High-order Evaluation Complexity of a Stochastic Adaptive Regularization Algorithm for Nonconvex Optimization Using Inexact Function Evaluations and Randomly Perturbed Derivatives

S. Bellavia*, G. Gurioli†, B. Morini‡and Ph. L. Toint§

7 May 2020

Abstract

A stochastic adaptive regularization algorithm allowing random noise in derivatives and inexact function values is proposed for computing strong approximate minimizers of any order for inexpensively constrained smooth optimization problems. For an objective function with Lipschitz continuous p-th derivative in a convex neighbourhood of the feasible set and given an arbitrary optimality order q, it is shown that this algorithm will, in expectation, compute such a point in at most $O\Big(\big(\min_{j\in\{1,\ldots,q\}}\epsilon_j\big)^{-\frac{p+1}{p-q+1}}\Big)$ inexact evaluations of f and its derivatives whenever $q=1$ and the feasible set is convex, or $q=2$ and the problem is unconstrained, where ϵ_j is the tolerance for jth order accuracy. This bound becomes at most $O\Big(\big(\min_{j\in\{1,\ldots,q\}}\epsilon_j\big)^{-\frac{q(p+1)}{p}}\Big)$ inexact evaluations in the other cases if all derivatives are Lipschitz continuous. Moreover these bounds are sharp in the order of the accuracy tolerances.

Keywords: evaluation complexity, regularization methods, inexact functions and derivatives, stochastic analysis.

1 Introduction

We consider the complexity of an adaptive regularization algorithm for computing approximate local minimizers of arbitrary order for the possibly constrained minimization problem of the form

$$\min_{x\in\mathcal{X}} f(x), \tag{1.1}$$

whose main characteristics can be summarized as follows.

*Department of Industrial Engineering of Florence, Università degli Studi di Firenze, Italy. Member of the INdAM Research Group GNCS. Email: stefania.bellavia@unifi.it

†Department of Mathematics and Computer Science "Ulisse Dini", Università degli Studi di Firenze, Italy. Member of the INdAM Research Group GNCS. Email: gianmarco.gurioli@unifi.it

‡Department of Industrial Engineering of Florence, Università degli Studi di Firenze, Italy. Member of the INdAM Research Group GNCS. Email: benedetta.morini@unifi.it

§Namur Center for Complex Systems (naXys), University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@unamur.be

- The values of the objective function's j-th derivatives $\nabla_x^j f$ are subject to random noise and can only be computed inexactly. This is to say that only the approximation

$$\overline{\nabla_x^j f}(x) = \nabla_x^j f(x;\xi), \tag{1.2}$$

  can be calculated, where the inexactness/noise is characterised by a random variable ξ, generated by an unknown probability distribution. Inexact values of the objective function are also allowed, but are assumed to follow a "dynamic accuracy" framework in which the accuracy of these evaluations is deterministically controlled in response to the inexact values of the derivatives.

- The feasible set $\mathcal{X} \subseteq \Re^n$ represents "inexpensive constraints", that is constraints such that their evaluation/enforcement has negligible cost compared to that of computing the objective function. Such constraints include (but are not limited to) bound constraints and other convex constraints with cheap projections. Unconstrained problems are obviously also covered by this definition.

Context. Solving optimization problems involving inexact evaluations is not a new topic and has already been investigated in two different frameworks. The first is that of (deterministic) dynamic accuracy, where it is assumed that the accuracy of f and its derivatives can be specified and fullfilled by the algorithm (see [8, Section 10.6], [11], [10] or [1] for example). In this context, adaptive conditions are derived that guarantee convergence to approximate solutions, and evaluation complexity of the resulting algorithms can be analyzed [1], indicating a very modest degradation of the worst-case performance compared with the case where evaluations are exact [4, 5]. A drawback of this approach is that nothing is said for the case where the requested accuracy requirement cannot be met or, as is often the case, cannot even be measured. This problem does not occur in the second case, in which the inexactness in the function (and possible derivatives) values can be seen as caused by some random noise ξ (as in (1.2)), in which case the algorithm/user is not able to specify an accuracy level and poor accuracy might result. The available analysis for this case differs by the assumptions made on the distribution of this noise. In [12], the authors consider a linesearch method for the unbiased case, in which $\mathbb{E}_\xi[\overline{f}(x;\xi)] = f(x)$, and estimate its evaluation complexity for finding approximate first-order critical points. It is assumed instead that the function values of f, as well as those of its derivatives, can be approximated within a prescribed accuracy with a fixed, sufficiently high probability, conditioned to the past. A similar context is considered in [2], where the objective function values are inexact but computed with accuracy guaranteed with probability one.

A trust-region variant is also proposed in [7], where it is proved to converge almost-surely to first-order critical points. The approach of [6] includes the use of random first-order models and directions within line search method as well as probabilistic second-order models in the Adaptive Cubic Regularization (ARC) framework. In both cases, the authors employ exact function evaluations. A general theory for global convergence rate analysis is also provided. More recently, [3] proposed a complexity analysis for a trust-region method (covering convergence to second-order points) using elegant properties of sub-martingales, and making no assumption on bias. A recent overview is proposed in [9].

Significantly for our purpose, all above references, except for [1, 4, 5], are restricted to unconstrained problems and have considered the evaluation complexity of computing approximate first- or second-order critical points. It is also the case that all methods described

in these references make use of a polynomial model, ranging from simply linear, when only approximate first-order points are sought, to quadratic or models of arbitrary degree, when approximate critical points of higher order are of interest.

We have included inexpensive constraints in the formulation of problem (1.1) because of their relevance. In particular, they cover most convex constraints, such as bounds on the variables. They are of special interest here since the evaluation complexity of solving inexpensively constrained problems is, for such constraints, well captured by the number of evaluations of the objective function and its derivatives. They have been considered from that point of view in [1, 4, 5].

Contributions. Having set the scene, we now make the contributions of this paper more precise.

- We consider problem (1.1) assuming that the objective function values can be computed within a prescribed accuracy, but allowing randomly inexact evaluations of its derivatives, thereby using a mix of the two strategies described above. Our probabilistic assumptions on the derivatives approximations extend those of [3]: in particular we do not assume unbiased estimations. Our choice of imposing dynamic accuracy framework for the objective function complements that of [2], allowing for more inaccuracy, but in a deterministic context.

- As in [1, 4, 5], we propose a regularization algorithm for the solution of the problem, which is based on polynomial models of arbitrary degree. This allows us to seek for first- and second-order critical points, but also for critical points of arbitrary order (we define what we mean by that in Section 2).

- We establish sharp worst-case bounds (in expectation) on the evaluation complexity of computing these (possibly high-order) approximate critical points, depending on the order and on the degree of the polynomial model used. These bounds correspond in order to the best known bounds for regularization algorithms using exact evaluations.

These results are obtained by building on the probabilistic framework of [6], and by merging approximation results in [1] with techniques of [5].

Outline. The paper is organized as follows. Section 2 discusses optimality measures for arbitrary order and introduces the regularization algorithm and the associated stochastic assumptions. Its evaluation complexity is then studied in Section 3. We finally present some conclusions and perspectives in Section 4.

Notations. Unless otherwise specified, $\|\cdot\|$ denotes the standard Euclidean norm for both vectors and matrices. For a general symmetric tensor S of order p, we define

$$\|S\| \stackrel{\rm def}{=} \max_{\|v\|=1} |S[v]^p| = \max_{\|v_1\|=\cdots=\|v_p\|=1} |S[v_1,\ldots,v_p]| \tag{1.3}$$

the induced Euclidean norm (see [13, Theorem 2.1] for a proof of the second equality). We denote by $\nabla_x^\ell f(x)$ the ℓ-th order derivative of f evaluated at x, noting that such a tensor is always symmetric for any $\ell \geq 2$. The notation $\nabla_x^\ell f(x)[s]^\ell$ denotes this ℓ-th derivative tensor applied to ℓ copies of the vector s. All inexact quantities are indicated by an overbar (as in (1.2)). For symmetric matrices M, $\lambda_{\min}(M)$ is the leftmost eigenvalue of M. We will also use

the function

$$\chi_j(t) \stackrel{\rm def}{=} \sum_{\ell=1}^{j} \frac{t^\ell}{\ell!} \quad (t \geq 0), \tag{1.4}$$

where $j \geq 1$. We use the notation $\mathbb{E}[X]$ to indicate the expected value of a random variable X. In addition, given a random event A, $\mathbb{P}{\rm r}(A)$ denotes the probability of A, while $\mathbb{1}_A$ refers to the indicator of the random event A occurring. The notation A^c indicates that event A does not occur.

2 A stochastic regularization algorithm with inexact evaluations

2.1 Preliminaries

We first make our framework more formal by describing our assumptions on problem (1.1).

AS.1 The function f is p times continuously differentiable in a convex neighbourhood of $\mathcal{X}$. Moreover, its j-th order derivative tensor is Lipschitz continuous for $j \in \{1, \ldots, p\}$ in the sense that there exist constants $L_{f,j} \geq 0$ such that, for all $j \in \{1, \ldots, p\}$ and all x, y in that neighbourhood,

$$\|\nabla_x^j f(x) - \nabla_x^j f(y)\| \leq L_{f,j}\|x - y\|. \tag{2.1}$$

AS.2 f is bounded below in $\mathcal{X}$, that is there exists a constant $f_{\rm low}$ such that $f(x) \geq f_{\rm low}$ for all $x \in \mathcal{X}$.

Note that, if $\mathcal{X}$ is convex, then AS.1 can be restricted to hold in an open neighbourhood of $\mathcal{X}$.

Under AS.1, the p-th order Taylor series of f taken at a point x and evaluated for a step s is well-defined and can be written as

$$T_{f,p}(x, s) \stackrel{\rm def}{=} f(x) + \sum_{\ell=1}^{p} \frac{1}{\ell!} \nabla_x^\ell f(x)[s]^\ell, \tag{2.2}$$

where $\nabla_x^\ell f(x)[s]^\ell$ denotes the scalar obtained by applying the ℓ-dimensional tensor $\nabla_x^\ell f(x)$ to ℓ copies of the vector s. We will make frequent use of the *Taylor increment* defined as

$$\Delta T_{f,p}(x, s) = T_{f,p}(x, 0) - T_{f,p}(x, s) = -\sum_{\ell=1}^{p} \frac{1}{\ell!} \nabla_x^\ell f(x)[s]^\ell \tag{2.3}$$

We will also rely on the following important consequence of AS.1.

