Trust-region algorithms: Probabilistic complexity and intrinsic noise with applications to subsampling techniques

S. Bellavia [a,*,1], G. Gurioli [b,1], B. Morini [a,1], Ph.L. Toint [c]

[a] *Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, Italy*
[b] *Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo",*
ISTI-CNR, Pisa, Italy
[c] *Namur Center for Complex Systems (naXys), University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium*

A R T I C L E I N F O

A B S T R A C T

A trust-region algorithm is presented for finding approximate minimizers of smooth unconstrained functions whose values and derivatives are subject to random noise. It is shown that, under suitable probabilistic assumptions, the new method finds (in expectation) an ϵ-approximate minimizer of arbitrary order $q \geq 1$ in at most $\mathcal{O}(\epsilon^{-(q+1)})$ inexact evaluations of the function and its derivatives, providing the first such result for general optimality orders. The impact of intrinsic noise limiting the validity of the assumptions is also discussed and it is shown that difficulties are unlikely to occur in the first-order version of the algorithm for sufficiently large gradients. Conversely, should these assumptions fail for specific realizations, then "degraded" optimality guarantees are shown to hold when failure occurs. These conclusions are then discussed and illustrated in the context of subsampling methods for finite-sum optimization.

* Corresponding author.
E-mail addresses: stefania.bellavia@unifi.it (S. Bellavia), gianmarco.gurioli@isti.cnr.it (G. Gurioli), benedetta.morini@unifi.it (B. Morini), philippe.toint@unamur.be (Ph.L. Toint).
[1] Member of the INdAM Research Group GNCS.

1. Introduction

This paper is concerned with trust-region methods for solving the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \qquad f : \mathbb{R}^n \to \mathbb{R}, \qquad (1.1)$$

where we assume that the *values of the objective function f and its derivatives are computed subject to random noise.* Our objective is twofold. Firstly, we introduce a version of the deterministic method proposed in [10] which is able to handle the random context and provide, under reasonable probabilistic assumptions, a sharp evaluation complexity bound (in expectation) for arbitrary optimality order. Secondly, we investigate the effect of intrinsic noise (that is noise whose level cannot be assumed to vanish) on a first-order version of our algorithm and prove "degraded" optimality, should this noise limit the validity of our assumptions. The new results are then detailed and illustrated in the framework of finite-sum minimization using subsampling.

Minimization algorithms using adaptive steplength and allowing for random noise in the objective function or derivatives' evaluations have already generated a significant literature (e.g. [1–3,6,7,13,16]). We focus here on trust-region methods, which are methods in which a trial step is computed by approximately minimizing a model of the objective function in a "trust region" where this model is deemed sufficiently accurate. The trial step is then accepted or rejected depending on whether a sufficient improvement in objective function value predicted by the model is obtained or not, the radius of trust-region being then reduced in the latter case. We refer the reader to [15] for an in-depth coverage of this class of algorithms and to [17] for a more recent survey. Trust-region methods involving stochastic errors in function/derivatives values were considered in particular in [1,12] and [7,13], the latter being the only methods (to the author's knowledge) handling random perturbations in both the objective function and its derivatives. The complexity analysis of the STORM (STochastic Optimization with Random Models) algorithm described in [7,13] is based on supermartingales and makes probabilistic assumptions on the accuracy of these evaluations which become tighter when the trust-region radius becomes small. It also hinges on the definition of a monotonically decreasing "merit function" associated with the stochastic process corresponding to the algorithm. The method proposed in this paper can be viewed as an alternative in the same context, but differs from the STORM approach in several aspects. The first is that the method discussed here uses a model whose degree is chosen adaptively at each iteration, requir-

ing the (noisy) evaluation of higher derivatives only when necessary. The second is that its scope is not limited to searching for first- and second-order approximate minimizers, but is capable of computing them to arbitrary optimality order. The third is that the probabilistic accuracy conditions on the derivatives' estimations no longer depend on the trust-region radius, but rather on the predicted reduction in objective function values, which may be less sensitive to problem conditioning. Finally, its evaluation complexity analysis makes no use of a merit function of the type used in [7].

