^*\| \geq \|x_{k+1}^h - x_k^h\| - \|x_k^h - x_k^*\| = \|s_k^h\| - \|x_k^h - x_k^*\|$, we obtain:

$$\lambda_k \|s_k^h\|^q \leq \theta \|s_k^h\|^q + L_2 \|x_k^h - x_k^*\|^2 + \|H(s_k^h)\| + L_2 \|x_k^* - x_k^h\| \|s_k^h\| + 2\lambda_k \|s_k^h\|^{q-1} \|x_k^* - x_k^h\|.$$

Notice that

$$\theta \|s_k^h\|^q + L_2 \|x_k^h - x_k^*\|^2 + \|H(s_k^h)\| + L_2 \|x_k^* - x_k^h\| \|s_k^h\| + 2\lambda_k \|s_k^h\|^{q-1} \|x_k^* - x_k^h\| \\ \geq L_2 \|x_k^h - x_k^*\|^2 + L_2 \|x_k^* - x_k^h\| \|s_k^h\|.$$

We can then study when the inequality holds

$$\lambda_k \|s_k^h\|^q \leq L_2 \|x_k^h - x_k^*\|^2 + L_2 \|x_k^* - x_k^h\| \|s_k^h\|.$$

The right hand side of the inequality is expressed as a polynomial of $\|s_k^h\|$ of order 1 with positive value in 0, so the inequality will be true if $\|s_k^h\|$ is small enough. We can then assume $\|s_k^h\| < 1$, so that $\|s_k^h\|^q \leq \|s_k^h\|^2$ if $q \geq 2$. Then, we have that

$$\lambda_k \|s_k^h\|^q \leq \lambda_k \|s_k^h\|^2.$$

We can then solve

$$L_2 \|x_k^h - x_k^*\|^2 + L_2 \|x_k^* - x_k^h\| \|s_k^h\| - \lambda_k \|s_k^h\|^2 \geq 0.$$

The solution leads to

$$\|s_k^h\| \leq C_f \|x_k^h - x_k^*\|, \quad C_f = \frac{1}{2\lambda_k} \left[L_2 + \sqrt{L_2^2 + 4L_2\lambda_k} \right]. \tag{35}$$

Let us now consider the case in which the lower level model is used. The idea is similar as in the previous case. Reminding (25) and that $R\nabla_x f^h(x_k^*) = 0$, we have:

$$\nabla_s \left(m_k^H(x_{0,k}^H, s_k^H) + \frac{\lambda_k}{q+1} \|s_k^H\|^{q+1} \right) = \nabla_x f^H(x_{0,k}^H + s_k^H) - \nabla_s T_{q,k}^H(x_{0,k}^H, s_k^H) + R\nabla_x f^h(x_k^*) + \\ \sum_{i=1}^{q} \frac{1}{(i-1)!} \mathcal{R}(\nabla_x^i f^h(x_k^h)) \underbrace{(s_k^H, \ldots, s_k^H)}_{i-1 \text{ times}} + \lambda_k \|s_k^H\|^{q-1} s_k^H.$$

Algebraic manipulations (adding $R\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^*)$ to both sides and expressing $(x_{k+1}^h - x_k^*) = s_k^h + (x_k^h - x_k^*)$) lead to:

$$R\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^*) = \nabla_s \left(m_k^H(x_{0,k}^H, s_k^H) + \frac{\lambda_k}{q+1} \|s_k^H\|^{q+1} \right) - \nabla_x f^H(x_{0,k}^H + s_k^H) \\ + \nabla_s T_{q,k}^H(x_{0,k}^H, s_k^H) + R\nabla_x f^h(x_k^*) - R\nabla_x f^h(x_k^h) - R\nabla_x^2 f^h(x_k^*)(x_k^* - x_k^h) \\ - H_H(s_k^H) + R(\nabla_x^2 f^h(x_k^*) - \nabla_x^2 f^h(x_k^h)) s_k^h - \lambda_k \|s_k^H\|^{q-1} s_k^H,$$

where

$$H_H(s_k^H) = \sum_{i=3}^{q} \frac{1}{(i-1)!} \left[\mathcal{R}(\nabla_x^i f^h(x_k^H)) \right] (\underbrace{s_k^H, \ldots, s_k^H}_{i-1 \text{ times}}).$$

Further, we can write $R\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^*) = R\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^h) + R\nabla_x^2 f^h(x_k^*)(x_k^h - x_k^*) = R\nabla_x^2 f^h(x_k^*) P s_k^H + R\nabla_x^2 f^h(x_k^*)(x_k^h - x_k^*)$:

$$(R\nabla_x^2 f^h(x_k^*) P + \lambda_k \|s_k^H\|^{q-1}) s_k^H = \nabla_s \left(m_{q,k}^H(x_k^H, s_k^H) + \frac{\lambda_k}{q+1} \|s_k^H\|^q \right) - \nabla_x f^H(x_{0,k}^H + s_k^H) \\ + \nabla_s T_{q,k}^H(x_{0,k}^H, s_k^H) + R\nabla_x f^h(x_k^*) - R\nabla_x f^h(x_k^h) - R\nabla_x^2 f^h(x_k^*)(x_k^* - x_k^h) \\ - H_H(s_k^H) + R(\nabla_x^2 f^h(x_k^*) - \nabla_x^2 f^h(x_k^h)) P s_k^H - R\nabla_x^2 f^h(x_k^*)(x_k^h - x_k^*).$$

We can again use relation (16) (applied to $f^H, T_{q,k}^H$ with constant L_H and to $f, T_{2,k}^h$ with constant L_2), (18), the fact that $R\nabla_x^2 f^h(x_k^*) P$ is still positive definite, and Assumption 1 together with relation $s_k^h = P s_k^H$, to deduce that:

$$\lambda_k \|s_k^H\|^q \leq (\theta + L_H) \|s_k^H\|^q + \kappa_R L_2 \|x_k^* - x_k^h\|^2 + \|H_H(s_k^H)\| \\ + \kappa_R^2 L_2 \|x_k^* - x_k^h\| \|s_k^H\| + \|R\nabla_x^2 f^h(x_k^*)(x_k^h - x_k^*)\|.$$

We remark that

$$(\theta + L_H)\|s_k^H\|^q + \kappa_R L_2 \|x_k^* - x_k^h\|^2 + \|H_H(s_k^H)\| + \kappa_R^2 L_2 \|x_k^* - x_k^h\| \|s_k^H\| \\ + \|R\nabla_x^2 f^h(x_k^*)(x_k^h - x_k^*)\| \geq \kappa_R L_2 \|x_k^* - x_k^h\|^2 + \kappa_R L_2 \|x_k^* - x_k^h\| \|s_k^H\|.$$

As previously, we can solve the following inequality:

$$\lambda_k \|s_k^H\|^2 \leq \kappa_R L_2 \|x_k^* - x_k^h\|^2 + \kappa_R^2 L_2 \|x_k^* - x_k^h\| \|s_k^H\|,$$

and conclude that:

$$\|s_k^H\| \leq C_c \|x_k^h - x_k^*\|, \quad C_c = \frac{\left[\kappa_R L_2 + \sqrt{\kappa_R^2 L_2^2 + 4 L_2 \kappa_R \lambda_k}\right]}{2\lambda_k}. \tag{36}$$

We can then use the fact that $\lambda_k \leq \lambda_{\max}$ for all k and that $\|s_k^h\| \leq \kappa_R \|s_k^H\|$ to conclude that in all cases it exists a constant C such that $\|s_k^h\| \leq C\|x_k^h - x_k^*\|$. □