Lemma 2.1 Suppose that AS.1 holds. Then, for all $x, s \in \Re^n$,

$$f(x+s) \leq T_{f,p}(x,s) + \frac{L_{f,p}}{(p+1)!}\|s\|^{p+1} \tag{2.4}$$

and

$$\|\nabla_x^\ell f(x+s) - \nabla_s^\ell T_{f,p}(x,s)\| \leq \frac{L_{f,p}}{(p-\ell+1)!}\|s\|^{p-\ell+1} \text{ for all } \ell \in \{1,\ldots,q\}. \tag{2.5}$$

Proof. See [4, Lemma 2.1]. □

We now turn to the important question of defining what we mean by (approximate) critical points of arbitrary order. We will use an extension of the definition of strong approximate minimizers discussed in [5] to the constrained case. Specifically, given "accuracy requests" $\epsilon = (\epsilon_1, \ldots, \epsilon_q)$ and "optimality radiuses" $\delta = (\delta_1, \ldots, \delta_q)$ with

$$\epsilon_j \in (0,1] \text{ and } \delta_j \in (0,1] \text{ for } j \in \{1,\ldots,q\}$$

we say that $x \in \mathcal{X}$ is a (strong) q-th order (ϵ,δ)-approximate minimizer for problem (1.1) if

$$\phi_{f,j}^{\delta_j}(x) \leq \epsilon_j \frac{\delta_j^j}{j!} \quad \text{for} \quad j \in \{1,\ldots,q\}, \tag{2.6}$$

where

$$\phi_{f,j}^{\delta_j}(x) \stackrel{\rm def}{=} f(x) - \min_{x+d\in\mathcal{X}, \|d\|\leq\delta_j} T_{f,j}(x,d) = \max_{x+d\in\mathcal{X}, \|d\|\leq\delta_j} \Delta T_{f,j}(x,d), \tag{2.7}$$

where, as is standard, the min and max are considered global. Note that $\phi_{f,j}^{\delta}(x)$, $\delta > 0$, is always non-negative. We stress that this notion of approximate minimizer is stronger than that of a "weak" approximate minimizer (used for instance in [1, 4]) where the condition $\phi_{f,q}^{\delta}(x) \leq \epsilon_\chi \chi_q(\delta)$ replaces (2.6) (ϵ_χ is now a scalar in $(0,1)$). That this is weaker than (2.6) is easily seen if one observes that, because χ_q is convex, $\chi_q(0) = 0$ and $\chi_q(1) \leq e - 1 < 2$,

$$\chi_q(\delta) \in [\delta, 2\delta) \quad \text{for any} \quad \delta \in [0,1]. \tag{2.8}$$

As a consequence, $\chi_q(\delta)$ is typically significantly larger that $\delta_j^j/j!$ for $j \in \{1,\ldots,q\}$. We refer the reader to [5] for a further discussion of the choice of (2.6), including the observation that this optimality measure reduces to the familiar necessary conditions for orders one and two, in which cases the value $\delta_j = 1$ for all $j \in \{1,\ldots,q\}$ is always acceptable.

2.2 The regularization algorithm

We are now in position to describe our stochastic adaptive regularization algorithm SARqp whose purpose is to compute a q-th order (ϵ,δ)-approximate minimizer for problem (1.1). The vector of accuracies ϵ and the optimality order $q \geq 1$ are given. A model degree $p \geq q$ is also given, corresponding to the maximum order of available derivatives. If the objective function f and its derivatives of orders ranging from one to p were known exactly, a typical

adaptive regularization method could be outlined as follows. At iteration k, a local model of the objective function's variation would first be defined by regularizing the Taylor series of degree p at the current iterate x_k, namely

$$m_k(s) = -\Delta T_{f,p}(x_k, s) + \frac{\sigma_k}{(p+1)!}\|s\|^{p+1}. \tag{2.9}$$

A step s_k would next be computed by approximately minimizing $m_k(s)$ for $x_k + s \in \mathcal{X}$ in the sense that $m_k(s_k) \leq m_k(0) = 0$ and

$$\phi_{m_k,j}^{\delta_{k,j}}(s_k) \leq \theta\epsilon_j \frac{\delta_{k,j}^j}{j!}, \tag{2.10}$$

for some $\theta \in (0, \frac{1}{2})$ and $\delta_k \in (0,1]^q$. In this condition, $\phi_{m_k,j}^{\delta_{k,j}}(s_k)$ is the j-th order optimality measure (2.7) for the model (2.9) computed at s_k, in which, for $j \in \{1, \ldots, q\}$,

$$T_{m_k,j}(s_k, d) = m_k(s_k) + \sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k, s_k)[d]^\ell + \frac{\sigma_k}{(p+1)!}\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell\|s_k\|^{p+1}[d]^\ell \tag{2.11}$$

and thus

$$\Delta T_{m_k,j}(s_k, d) = -\sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k, s_k)[d]^\ell - \frac{\sigma_k}{(p+1)!}\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell\|s_k\|^{p+1}[d]^\ell.$$

The values of $f(x_k + s_k)$ and $\{\nabla_x^\ell f(x_k + s_k)\}_{\ell=1}^p$ would then be computed and the trial point $x_k + s_k$ would then be accepted as the next iterate, provided the ratio

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{\Delta T_{f,p}(x_k, s_k)},$$

is sufficiently positive. The regularization parameter σ_k would then be updated before a new iteration is started. (See [4] for the complete description of such an algorithm using exact function and derivatives values.) The SARqp algorithm follows the same lines, except that the values of $f(x_k)$, $f(x_k + s_k)$ and $\Delta T_{f,p}(x_k, s_k)$ are not known exactly, the inexactness in the latter resulting from the inexactness of the derivatives $\{\nabla_x^\ell f(x_k)\}_{\ell=1}^p$. Instead, inexact values $\overline{f}(x_k)$, $\overline{f}(x_k + s_k)$ and $\overline{\Delta T}_{f,p}(x_k, s_k)$ are now computed and used to define the model

$$m_k(s) = -\overline{\Delta T}_{f,p}(x_k, s) + \frac{\sigma_k}{(p+1)!}\|s\|^{p+1}. \tag{2.12}$$

In particular, given $\alpha \in (0,1)$ and setting

$$\omega_k = \min\left[\frac{\alpha\eta}{2}, \frac{1}{\sigma_k}\right], \tag{2.13}$$

the approximations $\overline{f}(x_k)$ and $\overline{f}(x_k + s_k)$ are required to satisfy the following accuracy conditions:

$$\left|\overline{f}(x_k) - f(x_k)\right| \leq \omega_k\overline{\Delta T}_{f,p}(x_k, s_k), \tag{2.14}$$

$$\left|\overline{f}(x_k + s_k) - f(x_k + s_k)\right| \leq \omega_k\overline{\Delta T}_{f,p}(x_k, s_k). \tag{2.15}$$

In what follows, we will consistently denote inexact values by an overbar.

The model (2.12) is then approximately minimized by the feasible step s_k in the sense that the trial point $x_k + s_k \in \mathcal{X}$,

$$m_k(s_k) \leq m_k(0) = 0 \tag{2.16}$$

and

$$\overline{\phi}^{\delta_{k,j}}_{m_k,j}(s_k) = \max_{x_k+s_k+d\in\mathcal{X},\|d\|\leq\delta_{k,j}} \overline{\Delta T}_{m_k,j}(s_k,d) \leq \theta\epsilon_j \frac{\delta^j_{k,j}}{j!}, \tag{2.17}$$

for $j \in \{1,\ldots,q\}$ and some $\theta \in (0,\frac{1}{2})$ and $\delta_k \in (0,1]^q$. The values $\overline{f}(x_k)$, $\overline{f}(x_k+s_k)$ and $\overline{\Delta T}_{f,p}(x_k,s_k)$ are also used to compute the ratio ρ_k, the value of which decides of the acceptance of the trial point. The SARqp algorithm is detailed as Algorithm 2.1 on the following page.

We first verify that the algorithm is well-defined.

Lemma 2.2 A step s_k satisfying (2.16) and (2.17) for $j \in \{1,\ldots,q\}$ and some $\delta_k \in (0,1]^q$ always exists.

Proof. The proof is a direct extension of that of [5, Lemma 4.4] using inexact models. It is given in appendix for completeness. □

Some comments on this algorithm are useful at this stage.

1. It is important to observe that the algorithm is fully implementable with existing computational technology in the very frequent cases where $q = 1$ and $\mathcal{X}$ is convex or $q = 2$ and the problem is unconstrained ($\mathcal{X} = \Re^n$). Indeed the value of $\overline{\phi}^{\delta_{k,1}}_{m_k,1}$ can easily be obtained analytically for the unconstrained case, while it results from a simple linear/convex problem otherwise (for instance if $\mathcal{X}$ is defined by simple bounds). Thus the SARqp algorithm reduces to a very classical approach when $q = 1$. When $q = 2$ and the problem is unconstrained, the same comment obviously applies for $\overline{\phi}^{\delta_{k,1}}_{m_k,1}$, while the value $\overline{\phi}^{\delta_{k,2}}_{m_k,2}$ can be computed by a standard trust-region solver (whose cost is comparable to that of the more usual calculation of the most negative eigenvalue), again making the algorithm practical.

   In other cases, the computation $\overline{\phi}^{\delta_{k,j}}_{m_k,j}$ may be extremely expensive, making our approach mostly theoretical at this stage. However, we note that since evaluations of the objective function and its derivatives do not occur in this computation (once the approximate derivatives are known), its cost has no impact on the evaluation complexity of the SARqp algorithm.

2. We assume in what follows that, once the inexact model $m_k(s)$ is determined, then the computation of the pair (s_k,δ_k) (and thus of the trial point $x_k + s_k$) is deterministic. Moreover, we assume that the mechanism which ensures (2.14)-(2.15) in Step 3 of the algorithm is also deterministic, so that ρ_k and the fact that iteration k is successful

Algorithm 2.1: The SARqp Algorithm

Step 0: Initialization. An initial point $x_0 \in \mathcal{X}$ and an initial regularization parameter $\sigma_0 > 0$ are given, as well as a vector of accuracies $\epsilon \in (0,1]^q$. The constants $\theta \in (0, \frac{1}{2})$, $\eta \in (0,1)$, $\gamma > 1$, $\alpha \in (0,1)$, $\omega_0 = \min\left[\frac{1}{2}\alpha\eta, \frac{1}{\sigma_0}\right]$ and $\sigma_{\min} \in (0, \sigma_0)$ are also given. Set $k = 0$.

Step 1: Model construction. Compute approximate derivatives $\{\overline{\nabla_x^\ell f}(x_k)\}_{\ell\in\{1,\ldots,p\}}$ and form the model $m_k(s)$ defined in (2.12).

Step 2: Step calculation. Compute a step s_k satisfying $x_k + s_k \in \mathcal{X}$, (2.16) and (2.17) for $j \in \{1,\ldots,q\}$ and some $\delta_k \in (0,1]^q$. If $\overline{\Delta T}_{f,p}(x_k, s_k) = 0$, go to Step 4.

Step 3: Function estimates computation. Compute the approximations $\overline{f}(x_k)$ and $\overline{f}(x_k + s_k)$ of $f(x_k)$ and $f(x_k + s_k)$, respectively, such that (2.14)–(2.15) are satisfied.

Step 4: Acceptance test. Set

$$\rho_k = \begin{cases} \dfrac{\overline{f}(x_k) - \overline{f}(x_k + s_k)}{\overline{\Delta T}_{f,p}(x_k, s_k)} & \text{if } \ \overline{\Delta T}_{f,p}(x_k, s_k) > 0, \\ -\infty & \text{otherwise.} \end{cases} \tag{2.18}$$

If $\rho_k \geq \eta$ (*successful iteration*), then define $x_{k+1} = x_k + s_k$; otherwise (*unsuccessful iteration*) define $x_{k+1} = x_k$.