In [4], the impact of intrinsic random noise on the evaluation complexity of a *deterministic* "noise-aware" trust-region algorithm for unconstrained nonlinear optimization was investigated and contrasted with that of an inexact version where noise is fully controllable. While such results cover some practical occurrences, including noise due to floating point computation, the current paper considers the question in the more general probabilistic framework.

Even if the analysis presented below does not depend in any way on the choice of the optimality order q, the authors are well aware that, while requests for optimality of orders $q \in \{1, 2\}$ lead to practical, implementable algorithms, this may no longer be the case for $q > 2$. For high orders, the methods discussed in the paper therefore constitute an "idealized" setting (in which complicated subproblems can be approximately solved without affecting the evaluation complexity) and thus indicate the limits of achievable results.

The paper is organized as follows. After introducing the new stochastic trust-region algorithm in Section 2, its evaluation complexity is provided in the main Theorem 3.7 on page 22. Section 4 is then devoted to an in-depth discussion of the impact of noise on the first-order instantiation of the algorithm, with a particular emphasis on the case where noise is generated by subsampling in finite-sum minimization context. Conclusions and perspectives are finally proposed in Section 5. Because our contribution borrows ideas from [3], themselves being partly inspired by [12], repeating some material from these sources is necessary to keep our argument understandable. We have however done our best to limit this repetition as much as possible.

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

$$\|S\|_{[p]} \overset{\text{def}}{=} \max_{\|v\|=1} |S[v]^p| = \max_{\|v_1\|=\cdots=\|v_p\|=1} |S[v_1, \ldots, v_p]|$$

the induced Euclidean norm. We also denote by $\nabla_x^j f(x)$ the j-th order derivative tensor of f evaluated at x and note that such a tensor is always symmetric for any $j \geq 2$. $\nabla_x^0 f(x)$ is a synonym for $f(x)$. $\lceil \alpha \rceil$ denotes the smallest integer not smaller than α. Moreover, given a set $\mathcal{B}$, $|\mathcal{B}|$ denotes its cardinality, $\mathbb{1}_{\mathcal{B}}$ refers to its indicator function and $\mathcal{B}^c$ indicates its complement. All stochastic quantities live in a probability space denoted by $(\Omega, \mathcal{A}, \mathbb{P}r)$ with the probability measure $\mathbb{P}r$ and the σ-algebra $\mathcal{A}$ containing subsets of Ω.

We never explicitly define Ω, but specify it through random variables. $\Pr[\text{event}]$ finally denotes the probability of an event and $\mathbb{E}[X]$ the expectation of a random variable X.

2. A trust-region minimization method for problems with randomly perturbed function values and derivatives

We make the following assumptions on the optimization problem (1.1).

AS.1   The function f is q-times continuously differentiable in $\mathbb{R}^n$, for some $q \geq 1$. Moreover, its j-th order derivative tensor is Lipschitz continuous for $j \in \{1, \ldots, q\}$ in the sense that, for each $j \in \{1, \ldots, q\}$, there exists a constant $\vartheta_{f,j} \geq 0$ such that, for all $x, y \in \mathbb{R}^n$,

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

AS.2 f is bounded below in $\mathbb{R}^n$, that is there exists a constant f_{low} such that $f(x) \geq f_{\text{low}}$ for all $x \in \mathbb{R}^n$.

AS.2 ensures that the minimization problem (1.1) is well-posed. AS.1 is a standard assumption in evaluation complexity analysis.[2] It is important because we consider algorithms that are able to exploit all available derivatives of f and, as in many minimization methods, our approach is based on using the Taylor expansions (now of degree j for $j \in \{1, \ldots, q\}$) given by

$$t_{f,j}(x, s) \stackrel{\text{def}}{=} f(x) + \sum_{\ell=1}^{j} \frac{1}{\ell!} \nabla_x^\ell f(x)[s]^\ell. \tag{2.2}$$

AS.1 then has the following crucial consequence on the quantity $f(x + s) - t_{f,j}(x, s)$.

Lemma 2.1. *Suppose that AS.1 holds. Then for all $x, s \in \mathbb{R}^n$,*

$$|f(x + s) - t_{f,j}(x, s)| \leq \frac{\vartheta_{f,j}}{(j + 1)!} \|s\|^{j+1}. \tag{2.3}$$

Proof. See [9, Lemma 2.1] with $\beta = 1$. $\quad\square$

---

[2] It is well-known that requesting (2.1) to hold for all $x, y \in \mathbb{R}^n$ is strong. The weakest form of AS.1 which we could use in what follows is to require (2.1) to hold for all $x = x_k$ (the iterates of the minimization algorithm we are about to describe) and all $y = x_k + \xi s_k$ (where s_k is the associated step and ξ is arbitrary in [0,1]). However, ensuring this condition a priori, although maybe possible for specific applications, is hard in general, especially for a non-monotone algorithm with a random element.

At a given iterate x_k of our algorithm, we will be interested in finding a step $s \in \mathbb{R}^n$ which makes the *Taylor decrements*

$$\Delta t_{f,j}(x_k, s) \stackrel{\text{def}}{=} f(x_k) - t_{f,j}(x_k, s) = t_{f,j}(x_k, 0) - t_{f,j}(x_k, s) \tag{2.4}$$

large (note that $\Delta t_{f,j}(x, s)$ is independent of $f(x)$). When this is possible, we anticipate from the approximating properties of the Taylor expansion that some significant decrease is also possible in f. Conversely, if $\Delta t_{f,j}(x, s)$ cannot be made large in a neighborhood of x, we must be close to an approximate minimizer. More formally, we define, for some $\theta \in (0, 1]$ and some *optimality radius* $\delta \in (0, \theta]$, the measure

$$\phi_{f,j}^{\delta}(x) = \max_{\|d\| \leq \delta} \Delta t_{f,j}(x, d), \tag{2.5}$$

that is the maximal decrease in $t_{f,j}(x, d)$ achievable in a ball of radius δ centered at x. (The practical purpose of introducing θ is to avoid unnecessary computations, as discussed below.) We then define x to be a q-th order (ϵ, δ)-approximate minimizer (for some accuracy requests $\epsilon \in (0, 1]^q$) if and only if

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

(a vector d solving the optimization problem defining $\phi_{f,j}^{\delta}(x)$ in (2.5) is called an *optimality displacement*) [8,10]. This notion is coherent with standard optimality measures for low orders[3] and has the advantage of being well-defined and continuous in x for every order. Note that $\phi_{f,j}^{\delta}(x)$ is always non-negative. A more complete motivation for using $\phi_{f,j}^{\delta}(x)$ is given in Section 12.1.3 of [11].

This paper is concerned with the case where the values of the objective function f and of its derivatives $\nabla_x^j f$ are subject to random noise and can only be computed inexactly (our assumptions on random noise will be detailed below). Our notational convention will be to denote inexact quantities with an overbar, so $\overline{f}(x, \xi)$ and $\overline{\nabla_x^j f}(x, \xi)$ denote inexact values of $f(x)$ and $\nabla_x^j f(x)$, where ξ is a random variable causing inexactness. Thus (2.2) and (2.4) are unavailable, and we have to consider

$$\overline{t}_{f,j}(x_k, s, \xi) \stackrel{\text{def}}{=} \overline{f}(x_k, \xi) + \sum_{\ell=1}^{j} \frac{1}{\ell!} \overline{\nabla_x^\ell f}(x_k, \xi)[s]^\ell$$

and the associated decrement

$$\overline{\Delta t}_{f,j}(x, s_k, \xi) \stackrel{\text{def}}{=} \overline{t}_{f,j}(x_k, 0, \xi) - \overline{t}_{f,j}(x_k, s_k, \xi) = -\sum_{\ell=1}^{j} \frac{1}{\ell!} \overline{\nabla_x^\ell f}(x_k, \xi)[s]^\ell \tag{2.7}$$