Step 5: Regularization parameter update. Set

$$\sigma_{k+1} = \begin{cases} \max\left[\sigma_{\min}, \frac{1}{\gamma}\sigma_k\right], & \text{if } \ \rho_k \geq \eta, \\ \gamma\sigma_k, & \text{if } \ \rho_k < \eta. \end{cases} \tag{2.19}$$

Step 6: Relative accuracy update. Set

$$\omega_{k+1} = \min\left[\frac{1}{2}\alpha\eta, \frac{1}{\sigma_{k+1}}\right].$$

Increment k by one and go to Step 1.

are deterministic outcomes of the realization of the inexact model. Such a deterministic accuracy improvement mechanism may occur in several contexts: the discretization used for computing f may be refined or a tolerance in an iterative procedure internal to f may be tightened for example.

3. Observe that, because we have chosen m_k to be a model of the local variation in f rather than a model of f itself, $\overline{f}(x_k)$ is not needed (and not computed) in Steps 1 and 2 of the algorithm. This distinguishes the SARqp algorithm from the approaches of [3, 7].

In what follows, all random quantities are denoted by capital letters, while the use of small letters is reserved for their realization. In particular, let us denote a random model at iteration k as M_k, while we use the notation $m_k = M_k(\zeta_k)$ for its realizations, where ζ_k is a random realization at iteration k taken from a suitable context-dependent probability space. Similarly, the iterates X_k, as well as the regularization parameters Σ_k and the steps S_k are the random variables such that $x_k = X_k(\zeta_k)$, $\sigma_k = \Sigma_k(\zeta_k)$ and $s_k = S_k(\zeta_k)$. Moreover, each realization of the random vector δ_k and the threshold ω_k can be seen as $\delta_k = \Delta_k(\zeta_k)$ and $\omega_k = \Omega_k(\zeta_k)$. Hence, the SAR$qp$ Algorithm generates a random process

$$\{X_k, S_k, M_k, \Sigma_k, \Delta_k, \Omega_k\}. \tag{2.20}$$

In order to limit the notational burden, we will omit the explicit reference to ζ_k in what follows, making it implicit for all random variables considered.

2.3 The probabilistic setting

In view of our last comment, we now make our probabilistic assumptions on the SARqp algorithm explicit. For $k \geq 0$, our assumption on the past is formalized by considering $\mathcal{A}_{k-1}^M$ the $\hat{\sigma}$-algebra induced by the random variables M_0, M_1,..., M_{k-1}, with $\mathcal{A}_{-1}^M = \hat{\sigma}(x_0)$. In order to formalize our probabilistic assumptions we need a few more definitions. We define, at iteration k of an arbitrary realization,

$$d_{k,j} = \arg \max_{x_k+s_k+d\in\mathcal{X}, \|d\|\leq\delta_{k,j}} \Delta T_{m_k,j}(s_k, d) \tag{2.21}$$

the argument of the maximum in the definition of $\phi_{m_k,j}^{\delta_{k,j}}(x_k)$, and

$$\overline{d}_{k,j} = \arg \max_{x_k+s_k+d\in\mathcal{X}, \|d\|\leq\delta_{k,j}} \overline{\Delta T}_{m_k,j}(s_k, d) \tag{2.22}$$

that in the definition of $\overline{\phi}_{m_k,j}^{\delta_{k,j}}(s_k)$. We also define, at the end of Step 2 of iteration k, the events

$$\mathcal{M}_k = \begin{cases} \mathcal{M}_k^{(1)} \cap \bigcap_{j=1}^q \left(\mathcal{M}_{k,j}^{(2)} \cap \mathcal{M}_{k,j}^{(3)}\right) & \text{if } q = 1 \text{ and } \mathcal{X} \text{ is convex, or} \\ & \text{if } q = 2 \text{ and } \mathcal{X} = \Re^n, \\ \mathcal{M}_k^{(1)} \cap \mathcal{M}_k^{(4)} \cap \bigcap_{j=1}^q \left(\mathcal{M}_{k,j}^{(2)} \cap \mathcal{M}_{k,j}^{(3)}\right) & \text{otherwise,} \end{cases} \tag{2.23}$$

with

$$\begin{aligned} \mathcal{M}_k^{(1)} &= \left\{|\overline{\Delta T}_{f,p}(X_k, S_k) - \Delta T_{f,p}(X_k, S_k)| \leq \Omega_k \overline{\Delta T}_{f,p}(X_k, S_k)\right\}, \\ \mathcal{M}_{k,j}^{(2)} &= \left\{|\overline{\Delta T}_{m_k,j}(S_k, D_{k,j}) - \Delta T_{m_k,j}(S_k, D_{k,j})| \leq \Omega_k \overline{\Delta T}_{m_k,j}(S_k, D_{k,j})\right\}, \\ \mathcal{M}_{k,j}^{(3)} &= \left\{|\overline{\Delta T}_{m_k,j}(S_k, \overline{D}_{k,j}) - \Delta T_{m_k,j}(S_k, \overline{D}_{k,j})| \leq \Omega_k \overline{\Delta T}_{m_k,j}(S_k, \overline{D}_{k,j})\right\}, \\ \mathcal{M}_k^{(4)} &= \left\{ \max_{\ell\in\{2,\ldots,p\}} \|\overline{\nabla_x^\ell f}(X_k)\| \leq \Theta\right\}, \end{aligned}$$

for some $\Theta > 0$ independent of k. In what follows, we will say that iteration k is *accurate*, if $\mathbb{1}_{\mathcal{M}_k} = 1$, and iteration k is *inaccurate*, if $\mathbb{1}_{\mathcal{M}_k} = 0$. Immediately note that, as specified by (2.23), $\mathcal{M}_k^{(4)}$ is not required if $q = 1$ and $\mathcal{X}$ is convex, nor if $q = 2$ and the problem is unconstrained.

The conditions defining $\mathcal{M}_k$ may seem abstract at first sight, but we now motivate them by looking at what kind of accuracy on each derivative $\overline{\nabla_x^\ell f}(x_k)$ ensures that they hold.

Lemma 2.3 For each $k \geq 0$, we have the following.

1. Let
$$\tau_k \stackrel{\rm def}{=} \max\left[\|S_k\|, \max_{j\in\{1,\ldots,q\}}[\|D_{k,j}\|, \|\overline{D}_{k,j}\|]\right] \tag{2.24}$$
and
$$\overline{\Delta T}_{k,\min} \stackrel{\rm def}{=} \min\left[\overline{\Delta T}_{f,p}(X_k,S_k), \min_{j\in\{1,\ldots,q\}}\left[\overline{\Delta T}_{m_k,j}(S_k,D_{k,j}), \overline{\Delta T}_{m_k,j}(S_k,\overline{D}_{k,j})\right]\right]. \tag{2.25}$$
Then $\mathcal{M}_k^{(1)}$, $\{\mathcal{M}_{k,j}^{(2)}\}_{j=1}^q$ and $\{\mathcal{M}_{k,j}^{(3)}\}_{j=1}^q$ occur if
$$\|\overline{\nabla_x^\ell f}(X_k) - \nabla_x^\ell f(X_k)\| \leq \Omega_k \frac{\overline{\Delta T}_{k,\min}}{6\tau_k^\ell} \quad \text{for} \quad \ell \in \{1,\ldots,p\}. \tag{2.26}$$

2. Suppose that AS.1 holds. Then $\mathcal{M}_k^{(4)}$ occurs if
$$\|\overline{\nabla_x^\ell f}(X_k) - \nabla_x^\ell f(X_k)\| \leq \Theta_0 \quad \text{for} \quad \ell \in \{2,\ldots,p\} \tag{2.27}$$
and some constant $\Theta_0 \geq 0$ independent of k and ℓ.

Proof. Consider the first assertion. That $\mathcal{M}_k^{(1)}$ occurs follows from the inequalities
$$\begin{aligned}
|\overline{\Delta T}_{f,p}(X_k,S_k) - \Delta T_{f,p}(X_k,S_k)| &\leq \sum_{\ell=1}^p \frac{\|S_k\|^\ell}{\ell!}\,\|\overline{\nabla_x^\ell f}(X_k) - \nabla_x^\ell f(X_k)\| \\
&\leq \sum_{\ell=1}^p \frac{\tau_k^\ell}{\ell!}\,\|\overline{\nabla_x^\ell f}(X_k) - \nabla_x^\ell f(X_k)\| \\
&\leq \sum_{\ell=1}^p \frac{\Omega_k}{6\ell!}\,\overline{\Delta T}_{k,\min} \\
&\leq \sum_{\ell=1}^p \frac{\Omega_k}{6\ell!}\,\overline{\Delta T}_{f,p}(X_k,S_k) \\
&\leq \frac{1}{6}\chi_p(1)\,\Omega_k\,\overline{\Delta T}_{f,p}(X_k,S_k) \\
&< \Omega_k\,\overline{\Delta T}_{f,p}(X_k,S_k).
\end{aligned}$$
where we have used (2.24), (2.26), (2.25) and the fact that $\chi_p(1) \leq 2$. The verification that $\{\mathcal{M}_{k,j}^{(2)}\}_{j=1}^q$ and $\{\mathcal{M}_{k,j}^{(3)}\}_{j=1}^q$ also occur uses a very similar argument, with one additional

ingredient: employing the triangle inequality, (2.12), we have that, for all $\ell \in \{1, \ldots, p\}$,

$$\left\| \overline{\nabla_d^\ell T}_{m_k,j}(S_k, 0) - \nabla_d^\ell T_{m_k,j}(S_k, 0) \right\| \leq \sum_{t=\ell}^{p} \left\| \overline{\nabla_x^t f}(X_k) - \nabla_x^t f(X_k) \right\| \frac{\|S_k\|^{t-\ell}}{(t-\ell)!}.$$