---

[3] It is easy to verify that, irrespective of δ, (2.6) holds for $j = 1$ if and only if $\|\nabla_x^1 f(x)\| \leq \epsilon_1$ and that, if $\|\nabla_x^1 f(x)\| = 0$, $\lambda_{\min}[\nabla_x^2 f(x)] \geq -\epsilon_2$ if and only if $\phi_{f,2}^{\delta}(x) \leq \frac{1}{2}\epsilon_2 \delta^2$.

instead. For simplicity, we will often omit to mention the dependence of inexact values on the random variable ξ in what follows, so (2.7) is rewritten as

$$\overline{\Delta t}_{f,j}(x, s_k) \stackrel{\text{def}}{=} \overline{t}_{f,j}(x_k, 0) - \overline{t}_{f,j}(x_k, s_k) = -\sum_{\ell=1}^{j} \frac{1}{\ell!} \overline{\nabla_x^\ell f}(x_k)[s]^\ell. \tag{2.8}$$

This in turn would require that we measure optimality using

$$\overline{\phi}_{f,j}^\delta(x) \stackrel{\text{def}}{=} \max_{\|d\| \le \delta} \overline{\Delta t}_{f,j}(x, d) \tag{2.9}$$

instead of (2.5). However, computing this exact global maximizer may be costly, so we choose to replace the computation of (2.9) by an approximation, that is with the computation of an optimality displacement d with $\|d\| \le \delta$ such that $\varsigma \overline{\phi}_{f,j}^\delta(x) \le \overline{\Delta t}_{f,j}(x, d)$ for some constant $\varsigma \in (0, 1]$.

We can now present the Trust-Region with Noisy Evaluations (TRqNE) algorithm, see on the next page. The worst-case complexity result is given in Theorem 3.7 on page 22 and shows that, under suitable probabilistic assumptions, the TRqNE algorithm finds (in expectation) an (ϵ, δ)-approximate minimizer of arbitrary order $q \ge 1$ in at most $\mathcal{O}(\epsilon^{-(q+1)})$ inexact evaluations of the function and its derivatives. In addition to the standard requirements AS.1–AS.2, the proof of such a complexity result also needs a third assumption (AS.3 in the next section), assuming that inaccurate models or function evaluations do not occur too frequently and that, on successful iterations, the inexactness on function values still allows, on average, the decrease of the objective function. The trust region radius at iteration k is denoted by r_k instead of the standard notation Δ_k.

A feature of the TRqNE algorithm is that it uses an adaptive strategy (in Step 1) to choose the model's degree in view of the desired accuracy and optimality order. Indeed, the model of the objective function used to compute the step is $\overline{t}_{f,j_k}(x_k, s)$, whose degree j_k can vary from an iteration to the other, depending on the "order of (inexact) optimality" achieved at x_k (as determined by Step 1). Also observe that, if the trust-region radius is small (that is $r_k \le \theta$), the optimality displacement d_{k,j_k} is an approximate global minimizer of the model within the trust region, which justifies the choice $s_k = d_{k,j_k}$ in this case. If $r_k > \theta$, the step computation is allowed to be fairly approximate as the only requirement for a step in the trust region is (2.13). This can be interpreted as a generalization of the familiar notions of "Cauchy" and "eigen" points (see [15, Chapter 6]). In addition, note that, while nothing guarantees that $f(x_k) \ge f(x_{k+1})$, the mechanism of the algorithm ensures that $\overline{f}(x_k) \ge \overline{f}(x_{k+1})$.

The TRqNE algorithm generates a random process. Randomness occurs because of the random noise present in the Taylor decreases and objective function values, the former resulting itself from the randomly perturbed derivatives values and, as the algorithm proceeds, from the random realizations of the iterates x_k and steps s_k. In the following analysis, uppercase letters denote random quantities, while lowercase ones denote realizations of these random quantities. Thus, given $\omega \in \Omega$, $x_k = X_k(\omega)$, $g_k = G_k(\omega)$, etc.

Algorithm 2.1. The TRqNE algorithm.