Considering now $D = D_{k,j}$ or $D = \overline{D}_{k,j}$ and using the above inequality, (2.24), (2.26), (2.25) and the facts that $\chi_j(1) \leq 2$ and $\chi_{p-\ell}(1) \leq 2$, we have that

$$\begin{aligned}
&|\overline{\Delta T}_{m_k,j}(S_k, D) - \Delta T_{m_k,j}(S_k, D)| \\
&\qquad \leq \sum_{\ell=1}^{j} \frac{\|D\|^\ell}{\ell!} \|\overline{\nabla_d^\ell T_{m_k,j}}(S_k, 0) - \nabla_d^\ell T_{m_k,j}(S_k, 0)\| \\
&\qquad \leq \sum_{\ell=1}^{j} \frac{\|D\|^\ell}{\ell!} \sum_{t=\ell}^{p} \left\| \overline{\nabla_x^t f}(X_k) - \nabla_x^t f(X_k) \right\| \frac{\|S_k\|^{t-\ell}}{(t-\ell)!} \\
&\qquad \leq \sum_{\ell=1}^{j} \frac{1}{\ell!} \sum_{t=\ell}^{p} \left\| \overline{\nabla_x^t f}(X_k) - \nabla_x^t f(X_k) \right\| \frac{\tau_k^t}{(t-\ell)!} \\
&\qquad \leq \sum_{\ell=1}^{j} \frac{1}{\ell!} \sum_{t=\ell}^{p} \frac{1}{(t-\ell)!} \Omega_k \frac{\overline{\Delta T}_{k,\min}}{6} \\
&\qquad \leq \frac{1}{6} \Omega_k \, \overline{\Delta T}_{k,\min} \sum_{\ell=1}^{j} \frac{1}{\ell!} (1 + \chi_{p-\ell}(1)) \\
&\qquad \leq \Omega_k \, \overline{\Delta T}_{m_k,j}(S_k, D),
\end{aligned}$$

as desired. To prove the second assertion, observe that AS.1 implies that $\|\nabla_x^\ell f(X_k)\| \leq L_{f,\ell-1}$ for $j \in \{2, \ldots, p\}$, and thus, using (2.27), that, for $\ell \in \{2, \ldots, p\}$,

$$\begin{aligned}
\|\overline{\nabla_x^\ell f}(X_k)\| &\leq \|\nabla_x^\ell f(X_k)\| + \|\overline{\nabla_x^\ell f}(X_k) - \nabla_x^\ell f(X_k)\| \\
&\leq L_{f,\ell-1} + \Theta_0.
\end{aligned}$$

This gives the desired conclusion with the choice $\Theta = \max_{\ell \in \{2,\ldots,p\}} L_{f,\ell-1} + \Theta_0$. □

Of course, the conditions stated in Lemma 2.3 are sufficient but by no means necessary to ensure $\mathcal{M}_k$. In particular, they make no attempt to exploit a possible favourable balance between the errors made on derivatives at different degrees, nor do they take into account that $\mathcal{M}_k^{(1)}$, $\mathcal{M}_{k,j}^{(2)}$ and $\mathcal{M}_{k,j}^{(3)}$ only specify conditions on model accuracy in a finite, dimension-independent subset of directions. Despite these limitations, (2.26) and (2.27) allow the crucial conclusion that $\mathcal{M}_k$ does occur if the derivatives $\overline{\nabla_x^j f}(X_k)$ are sufficiently accurate compared to the model decrease. Moreover, since one would expect that, as an approximate minimizer is approached, $\|S_k\|$, $\|D_{k,j}\|$ and $\|\overline{D}_{k,j}\|$ (and thus τ_k) become small, they also show the accuracy requirement becomes looser for derivatives of higher degree.
We now formalize our assumption on the stochastic process generated by the SAR*qp* algorithm.

AS.3

For all $k \geq 0$, the event $\mathcal{M}_k$ satisfies the condition

$$p_{\mathcal{M},k} = \mathbb{P}\mathrm{r}(\mathcal{M}_k | \mathcal{A}_{k-1}^M) = \mathbb{E}[\mathbb{1}_{\mathcal{M}_k} | \mathcal{A}_{k-1}^M] \geq p_* \tag{2.28}$$

for some $p_* \in (\frac{1}{2}, 1]$ independent of k.

We observe that the introduction of Ω_k in the definitions of $\mathcal{M}_k$ is similar in spirit to the use of the trust-region radius in the definition of fully linear or quadratic models in [3, 6, 7] or of the steplength in [12]. However, the definition of $\mathcal{M}_k$ does not require the model to be "linearly/quadratically" accurate everywhere in a ball around x_k of radius at least $\|s_k\|$, but merely that their variation is accurate enough along s_k (as specified in $\mathcal{M}_k^{(1)}$) and along $d_{k,j}$ and $\overline{d}_{k,j}$ (as specified in $\mathcal{M}_{k,j}^{(2)}$ and $\mathcal{M}_{k,j}^{(3)}$)[(1)] for all $j \in \{1, \ldots, q\}$. The need to consider $\mathcal{M}_{k,j}^{(2)}$ and $\mathcal{M}_{k,j}^{(3)}$ for $j \in \{1, \ldots, q\}$ in the definition of $\mathcal{M}_k$ results from our insistence that q-th order approximate optimality must include j-th order approximate optimality for all such j. AS.3 also parallels assumptions in [3, 6, 7, 12] where accuracy in function values is measured using the guaranteed model decrease or proxies given by the $(p+1)$-st power of the trust-region radius or the steplength. Finally, the conditions imposed by $\mathcal{M}_{k,j}^{(2)}$ and $\mathcal{M}_{k,j}^{(3)}$ are only used whenever considering the value of $\overline{\phi}_{m_k,j}^{\delta_{k,j}}(s_k)$, that is in Lemma 3.1, itself only called upon in Lemma 3.3 in the case where $\|S_k\| \leq 1$. As a consequence, they are irrelevant when long steps are taken ($\|S_k\| > 1$).

3 Worst-case evaluation complexity

Having set the stage and stated our assumptions, we may now consider the worst-case evaluation complexity of the SARqp algorithm. Our aim is to derive a bound on the expected number of iterations $\mathbb{E}(N_\epsilon)$ which is needed, in the worst-case, to reach an (ϵ, δ)-approximate q-th-order-necessary minimizer. Specifically, N_ϵ is the number of iterations required until (2.6) holds for the first time, i.e.,

$$N_\epsilon = \inf \left\{ k \geq 0 \ \mid \ \phi_{f,j}^{\Delta_{k-1,j}}(X_k) \leq \epsilon_j \frac{\Delta_{k-1,j}^j}{j!} \ \text{ for } \ j \in \{1, \ldots, q\} \right\}. \tag{3.1}$$

Note that $\phi_{f,j}^{\Delta_{k-1,j}}(X_k)$, the j-th order optimality measure at iteration k, uses the optimality radiuses $\Delta_{k-1,j}$ resulting from the step computation at iteration $k-1$, as is the case in [1, 4]. Now recall that the trial point $X_{k-1} + S_{k-1}$ and the vector of radii Δ_{k-1} are deterministic once the inexact model at iteration $k-1$ is known. Thus these variables are measurable for $\mathcal{A}_{k-1}^M$ and because of our deterministic assumptions on the accuracy of f, the event $\{X_k = X_{k-1} + S_{k-1}\}$ (which occur when iteration $k-1$ is successful) is also measurable for $\mathcal{A}_{k-1}^M$. As a consequence and since $\phi_{f,j}^{\Delta_{k-1,j}}(X_k)$ uses exact derivatives of f, the event $\{N_\epsilon = k\}$ is measurable with respect to $\mathcal{A}_{k-1}^M$. The definition (3.1) can thus be viewed as that of a family of ϵ-dependent stopping times for the stochastic process generated by the SARqp algorithm (see, e.g., [6, section 2.3]).

3.1 General properties of the SARqp algorithm

We first consider properties of "accurate" iterations, in the sense that $\mathcal{M}_k$ occurs, and start with the relation between $\phi_{m_k,j}^{\delta_{k,j}}(s_k)$ and its approximation. The next lemma is inspired by Lemma 3.2 in [1], but significantly differs in that it now requires considering both directions $d_{k,j}$ and $\overline{d}_{k,j}$.

[(1)] A slightly stronger assumption would be to require a sufficient relative accuracy along s_k and in a (typically small) neighbourhood of s_k.

Lemma 3.1 Consider any realization of the algorithm and assume that $\mathcal{M}_k$ occurs. Then, for $j \in \{1, \ldots, q\}$,

$$\big(1-\omega_k\big)\overline{\phi}^{\delta_{k,j}}_{m_k,j}(s_k) \leq \phi^{\delta_{k,j}}_{m_k,j}(s_k) \leq \big(1+\omega_k\big)\overline{\phi}^{\delta_{k,j}}_{m_k,j}(s_k) \tag{3.2}$$

Proof. Let $j \in \{1, \ldots, q\}$. Consider $d_{k,j}$ defined in (2.21). From (2.17), we have that

$$\begin{aligned}
\Delta T_{m_k,j}(s_k, d_{k,j}) &\leq \overline{\Delta T}_{m_k,j}(s_k, d_{k,j}) + |\Delta T_{m_k,j}(s_k, d_{k,j}) - \overline{\Delta T}_{m_k,j}(s_k, d_{k,j})| \\
&\leq \big(1+\omega_k\big)\overline{\Delta T}_{m_k,j}(s_k, d_{k,j}) \\
&\leq \big(1+\omega_k\big) \max_{\|d\|\leq\delta_{k,j}, x_k+s_k+d\in\mathcal{X}} \overline{\Delta T}_{m_k,j}(s_k, d) \\
&= \big(1+\omega_k\big)\overline{\Delta T}_{m_k,j}(s_k, \overline{d}_{k,j})
\end{aligned}$$

where we used the fact that $\mathcal{M}_k$ occurs to derive the second inequality and considered $\overline{d}_{k,j}$ defined in (2.22). Therefore

$$\phi^{\delta_{k,j}}_{m_k,j}(s_k) = \Delta T_{m_k,j}(s_k, d_{k,j}) \leq \big(1+\omega_k\big)\overline{\phi}^{\delta_{k,j}}_{m_k,j}(s_k).$$

This proves the rightmost inequality of (3.2). Similarly, using our assumption that $\mathcal{M}_k$ occurs, we obtain that

$$\begin{aligned}
\Delta T_{m_k,j}(s_k, \overline{d}_{k,j}) &\geq \overline{\Delta T}_{m_k,j}(s_k, \overline{d}_{k,j}) - |\Delta T_{m_k,j}(s_k, \overline{d}_{k,j}) - \overline{\Delta T}_{m_k,j}(s_k, \overline{d}_{k,j})| \\
&\geq \big(1-\omega_k\big)\overline{\Delta T}_{m_k,j}(s_k, \overline{d}_{k,j})
\end{aligned}$$

and hence, from (2.7) and (2.17), that

$$\big(1-\omega_k\big)\overline{\phi}^{\delta_{k,j}}_{m_k,j}(s_k) \leq \max_{\|d\|\leq\delta_{k,j}, x_k+s_k+d\in\mathcal{X}} \Delta T_{m_k,j}(s_k, d) = \phi^{\delta_{k,j}}_{m_k,j}(s_k),$$

which concludes the proof of (3.2). □

The next step is to adapt an important property of $\Delta_{k,j}$ in the exact case to our inexact framework.

Lemma 3.2 Suppose that AS.1 holds. Then, for any $j \in \{1, \ldots, q\}$,

1. if $j = 1$ and $\mathcal{X}$ is convex or if $j = 2$ and $\mathcal{X} = \Re^n$, $\Delta_{k,j}$ can always be chosen equal to one;
2. in the other cases, and assuming that $\mathcal{M}_k$ occurs, then, either $\|s_k\| > 1$ or $\Delta_{k,j} \leq 1$ can be chosen such that
$$\Delta_{k,j} \geq \kappa_\delta(\sigma_k)\epsilon_j, \tag{3.3}$$
where $\kappa_\delta(\sigma) \in (0,1)$ is independent of ϵ and decreasing with σ.

Proof. The proof broadly follows the developments of [5, Lemmas 4.3 and 4.4], except that it now uses the model involving approximate derivatives and that L_f, the upper bound of the derivatives of f at x_k derived from AS.1 is now replaced by Θ, as guaranteed by $\mathcal{M}_k^{(4)}$. The details are provided in appendix. □

In what follows, we will assume that, whenever $q > 2$ or $\mathcal{X}$ is nonconvex or $q = 2$ and $\mathcal{X} \subset \Re^n$, the SAR$qp$ algorithm computes a pair (s_k, δ_k) such that, for each $j \in \{1, \ldots, q\}$, $\delta_{k,j}$ is always within a fraction of its maximal value, thereby ensuring (3.3). We now prove a crucial inequality relating the step length to the accuracy requirements.

Lemma 3.3 Consider any realization of the algorithm. Assume that $\mathcal{M}_k$ occurs, that iteration k is successful and that, for some $j \in \{1, \ldots, q\}$, (2.6) fails for $(x_{k+1}, \delta_{k,j})$. Then either $\|s_k\| > 1$ or

$$(1 - 2\theta)\epsilon_j \frac{\delta_{k,j}^j}{j!} \leq \frac{L_{f,p} + \sigma_k}{(p - q + 1)!} \sum_{\ell=1}^{j} \frac{\delta_{k,j}^\ell}{\ell!} \|s_k\|^{p-\ell+1} \tag{3.4}$$

Proof. [See [5, Lemma 5.3] for the composite unconstrained Lipschitz continuous case.] Suppose that $\|s_k\| \leq 1$. Since (2.6) fails at $(x_{k+1}, \delta_{k,j})$, we must have that

$$\phi_{f,j}^{\delta_{k,j}}(x_{k+1}) > \epsilon_j \frac{\delta_{k,j}^j}{j!} > 0 \tag{3.5}$$

for some $j \in \{1, \ldots, q\}$. Define d to be the argument of the minimum in the definition of $\phi_{f,j}^{\delta_{k,j}}(x_{k+1})$. Hence,

$$0 < \|d\| \leq \delta_{k,j} \tag{3.6}$$

and $x_k + d \in \mathcal{X}$. Using (3.5), (2.7) and the triangle inequality, we thus obtain that

$$\phi_{f,j}^{\delta_{k,j}}(x_{k+1}) = \Delta T_{f,j}(x_{k+1}, d) \leq |\Delta T_{f,j}(x_{k+1}, d) - \Delta T_{m_k,j}(s_k, d)| + \Delta T_{m_k,j}(s_k, d). \tag{3.7}$$

Recalling now from [4, Lemma 2.4]) that

$$\|\nabla_s^\ell \|s_k\|^{p+1}\| = \frac{(p+1)!}{(p-\ell+1)!} \|s_k\|^{p-\ell+1},$$

we may now use the fact that $x_{k+1} = x_k + s_k$ since iteration k is successful, (2.5) in Lemma 2.1, (2.11), (3.6) and the triangle inequality to obtain that

$$\begin{aligned} |\Delta T_{f,j}(x_{k+1}, d) - \Delta T_{m_k,j}(s_k, d)| &\leq \sum_{\ell=1}^{j} \frac{\delta_{k,j}^\ell}{\ell!} \|\nabla_x^\ell f(x_{k+1}) - \nabla_s^\ell T_{f,p}(x_k, s_k)\| \\ &\quad + \frac{\sigma_k}{(p+1)!} \sum_{\ell=1}^{j} \frac{\delta_{k,j}^\ell}{\ell!} \|\nabla_s^\ell \|s_k\|^{p+1}\| \qquad (3.8) \\ &\leq \frac{L_{f,p} + \sigma_k}{(p - q + 1)!} \sum_{\ell=1}^{j} \frac{\delta_{k,j}^\ell}{\ell!} \|s_k\|^{p-\ell+1} \end{aligned}$$

Moreover, using (2.17), (3.2) and the fact that $\omega_k \leq 1$ (see (2.13)), we deduce that

$$\Delta T_{m_k,j}(s_k,d) \leq \phi^{\delta_{k,j}}_{m_k,j}(s_k) \leq \big(1+\omega_k\big)\overline{\phi}^{\delta_{k,j}}_{m_k,j}(s_k) \leq 2\theta\epsilon_j\frac{\delta^j_{k,j}}{j!}. \tag{3.9}$$

Substituting (3.8) and (3.9) into (3.7) and using (3.6) and (3.5), we obtain (3.4). □

Lemma 3.4 Suppose that AS.1 holds and consider any realization of the algorithm. Suppose also that $\mathcal{M}_k$ occurs, that iteration k is successful and that, for some $j \in \{1,\ldots,q\}$, (2.6) fails for $(x_{k+1},\delta_{k,j})$. Then

$$\|s_k\|^{p+1} \geq \psi(\sigma_k)\epsilon_j^{\pi} \tag{3.10}$$

where

$$\pi = \begin{cases} \dfrac{p+1}{p-q+1} & \text{if } q=1 \text{ and } \mathcal{X} \text{ is convex or if } q=2 \text{ and } \mathcal{X}=\Re^n, \\ \dfrac{q(p+1)}{p} & \text{otherwise.} \end{cases} \tag{3.11}$$

and

$$\psi(\sigma) = \begin{cases} \min\left[1,\left(\dfrac{(1-2\theta)(p-q+1)!}{q!(L_{f,p}+\sigma)}\right)^{\pi}\right] & \text{if } q=1 \text{ and } \mathcal{X} \text{ is convex, or} \\ & \text{if } q=2 \text{ and } \mathcal{X}=\Re^n, \\ \min\left[1,\left(\dfrac{(1-2\theta)(p-q+1)!\,\kappa_\delta(\sigma)^{q-1}}{q!(L_{f,p}+\sigma)}\right)^{\pi}\right] & \text{otherwise.} \end{cases} \tag{3.12}$$

Proof. [See [5, Lemma 5.4] for the unconstrained case.] If $\|s_k\| > 1$, the conclusion immediately follows. Suppose therefore that $\|s_k\| \leq 1$ and consider j such that (3.4) holds. Recalling the definition of χ_j in (1.4), (3.4) can be rewritten as

$$\alpha_k\,\epsilon_j\,\delta^j_{k,j} \leq \|s_k\|^{p+1}\chi_j\left(\frac{\delta_{k,j}}{\|s_k\|}\right) \tag{3.13}$$

where we have set

$$\alpha_k = \frac{(1-2\theta)(p-q+1)!}{q!(L_{f,p}+\sigma_k)}.$$

In particular, since $\chi_j(t) \leq 2t^j$ for $t \geq 1$, we have that, when $\|s_k\| \leq \delta_{k,j}$,

$$\alpha_k\,\epsilon_j \leq 2\|s_k\|^{p+1}\left(\frac{1}{\|s_k\|}\right)^j = 2\|s_k\|^{p-j+1}. \tag{3.14}$$

Suppose first that $q=1$ and $\mathcal{X}$ is convex or $q=2$ and $\mathcal{X}=\Re^n$. Then, from our assumptions and Lemma 3.2, $\delta_{k,j}=1$ and $\|s_k\| \leq 1 = \delta_{k,j}$. Thus (3.14) yields the first case of (3.11)–(3.12). Suppose now that $q>2$ or that $\mathcal{X}$ is not convex or that $q=2$

and $\mathcal{X} \subset \Re^n$. Then our assumptions imply that (3.3) holds. If $\|s_k\| \leq \delta_{k,j}$, we may again deduce from (3.14) that the first case of (3.11)–(3.12) holds, which implies, because $\kappa_\delta(\sigma) < 1$ and $1/(p-j+1) \leq j/p$, that the second case also holds. Consider therefore the case where $\|s_k\| > \delta_{k,j}$. Then (3.13) and the fact that $\chi_j(t) < 2t$ for $t \in [0,1]$ give that

$$\alpha_k \, \epsilon_j \, \delta_{k,j}^j \leq 2\|s_k\|^{p+1} \left(\frac{\delta_{k,j}}{\|s_k\|} \right),$$

which, with (3.3), implies the second case of (3.11)–(3.12) as requested. □

Note that $\psi(\sigma)$ is decreasing as a function of σ in both cases of (3.12). We now investigate the decrease of the exact objective function values at successful iterations.

Lemma 3.5 Suppose that AS.1 holds and consider any realization of the algorithm. Then

$$\overline{\Delta T}_{f,p}(x_k, s_k) \geq \frac{\sigma_k}{(p+1)!}\|s_k\|^{p+1} \geq \frac{\sigma_{\min}}{(p+1)!}\|s_k\|^{p+1} \geq 0. \tag{3.15}$$

Moreover, if iteration k is successful, then

$$f(x_k) - f(x_{k+1}) \geq \frac{\eta\sigma_{\min}(1-\alpha)}{(p+1)!}\|s_k\|^{p+1} \geq 0. \tag{3.16}$$

Proof. The inequality (3.15) immediately follows from (2.12), (2.16), (2.19). Now the fact that iteration k is successful, together with (2.13) and (2.14)–(2.15), imply that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \overline{f}(x_k) - \overline{f}(x_{k+1}) - 2\omega_k \overline{\Delta T}_{f,p}(x_k, s_k) \\ &\geq \eta\overline{\Delta T}_{f,p}(x_k, s_k) - \alpha\eta\overline{\Delta T}_{f,p}(x_k, s_k), \end{aligned}$$

yielding (3.16) using (3.15). □

We finally conclude our analysis of "accurate" iterations by proving a standard result in the analysis of adaptive regularization methods. A similar version of this result was presented in [1, Lemma 4.2] for the case where both function values and models are sufficiently accurate.

Lemma 3.6 Suppose that AS.1 holds. For any realization of the algorithm, if iteration k is such that $\mathcal{M}_k$ occurs and

$$\sigma_k \geq \sigma_s \stackrel{\rm def}{=} \max\left[\sigma_0, \frac{L_{f,p}+3}{1-\eta}\right], \tag{3.17}$$

then iteration k is successful.