Step 0: Initialization. A criticality order q, a starting point x_0 and accuracy levels $\epsilon \in (0,1)^q$ are given. For a given constant $\eta \in (0,1)$, define

$$\epsilon_{\min} \overset{\text{def}}{=} \min_{j \in \{1,\dots,q\}} \epsilon_j \quad \text{and} \quad \nu \overset{\text{def}}{=} \min\left[\tfrac{1}{2}\eta, \tfrac{1}{4}(1-\eta)\right]. \tag{2.10}$$

The constants $\theta \in [\epsilon_{\min}, 1]$, $\varsigma \in (0,1]$, $\gamma > 1$, $r_{\max} \geq 1$ and an initial trust-region radius $r_0 \in (\epsilon_{\min}, r_{\max}]$ are also given. Set $k = 0$.

Step 1: Derivatives estimation. Set $\delta_k = \min[r_k, \theta]$.

For $j = 1, \dots, q$,

1. Compute derivatives' estimates $\overline{\nabla_x^j f}(x_k)$ and find an optimality displacement $d_{k,j}$ with $\|d_{k,j}\| \leq \delta_k$ such that

$$\varsigma \overline{\phi}_{f,j}^{\delta_k}(x_k) \leq \overline{\Delta t}_{f,j}(x_k, d_{k,j}). \tag{2.11}$$

2. If

$$\overline{\Delta t}_{f,j}(x_k, d_{k,j}) > \left(\frac{\varsigma \epsilon_j}{1+\nu}\right) \frac{\delta_k^j}{j!}, \tag{2.12}$$

go to Step 2 with $j_k = j$.

Set $j_k = q$.

Step 2: Step computation. If $r_k = \delta_k$, set $s_k = d_{k,j_k}$ and $\overline{\Delta t}_{f,j}(x_k, s_k) = \overline{\Delta t}_{f,j}(x_k, d_{k,j_k})$. Otherwise, compute a step s_k such that $\|s_k\| \leq r_k$ and

$$\overline{\Delta t}_{f,j}(x_k, s_k) \geq \overline{\Delta t}_{f,j}(x_k, d_{k,j_k}). \tag{2.13}$$

Step 3: Function decrease estimation. Compute the estimate $\overline{f}(x_k) - \overline{f}(x_k + s_k)$ of $f(x_k) - f(x_k + s_k)$.

Step 4: Test of acceptance. Compute

$$\rho_k = \frac{\overline{f}(x_k) - \overline{f}(x_k + s_k)}{\overline{\Delta t}_{f,j}(x_k, s_k)}. \tag{2.14}$$

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

Step 5: Trust-region radius update. Set

$$r_{k+1} = \begin{cases} \frac{1}{\gamma} r_k, & \text{if } \rho_k < \eta, \\ \min[r_{\max}, \gamma r_k], & \text{if } \rho_k \geq \eta, \end{cases}$$

Increment k by one and go to Step 1.

In particular, we distinguish

- $\Delta t_{f,j}(x,s)$, the value at a (deterministic) x, s of the exact Taylor decrement, that is of the Taylor decrement using the exact values of its derivatives at x;
- $\overline{\Delta t}_{f,j}(x,s) = \overline{\Delta t}_{f,j}(x,s,\xi)$, the value at a (deterministic) x, s of an inexact Taylor decrement, that is of a Taylor decrement using the inexact values of its derivatives (at x) resulting from the realization of random noise;

- $\Delta t_{f,j}(X, S)$, the random variable corresponding to the exact Taylor decrement taken at the random variables X, S;
- $\Delta T_{f,j}(X, S)$, the random variable giving the value of the Taylor decrement using randomly perturbed derivatives, taken at the random variables X, S.

Analogously, $F_k^0 \stackrel{\text{def}}{=} F(X_k)$ and $F_k^s \stackrel{\text{def}}{=} F(X_k + S_k)$ denote the random variables associated with the estimates of $f(X_k)$ and $f(X_k + S_k)$, with their realizations $f_k^0 = \bar{f}(x_k) = \bar{f}(x_k, \xi)$ and $f_k^s = \bar{f}(x_k + s_k) = \bar{f}(x_k + s_k, \xi)$. Similarly, the iterates X_k, as well as the trust-region radiuses R_k, the indices J_k, the optimality radiuses Δ_k, displacements $D_{k,j}$ and the steps S_k are random variables while x_k, r_k, j_k, δ_k, $d_{k,j}$, and s_k denote their realizations. Hence, the TRqNE algorithm generates the random process

$$\{X_k, R_k, M_k, J_k, \Delta_k, \{D_{k,j}\}_{j=1}^{J_k}, S_k, F_k\} \tag{2.15}$$

in which $X_0 = x_0$ (the initial guess) and $R_0 = r_0$ (the initial trust-region radius) are deterministic quantities, and where

$$M_k = \{\overline{\nabla_x^1 f}(X_k), \dots, \overline{\nabla_x^{j_k} f}(X_k)\} \text{ and } F_k = \{F(X_k), F(X_k + S_k)\}.$$

2.1. The probabilistic setting

We now state our probabilistic assumptions on the TRqNE algorithm. For $k \geq 0$, our assumption on the past is formalized by considering $\mathcal{A}_{k-1}$ the σ-algebra induced by the random variables $M_0, M_1, \dots, M_{k-1}$ and $F_0^0, F_0^s, F_1^0, F_1^s, \dots, F_{k-1}^0, F_{k-1}^s$ and let $\mathcal{A}_{k-1/2}$ be that induced by $M_0, M_1, \dots, M_k$ and $F_0^0, F_0^s, \dots, F_{k-1}^0, F_{k-1}^s$, with $\mathcal{A}_{-1} = \sigma(x_0)$.

We first define an event ensuring that the model is accurate enough at iteration k. At the end of Step 2 of this iteration and given $J_k \in \{1, \dots, q\}$, we now define,

$$\mathcal{M}_{k,j}^{(1)} = \left\{ \phi_{f,j}^{\Delta_k}(X_k) \leq \left(\frac{1+\nu}{\varsigma}\right) \Delta T_{f,j}(X_k, D_{k,j}) \right\} \quad (j \in \{1, \dots, J_k\}),$$

$$\mathcal{M}_k^{(2)} = \left\{ (1-\nu)\Delta T_{f,J_k}(X_k, S_k) \leq \Delta t_{f,J_k}(X_k, S_k) \leq (1+\nu)\Delta T_{f,J_k}(X_k, S_k) \right\},$$

$$\mathcal{M}_k = \left(\bigcap_{j \in \{1, \dots, J_k\}} \mathcal{M}_{k,j}^{(1)} \right) \cap \mathcal{M}_k^{(2)}. \tag{2.16}$$

The event $\mathcal{M}_{k,j}^{(1)}$ occurs when the j-th order optimality measure $(j \leq j_k)$ at iteration k is meaningful, while $\mathcal{M}_k^{(2)}$ occurs when this is the case for the model decrease. At first sight, these events may seem only vaguely related to the accuracy of the function's derivatives but a closer examination gives the following sufficient condition for $\mathcal{M}_k$ to happen.

Lemma 2.2. *At iteration k of any realization, the inequalities defining the event $\mathcal{M}_k$ are satisfied if, for $j \in \{1, \ldots, j_k\}$ and $\ell \in \{1, \ldots, j_k\}$*

$$\left| \left(\overline{\nabla_x^\ell f}(x_k) - \nabla_x^\ell f(x_k) \right) [s_k]^\ell \right| \leq \frac{\nu}{2} \overline{\Delta t}_{f,j_k}(x_k, s_k) \tag{2.17}$$

and

$$\left| \left(\overline{\nabla_x^\ell f}(x_k) - \nabla_x^\ell f(x_k) \right) [\hat{d}_{k,j}]^\ell \right| \leq \frac{\nu}{2} \overline{\Delta t}_{f,j}(x_k, \hat{d}_{k,j}), \tag{2.18}$$

where

$$\hat{d}_{k,j} = \underset{\|d\| \leq \delta_k}{\arg \max} \, \Delta t_{f,j}(x_k, d). \tag{2.19}$$