Proof. Suppose that (3.17) holds. Thus, using successively (2.18), the triangle inequality, the fact that $\mathcal{M}_k$ occurs, (2.4), (3.15), (2.13), (2.14)–(2.15) and (3.17), we deduce

that

$$\begin{aligned}
|\rho_k - 1| &\leq \frac{1}{\overline{\Delta T}_{f,p}(x_k, s_k)} \Big[\Big(\overline{f}(x_k) - f(x_k) \Big) + \Big(f(x_k + s_k) - \overline{f}(x_k + s_k) \Big) \\
&\qquad + \Big(- f(x_k + s_k) + f(x_k) - \Delta T_{f,p}(x_k, s_k) \Big) \\
&\qquad + \Big(\Delta T_{f,p}(x_k, s_k) - \overline{\Delta T}_{f,p}(x_k, s_k) \Big) \Big] \\
&\leq \frac{1}{\overline{\Delta T}_{f,p}(x_k, s_k)} \left[|f(x_k + s_k) - T_{f,p}(x_k, s_k)| + 3\omega_k |\overline{\Delta T}_{f,p}(x_k, s_k)| \right] \\
&\leq \frac{1}{\overline{\Delta T}_{f,p}(x_k, s_k)} \left[\frac{L_{f,p}}{(p+1)!} \|s_k\|^{p+1} + \frac{3|\overline{\Delta T}_{f,p}(x_k, s_k)|}{\sigma_k} \right] \\
&\leq \frac{L_{f,p}}{\sigma_k} + \frac{3}{\sigma_k} \\
&\leq 1 - \eta.
\end{aligned}$$

Therefore $\rho_k \geq \eta$ and iteration k is successful. □

3.2 Bounding the expected number of steps with $\Sigma_k \geq \sigma_s$

We now return to the general stochastic process generated by the SAR*qp* algorithm aiming at bounding from above the expected number of steps in the process generated by the algorithm with $\Sigma_k \geq \sigma_s$. To this purpose, for all $0 \leq k \leq \ell$, given $\ell \in \{0, \ldots, N_\epsilon - 1\}$, let us define the events

$$\Lambda_k = \{ \text{ iteration } k \text{ is such that } \Sigma_k < \sigma_s \ \}, \qquad \Lambda_k^c = \{ \text{ iteration } k \text{ is such that } \Sigma_k \geq \sigma_s \ \}$$
$$\mathcal{S}_k = \{ \text{ iteration } k \text{ is successful } \},$$

and let

$$N_\Lambda \stackrel{\rm def}{=} \sum_{k=0}^{N_\epsilon - 1} \mathbb{1}_{\Lambda_k}, \qquad N_{\Lambda^c} \stackrel{\rm def}{=} \sum_{k=0}^{N_\epsilon - 1} \mathbb{1}_{\Lambda_k^c}, \tag{3.18}$$

be the number of steps, in the stochastic process induced by the SAR*qp* algorithm, with $\Sigma_k < \sigma_s$ and $\Sigma_k \geq \sigma_s$, before N_ϵ is met, respectively. In what follows we suppose that AS.1–AS.3 hold, as well the assumption below.

AS.4

With reference to the stochastic process generated by the SAR*qp* algorithm and the definition of σ_s in (3.17), we assume that

$$\Sigma_0 = \gamma^{-i} \sigma_s, \tag{3.19}$$

for some positive integer i.

We may now follow the argument of [6] to derive an upper bound on $\mathbb{E}\big[N_{\Lambda^c}\big]$. In particular, the argument unfolds as follows:

(i) we apply [6, Lemma 2.2] to deduce that, for any $\ell \in \{0, \ldots, N_\epsilon - 1\}$ and for all realizations of the SAR*qp* algorithm, one has that

$$\sum_{k=0}^{\ell} \mathbb{1}_{\Lambda_k^c} \mathbb{1}_{\mathcal{S}_k} \leq \frac{\ell + 1}{2}; \tag{3.20}$$

(ii) as in [6], we note that both $\hat{\sigma}(\mathbb{1}_{\Lambda_k})$ and $\hat{\sigma}(\mathbb{1}_{\Lambda_k^c})$ belong to $\mathcal{A}_{k-1}^M$, as the random variable Λ_k is fully determined by the first $k-1$ iterations of the SARqp algorithm. Then, setting $\ell = N_\epsilon - 1$ we can rely on [6, Lemma 2.1] (with $W_k = \mathbb{1}_{\Lambda_k^c}$) and (2.28) to deduce that

$$\mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\Lambda_k^c}\mathbb{1}_{\mathcal{M}_k}\right] \geq \mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1} p_{\mathcal{M},k}\mathbb{1}_{\Lambda_k^c}\right] \geq p_* \,\mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\Lambda_k^c}\right]; \tag{3.21}$$

(iii) as a consequence, given that Lemma 3.6 ensures that each iteration k where $\mathcal{M}_k$ occurs and $\sigma_k \geq \sigma_s$ is successful, we have that

$$\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\Lambda_k^c}\mathbb{1}_{\mathcal{M}_k} \leq \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\Lambda_k^c}\mathbb{1}_{\mathcal{S}_k} \leq \frac{N_\epsilon}{2},$$

in which the last inequality follows from (3.20), with $\ell = N_\epsilon - 1$. Taking expectation in the above inequality, using (3.21) and recalling the rightmost definition in (3.18), we obtain, as in [6, Lemma 2.3], that, for any realization,

$$\mathbb{E}[N_{\Lambda^c}] \leq \frac{1}{2p_*}\mathbb{E}[N_\epsilon]. \tag{3.22}$$

The remaining upper bound on $\mathbb{E}[N_\Lambda]$ will be the focus of the next section.

3.3 Bounding the expected number of steps with $\Sigma_k < \sigma_s$

For analyzing $\mathbb{E}[N_\Lambda]$, where N_Λ is defined in (3.18), we now introduce the following variables.

Definition 1 *With reference to the process* (2.20) *generated by the* SARqp *algorithm, let us define:*

- $\overline{\Lambda}_k$ = {*iteration k is such that $\Sigma_k \leq \sigma_s$*};
- $N_I = \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{M}_k^c}$: *the number of inaccurate iterations with $\Sigma_k \leq \sigma_s$*;
- $N_A = \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{M}_k}$: *the number of accurate iterations with $\Sigma_k \leq \sigma_s$*;
- $N_{AS} = \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{M}_k}\mathbb{1}_{\mathcal{S}_k}$: *the number of accurate successful iterations with $\Sigma_k \leq \sigma_s$*;
- $N_{AU} = \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\Lambda_k}\mathbb{1}_{\mathcal{M}_k}\mathbb{1}_{\mathcal{S}_k^c}$: *the number of accurate unsuccessful iterations with $\Sigma_k < \sigma_s$*;
- $N_{IS} = \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{M}_k^c}\mathbb{1}_{\mathcal{S}_k}$: *the number of inaccurate successful iterations with $\Sigma_k \leq \sigma_s$*;
- $N_S = \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{S}_k}$: *the number of successful iterations with $\Sigma_k \leq \sigma_s$*;
- $N_U = \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\Lambda_k}\mathbb{1}_{\mathcal{S}_k^c}$: *the number of unsuccessful iterations with $\Sigma_k < \sigma_s$*.

(3.23)

Observe that $\overline{\Lambda}_k$ is the "closure" of Λ_k in that the inequality in its definition is no longer strict.

We immediately notice that an upper bound on $\mathbb{E}[N_\Lambda]$ is available, once an upper bound on $\mathbb{E}[N_I] + \mathbb{E}[N_A]$ is known, since

$$\mathbb{E}[N_\Lambda] \leq \mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\right] = \mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{M}_k^c} + \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{M}_k}\right] = \mathbb{E}[N_I] + \mathbb{E}[N_A]. \tag{3.24}$$

Using again [6, Lemma 2.1] (with $W_k = \mathbb{1}_{\overline{\Lambda}_k}$) to give an upper bound on $\mathbb{E}[N_I]$, we obtain the following result.

Lemma 3.7 [6, Lemma 2.6] Let $\mathcal{M}_k$ be the sequence of events in (2.23) and assume that (2.28) holds. Let N_I, N_A be defined as in Definition 1 in the context of the stochastic process (2.20) generated by the SAR*qp* algorithm. Then

$$\mathbb{E}[N_I] \leq \frac{1-p_*}{p_*}\,\mathbb{E}[N_A]. \tag{3.25}$$

Turning to the upper bound for $\mathbb{E}[N_A]$, we observe that

$$\mathbb{E}[N_A] = \mathbb{E}[N_{AS}] + \mathbb{E}[N_{AU}] \leq \mathbb{E}[N_{AS}] + \mathbb{E}[N_U]. \tag{3.26}$$

Hence, bounding $\mathbb{E}[N_I]$ can be achieved by providing upper bounds on $\mathbb{E}[N_{AS}]$ and $\mathbb{E}[N_U]$. Regarding the latter, we first note that the process induced by the SAR*qp* algorithm ensures that Σ_k is decreased by a factor γ on successful steps and increased by the same factor on unsuccessful ones, provided that (3.19) holds. Consequently, by virtue of [6, Lemma 2, 5], we obtain the following bound.

Lemma 3.8 [6, Lemma 2.5] Suppose that AS.4 holds. For any $\ell \in \{0, ..., N_\epsilon - 1\}$ and for all realisations of the SAR*qp* algorithm, we have that

$$\sum_{k=0}^{\ell} \mathbb{1}_{\Lambda_k}\mathbb{1}_{\mathcal{S}_k^c} \leq \sum_{k=0}^{\ell} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{S}_k} + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right).$$

From this inequality with $\ell = N_\epsilon - 1$, recalling Definition 1 and taking expectations, we therefore obtain that

$$\mathbb{E}[N_U] \leq \mathbb{E}[N_S] + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right) = \mathbb{E}[N_{AS}] + \mathbb{E}[N_{IS}] + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right). \tag{3.27}$$

An upper bound on $\mathbb{E}[N_{AS}]$ is given by the following lemma.

Lemma 3.9 Let Assumption AS.1 and AS.2 hold. For all realizations of the SARqp algorithm we have that

$$\mathbb{E}[N_{AS}] \leq \frac{(f_0 - f_{\rm low})(p+1)!}{\eta\sigma_{\min}(1-\alpha)\psi(\sigma_s)} \left(\min_{j\in\{1,\ldots,q\}} \epsilon_j\right)^{-\pi} + 1, \tag{3.28}$$

where π, $\psi(\sigma)$ and σ_s are defined in (3.11), (3.12) and (3.17), respectively.

Proof. For all realizations of the SARqp algorithm we have that:

- if iteration k is successful, then (3.16) holds;
- if iteration k is successful and accurate (i.e., $\mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{M}_k} = 1$) and (2.6) fails for $(x_{k+1}, \delta_{k,j})$, then (3.10) holds;
- if iteration k is unsuccessful, the mechanism of the SARqp algorithm guarantees that $x_k = x_{k+1}$ and, hence, that $f(x_{k+1}) = f(x_k)$.

Therefore, for any $\ell \in \{0, ..., N_\epsilon - 1\}$,

$$\begin{aligned}
f_0 - f_{\rm low} &\geq f_0 - f(X_{\ell+1}) = \sum_{k=0}^{\ell} \mathbb{1}_{\mathcal{S}_k}(f(X_k) - f(X_{k+1})) \geq \sum_{k=0}^{\ell} \mathbb{1}_{\mathcal{S}_k} \frac{\eta\sigma_{\min}(1-\alpha)}{(p+1)!}\|S_k\|^{p+1} \\
&\geq \sum_{k=0}^{\ell-1} \mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{M}_k} \frac{\eta\sigma_{\min}(1-\alpha)}{(p+1)!}\|S_k\|^{p+1} \qquad (3.29) \\
&\geq \sum_{k=0}^{\ell-1} \mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{M}_k} \frac{\eta\sigma_{\min}(1-\alpha)}{(p+1)!}\psi(\Sigma_k)\left(\min_{j\in\{1,\ldots,q\}} \epsilon_j\right)^{\pi} \\
&\geq \sum_{k=0}^{\ell-1} \mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{M}_k}\mathbb{1}_{\overline{\Lambda}_k} \frac{\eta\sigma_{\min}(1-\alpha)}{(p+1)!}\psi(\Sigma_k)\left(\min_{j\in\{1,\ldots,q\}} \epsilon_j\right)^{\pi} \\
&\geq \frac{\eta\sigma_{\min}(1-\alpha)}{(p+1)!}\psi(\sigma_s)\left(\min_{j\in\{1,\ldots,q\}} \epsilon_j\right)^{\pi} \left(\sum_{k=0}^{\ell-1} \mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{M}_k}\mathbb{1}_{\overline{\Lambda}_k}\right), \qquad (3.30)
\end{aligned}$$

having set $f_0 \stackrel{\rm def}{=} f(X_0)$ and where the last inequality is due to fact that $\psi(\sigma)$ is a decreasing function. We now notice that, by Definition 1,

$$N_{AS} - 1 \leq \sum_{k=0}^{N_\epsilon - 2} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{M}_k}\mathbb{1}_{\mathcal{S}_k}.$$

Hence, letting $\ell = N_\epsilon - 1$ and taking expectations in (3.30), we conclude that

$$f_0 - f_{\rm low} \geq (\mathbb{E}[N_{AS}] - 1)\frac{\eta\sigma_{\min}(1-\alpha)}{(p+1)!}\psi(\sigma_s)\left(\min_{j\in\{1,\ldots,q\}} \epsilon_j\right)^{\pi},$$

which is equivalent to (3.28). □

While inequalities (3.27) and (3.28) provide upper bounds on $\mathbb{E}[N_{AS}]$ and $\mathbb{E}[N_U]$, as desired, the latter still depends on $\mathbb{E}[N_{IS}]$, which has to be bounded from above as well. This can be done by following [6] once more: Definition 1, (3.25) and (3.26) directly imply that

$$\mathbb{E}[N_{IS}] \leq \mathbb{E}[N_I] \leq \frac{1-p_*}{p_*}\mathbb{E}[N_A] \leq \frac{1-p_*}{p_*}\left(\mathbb{E}[N_{AS}] + \mathbb{E}[N_U]\right) \tag{3.31}$$

and hence

$$\mathbb{E}[N_{IS}] \leq \frac{1-p_*}{2p_*-1}\left(2\mathbb{E}[N_{AS}] + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right)\right) \tag{3.32}$$

follows from (3.27) (remember that $\frac{1}{2} < p_* \leq 1$). Thus, the right-hand side in (3.25) is in turn upper bounded by virtue of (3.26), (3.27), (3.32) and (3.28), giving

$$\begin{aligned}
\mathbb{E}[N_A] &\leq \mathbb{E}[N_{AS}] + \mathbb{E}[N_U] \leq 2\mathbb{E}[N_{AS}] + \mathbb{E}[N_{IS}] + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right) \\
&\leq \left(\frac{1-p_*}{2p_*-1}+1\right)\left(2\mathbb{E}[N_{AS}] + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right)\right) \\
&= \frac{p_*}{2p_*-1}\left[2\mathbb{E}[N_{AS}] + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right)\right] \\
&\leq \frac{p_*}{2p_*-1}\left[\frac{2(f_0-f_{\rm low})(p+1)!}{\eta\sigma_{\min}(1-\alpha)\psi(\sigma_s)}\left(\min_{j\in\{1,\ldots,q\}}\epsilon_j\right)^{-\pi} + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right) + 2\right],
\end{aligned} \tag{3.33}$$

This inequality, together with (3.24) and (3.25), finally gives the desired bound on $\mathbb{E}[N_\Lambda]$:

$$\mathbb{E}[N_\Lambda] \leq \frac{1}{p_*}\mathbb{E}[N_A] \leq \frac{1}{2p_*-1}\left[\frac{2(f_0-f_{\rm low})(p+1)!}{\eta\sigma_{\min}(1-\alpha)\psi(\sigma_s)}\left(\min_{j\in\{1,\ldots,q\}}\epsilon_j\right)^{-\pi} + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right) + 2\right]. \tag{3.34}$$

We can now express our final complexity result in full.

Theorem 3.10 Suppose that AS.1–AS.4 hold. Then the following conclusions also hold.

1. If $q = 1$ and $\mathcal{X}$ is convex or if $q = 2$ and $\mathcal{X} = \Re^n$, then

$$\mathbb{E}[N_\epsilon] \leq \kappa(p_*)\left(\frac{2(f_0-f_{\rm low})(p+1)!}{\eta\sigma_{\min}(1-\alpha)\psi(\sigma_s)}\left(\min_{j\in\{1,\ldots,q\}}\epsilon_j\right)^{-\frac{p+1}{p-q+1}} + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right) + 2\right),$$

2. If $q > 2$ or $\mathcal{X}$ is non-convex or $q = 2$ and $\mathcal{X} \subset \Re^n$, then

$$\mathbb{E}[N_\epsilon] \leq \kappa(p_*)\left(\frac{2(f_0-f_{\rm low})(p+1)!}{\eta\sigma_{\min}(1-\alpha)\psi(\sigma_s)}\left(\min_{j\in\{1,\ldots,q\}}\epsilon_j\right)^{-\frac{q(p+1)}{p}} + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right) + 2\right),$$

with $\kappa(p_*) \stackrel{\rm def}{=} \frac{2p_*}{(2p_*-1)^2}$ and N_ϵ, $\psi(\sigma)$, σ_s defined as in (3.1), (3.12), (3.17), respectively.

Proof. Recalling the definitions (3.18) and the bound (3.22), we obtain that

$$\mathbb{E}[N_\epsilon] = \mathbb{E}[N_\Lambda^c] + \mathbb{E}[N_\Lambda] \leq \frac{\mathbb{E}[N_\epsilon]}{2p_*} + \mathbb{E}[N_\Lambda],$$

which implies, using (3.34), that

$$\frac{2p_*-1}{2p_*}\mathbb{E}[N_\epsilon] \leq \frac{1}{2p_*-1}\left(\frac{2(f_0-f_{\rm low})(p+1)!}{\eta\sigma_{\min}(1-\alpha)\psi(\sigma_s)}\left(\min_{j\in\{1,\ldots,q\}}\epsilon_j\right)^{-\pi} + \log_\gamma\left(\frac{\sigma_s}{\sigma_0}\right)+2\right).$$

This bound and the inequality $\frac{1}{2} < p_* \leq 1$ yield the desired result. □

Since the SARqp algorithm requires at most two function evaluations and one evaluation of the derivatives of orders one to p per iteration, the bounds stated in the above theorem effectively provide an upper bound on the average evaluation complexity of finding (ϵ,δ)-approximate q-th order minimizers. Theorem 3.10 generalizes the complexity bounds stated in [5, Theorem 5.5] to the case where evaluations of f and its derivatives are inexact, under probabilistic assumptions on the accuracies of the latter. As it was shown in [5, Theorems 6.1 and 6.4] that the evaluation complexity bounds are sharp in order of the tolerance ϵ for exact evaluations and Lipschitz continuous derivatives of f, this is also the case for the bounds of Theorem 3.10.

4 Conclusions and perspectives

We have shown that the SARqp algorithm, a stochastic inexact adaptive regularization algorithm using derivatives of order up to p, computes an (ϵ,δ)-approximate q-th order minimizer of problem (1.1) in at most $O(\epsilon^{-\frac{p+1}{p-q+1}})$ iterations in expectation if q is either one and $\mathcal{X}$ is convex, or two and the problem is unconstrained, while it may need $O(\epsilon^{-\frac{q(p+1)}{p}})$ iterations in expectation in the other cases[(2)]. Moreover, these bounds are sharp in the order of ϵ (see [5]). We therefore conclude that, if the probabilities $p_{\mathcal{M},k}$ in AS.3 are suitably large, the evaluation complexity of the SARqp algorithm is identical (in order) to that of the exact algorithm in [5].

We also note that the full power of AS.1 is only required for Lemma 3.2, while Lipschitz continuity of $\nabla_x^p f(x)$ is sufficient for all subsequent derivations. Thus if suitable lower bounds on $\Delta_{k,j}$ can be ensured in some other way, our development remains valid (although the precise complexity bounds will depend on the new bounds on $\Delta_{k,j}$). In AS.1, we have also required (Lipschitz) continuity of f and its derivatives in a convex neighbourhood of $\mathcal{X}$. This is somewhat too strong as requiring this assumption only on the "tree of iterates" $\cup_{k\geq 0}[x_k, x_k+s_k]$ is sufficient, but often impossible to verify *a priori*. In the same vein, it also is possible to avoid requiring that (3.3) is always ensured by the SARqp algorithm whenever $q>2$ or $\mathcal{X}$ is nonconvex or $q=2$ and $\mathcal{X}\subset\Re^n$ by instead redefining $\mathcal{M}_k$ to also include the satisfaction of this condition. We have preferred using an explicit assumption because this approach better differentiates deterministic requirements on the algorithm from stochastic assumptions more related to the problem itself.

We finally recall that [5] also derives complexity bounds for the (possibly non-smooth) composite optimization problem. We expect that the theory presented here can be extended to also cover this case.

[(2)] These simplified order bounds assume that $\epsilon_j = \epsilon$ for $j \in \{1,\ldots,q\}$.

An analysis covering adaptive regularization algorithms where the objective function evaluations are also subject to general random noise, parallel to that provided for trust-region methods for low order minimizers in [3], remains for now an open and challenging question.

Acknowledgment

INdAM-GNCS partially supported the first and third authors under Progetti di Ricerca 2019. The second author was partially supported by INdAM through a GNCS grant. The last author gratefully acknowledges the support and friendly environment provided by the Department of Industrial Engineering at the Università degli Studi di Firenze (Italy) during his visit in the fall of 2019.

References

[1] S. Bellavia, G. Gurioli, B. Morini, and Ph. L. Toint. Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM Journal on Optimization*, 29(4):2881–2915, 2019.

[2] A. Berahas, L. Cao, and K. Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. arXiv:1910.04055, 2019.

[3] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust region method via supermartingales. *INFORMS Journal on Optimization*, 1(2):92–119, 2019.

[4] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. *SIAM Journal on Optimization*, 30(1):513–541, 2020.

[5] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Strong evaluation complexity bounds for arbitrary-order optimization of nonconvex nonsmooth composite functions. arXiv:2001.10802, 2020.

[6] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming, Series A*, 159(2):337–375, 2018.

[7] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming, Series A*, 169(2):447–487, 2018.

[8] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.

[9] F. E. Curtis and K. Scheinberg. Adaptive stochastic optimization. arXiv:2001:06699, 2020.

[10] S. Gratton, E. Simon, and Ph. L. Toint. An algorithm for the minimization of nonsmooth nonconvex functions using inexact evaluations and its worst-case complexity. *Mathematical Programming, Series A*, (to appear), 2020.

[11] D. P. Kouri, M. Heinkenscloss, D. Rizdal, and B. G. van Bloemen-Waanders. Inexact objective function evaluations in a trust-region algorithm for PDE-constrained optimization under uncertainty. *SIAM Journal on Scientific Computing*, 36(6):A3011–A3029, 2014.

[12] C. Paquette and K. Scheinberg. A stochastic line search method with convergence rate analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

[13] X. Zhang, C. Ling, and L. Qi. The best rank-1 approximation of a symmetric tensor and related spherical optimization problems. *SIAM Journal on Matrix Analysis*, 33(3):806–821, 2012.

Proof of Lemma 2.2

Let s_k^* be a global minimizer of $m_k(s)$ in $\mathcal{X}$. By Taylor's theorem, we have that, for all d such that $x_k + s_k^* + d \in \mathcal{X}$,

$$\begin{aligned} 0 \;\leq\; m_k(s_k^*+d) - m_k(s_k^*) &= \sum_{\ell=1}^{p} \frac{1}{\ell!} \nabla_s^\ell \overline{T}_{f,p}(x_k, s_k^*)[d]^\ell \\ &\quad + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=1}^{p} \frac{1}{\ell!}\nabla_s^\ell \left(\|s_k^*\|^{p+1}\right)[d]^\ell + \frac{1}{(p+1)!}\nabla_s^{p+1}\left(\|s_k^*+\tau d\|^{p+1}\right)[d]^{p+1}\right] \end{aligned} \tag{A.1}$$

for some $\tau \in (0,1)$. We may now use the expression of $\nabla_s^\ell \left(\|s_k^*\|^{p+1}\right)$ given by [4, Lemma 2.4] in (A.1) and deduce that, for any $j \in \{1,\ldots,q\}$ and all d such that $x_k + s_k^* + d \in \mathcal{X}$,

$$\begin{aligned} &-\sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_s^\ell \overline{T}_{f,p}(x_k, s_k^*)[d]^\ell - \frac{\sigma_k}{(p+1)!}\sum_{\ell=1}^{j} \nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell \\ &\qquad \leq \sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell \overline{T}_{f,p}(x_k, s_k^*)[d]^\ell + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell + \|d\|^{p+1}\right]. \end{aligned} \tag{A.2}$$

It is now possible to choose $\delta_{k,j} \in (0,1]$ such that, for every d with $\|d\| \leq \delta_{k,j}$,

$$\begin{aligned} &\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell \overline{T}_{f,p}(x_k, s_k^*)[d]^\ell + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell + \|d\|^{p+1}\right] \\ &\qquad \leq \frac{1}{2}\theta\epsilon_j \frac{\delta_{k,j}^j}{j!}. \end{aligned} \tag{A.3}$$

We therefore obtain that if $\delta_{k,j}$ is small enough to ensure (A.3), then (A.2) implies that

$$-\sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_s^\ell \overline{T}_{f,p}(x_k, s_k^*)[d]^\ell - \frac{\sigma_k}{(p+1)!}\sum_{\ell=1}^{j} \nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell \leq \frac{1}{2}\theta\epsilon_j \frac{\delta_{k,j}^j}{j!}. \tag{A.4}$$

and therefore that, for all $j \in \{1,\ldots,q\}$,

$$\max_{x_k+s_k^*+d\in\mathcal{X}, \|d\|\leq\delta_{k,j}} \overline{\Delta T}_{m_k,j}(s_k^*, d) \leq \frac{1}{2}\theta\epsilon_j \frac{\delta_{k,j}^j}{j!}.$$

Thus the pair (s_k^*, δ_k) is acceptable for Step 2 of the algorithm. If we assume now that $x_k + s_k^*$ is not an isolated feasible point, the above inequality and continuity of $\overline{T}_{f,p}(x_k, s)$ and its derivatives with respect to s then ensure the existence of a feasible neighbourhood $\mathcal{N}_k^*$ of s_k^* in which

$$\max_{x_k+s+d\in\mathcal{X}, \|d\|\leq\delta_{k,j}} \overline{\Delta T}_{m_k,j}(s, d) \leq \theta\epsilon_j \frac{\delta_{k,j}^j}{j!}. \tag{A.5}$$

for all $s \in \mathcal{N}_k^*$. We may then choose any s_k in $\mathcal{N}_k^*$ such that, in addition to satisfying (A.5) and being such that $x_k + s_k$ is feasible, (2.16) also holds. Thus the definition of $\overline{\phi}_{m_k,j}^{\delta_{k,j}}(s_k)$ in (2.17) gives that

$$\overline{\phi}_{m_k,j}^{\delta_{k,j}}(s_k) \leq \theta\epsilon_j \frac{\delta_{k,j}^j}{j!} \tag{A.6}$$

and every such (s_k, δ_k) is also acceptable for Step 2 of the algorithm.

Proof of Lemma 3.2

Let s_k^* be a global minimizer of $m_k(s)$. We first consider the case where $q = 1$ and $\mathcal{X}$ is convex or $q = 2$ and $\mathcal{X} = \Re^n$. Then it is easy to verify that, for each $j \in \{1, \ldots, q\}$, the optimization problem involved in the definition of $\overline{\phi}_{m_k,j}^{\delta_{k,j}}(s_k^*)$ (in (2.17)) is convex and therefore that $\delta_{k,j}$ can be chosen arbitrarily in $(0, 1]$. The first case of Lemma 3.2 then follows from the continuity of $\overline{\phi}_{m_k,j}^{\delta_{k,j}}(s)$ with respect to s.

In order to prove the second case, we now pursue the reasoning of the proof of Lemma 2.2. We start by supposing that $\|s_k^*\| > 1$. We may then reduce the neighbourhood of s_k^* in which s_k can be chosen enough to guarantee that $\|s_k\| \geq 1$, which then gives the desired result because of (A.5). Suppose therefore that $\|s_k^*\| \leq 1$. The triangle inequality then implies that

$$\|\nabla_s^\ell \overline{T}_{f,p}(x_k, s_k^*)\| \leq \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!} \|\overline{\nabla_x^i f}(x_k)\| \, \|s_k^*\|^{i-\ell},$$

for $\ell \in \{q+1, \ldots, p\}$, and thus, using, AS.1 and [4, Lemma 2.4], we deduce that

$$\begin{aligned}\sum_{\ell=j+1}^{p} \frac{1}{\ell!} \nabla_s^\ell \overline{T}_{f,p}(x_k, s_k^*)[d]^\ell + \frac{\sigma_k}{(p+1)!} \left[\sum_{\ell=j+1}^{p} \nabla_s^\ell \|s_k^*\|^{p+1} [d]^\ell \right] \\ \leq \sum_{\ell=j+1}^{p} \frac{\|d\|^\ell}{\ell!} \left[\sum_{i=\ell}^{p} \frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!} \|\overline{\nabla_x^i f}(x_k)\| + \frac{\sigma_k \|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!} \right].\end{aligned}$$

We now call upon the fact that, since $q \geq 3$ or $\mathcal{X}$ is not convex or $q = 2$ and $\mathcal{X} \subset \Re^n$ and $\mathcal{M}_k$ occurs by assumption, $\mathcal{M}_k^{(4)}$ also occurs. Thus

$$\begin{aligned}\sum_{\ell=j+1}^{p} \frac{1}{\ell!} \nabla_s^\ell \overline{T}_{f,p}(x_k, s_k^*)[d]^\ell + \frac{\sigma_k}{(p+1)!} \left[\sum_{\ell=j+1}^{p} \nabla_s^\ell \|s_k^*\|^{p+1} [d]^\ell \right] \\ \leq \sum_{\ell=j+1}^{p} \frac{\|d\|^\ell}{\ell!} \left[\Theta \sum_{i=\ell}^{p} \frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!} + \frac{\sigma_k \|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!} \right].\end{aligned}$$

We therefore obtain from (A.3) that any pair $(s_k^*, \delta_{s,j})$ satisfies (A.4) for $\|d\| \leq \delta_{s,j}$ if

$$\sum_{\ell=j+1}^{p} \frac{\delta_{s,j}^\ell}{\ell!} \left[\Theta \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!} \|s_k^*\|^{i-\ell} + \frac{\sigma_k \|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!} \right] + \sigma_k \frac{\delta_{s,j}^{p+1}}{(p+1)!} \leq \frac{1}{2} \theta \epsilon_j \frac{\delta_{s,j}^j}{j!}. \tag{A.7}$$

which, because $\|s_k^*\| \leq 1$, is in turn ensured by the inequality

$$\sum_{\ell=j+1}^{p} \frac{\delta_{s,j}^\ell}{\ell!} \left[\Theta \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!} + \sigma_k \right] + \sigma_k \frac{\delta_{s,j}^{p+1}}{(p+1)!} \leq \frac{1}{2} \theta \epsilon_j \frac{\delta_{s,j}^j}{j!}. \tag{A.8}$$

Observe now that, since $\delta_{s,j} \in [0, 1]$, $\delta_{s,j}^\ell \leq \delta_{s,j}^{j+1}$ for $\ell \in \{j+1, \ldots, p\}$. Moreover, we have that,

$$\sum_{i=\ell}^{p} \frac{1}{(i-\ell)!} \leq e < 3, \quad (\ell \in \{j+1, \ldots, p+1\}), \qquad \sum_{\ell=j+1}^{p+1} \frac{1}{\ell!} \leq e - 1 < 2$$

and therefore (A.8) is guaranteed by the condition

$$j!(6\Theta + 2\sigma_k)\,\delta_{s,j} \leq \frac{1}{2}\theta\epsilon_j, \tag{A.9}$$

which means that the pair (s_k^*, δ_s) satisfies (A.4) for all $j \in \{1, \ldots, q\}$ whenever,

$$\delta_{s,j} \leq \frac{1}{2}\delta_{\min,k} \stackrel{\rm def}{=} \frac{\theta\epsilon_j}{2q!(6\Theta + 2\sigma_k)}.$$

As in the proof of Lemma 2.2, we may invoke continuity of the derivatives of $m_k(s)$ with respect to s to deduce that there exists a neighbourhood $\mathcal{N}_k^*$ of s_k^* such that (A.5) holds for every $s \in \mathcal{N}_k^*$ and every $\delta_{k,j} \leq \delta_{\min,k}$. Choosing now s_k to ensure (2.16) and $x_k + s_k \in \mathcal{X}$ in addition to (A.5), we obtain that the pair $(s_k, \delta_{k,j})$ satisfies both (2.16) and

$$\overline{\phi}_{m_k,j}^{\delta_{k,j}}(s_k) \leq \theta\epsilon_j\,\frac{\delta_{k,j}^j}{j!}.$$

The desired conclusion then follows with

$$\kappa_\delta(\sigma) = \frac{\nu\theta}{q!(6\Theta + 2\sigma)}$$

for any constant $\nu \in (0, 1)$. Moreover, $\kappa_\delta(\sigma)$ is clearly a decreasing function of σ.