Proof. If (2.18) holds, we have that, for every $j \in \{1, \ldots, j_k\}$ and $v_k \in \{\hat{d}_{k,1}, \ldots, \hat{d}_{k,j}\}$, with $\hat{d}_{k,\ell}, \, \ell = 1, \ldots, j$, given in (2.19),

$$\begin{aligned}
|\overline{\Delta t}_{f,j}(x_k, v_k) - \Delta t_{f,j}(x_k, v_k)| &\leq \sum_{\ell=1}^{j} \frac{1}{\ell!} \left| \left(\overline{\nabla_x^\ell f}(x_k) - \nabla_x^\ell f(x_k) \right) [v_k]^\ell \right| \\
&\leq \frac{1}{2} \nu \overline{\Delta t}_{f,j}(x_k, v_k) \sum_{\ell=1}^{j} \frac{1}{\ell!} \\
&< \nu \, \overline{\Delta t}_{f,j}(x_k, v_k)
\end{aligned} \tag{2.20}$$

where we have used the bound $\sum_{\ell=1}^{j} \frac{1}{\ell!} < e - 1 < 2$. Now note that the definition of $\overline{\phi}_{f,j}^{\delta_k}(x_k)$ in (2.9), (2.20) for $v_k = \hat{d}_{k,j}$ and (2.11) imply that, for any $j \in \{1, \ldots, j_k\}$,

$$\begin{aligned}
\phi_{f,j}^{\delta_k}(x_k) = \Delta t_{f,j}(x_k, \hat{d}_{k,j}) &\leq \overline{\Delta t}_{f,j}(x_k, \hat{d}_{k,j}) + |\Delta t_{f,j}(x_k, \hat{d}_{k,j}) - \overline{\Delta t}_{f,j}(x_k, \hat{d}_{k,j})| \\
&\leq (1+\nu) \overline{\Delta t}_{f,j}(x_k, \hat{d}_{k,j}) \\
&\leq (1+\nu) \max_{\|d\| \leq \delta_k} \overline{\Delta t}_{f,j}(x_k, d) \\
&= (1+\nu) \overline{\phi}_{f,j}^{\delta_k}(x_k) \\
&\leq \left(\frac{1+\nu}{\varsigma} \right) \overline{\Delta t}_{f,j}(x_k, d_{k,j}).
\end{aligned}$$

Hence the inequality in the definition of $\mathcal{M}_{k,j}^{(1)}$ holds for $j \in \{1, \ldots, j_k\}$. The proof of the inequalities defining $\mathcal{M}_k^{(2)}$ is analog to that of (2.20). We have from (2.17) that

$$
\begin{aligned}
|\overline{\Delta t}_{f,j_k}(x_k,s_k) - \Delta t_{f,j_k}(x_k,s_k)| &\leq \sum_{\ell=1}^{j_k} \frac{1}{\ell!} \left| \left(\overline{\nabla_x^\ell f}(x_k) - \nabla_x^\ell f(x_k) \right) [s_k]^\ell \right| \\
&\leq \tfrac{1}{2} \nu \overline{\Delta t}_{f,j_k}(x_k,s_k) \sum_{\ell=1}^{j_k} \frac{1}{\ell!} \\
&< \nu \, \overline{\Delta t}_{f,j_k}(x_k,s_k)
\end{aligned}
\tag{2.21}
$$

where we have again used the bound $\sum_{\ell=1}^{j_k} \dfrac{1}{\ell!} < 2$. $\quad \square$

This result immediately suggests a few comments.

- The conditions (2.17)-(2.18) are merely sufficient, not necessary. In particular, they ignore any possible cancellation of errors between terms of the Taylor expansion of different degree.
- We note that (2.17)-(2.18) require the ℓ-th derivative to be relatively accurate along a finite and limited set of directions, independent of problem dimension.
- Since $\|d_{k,j}\|$ and $\|\hat{d}_{k,j}\|$ are bounded by $\delta_k \leq \theta \leq 1$, we also note that the accuracy required by these conditions decreases when the degree ℓ increases. Moreover, for a fixed degree, the request is weaker for small displacements (a typical situation when a solution is approached) than for large ones.
- Requiring

$$
\|\overline{\nabla_x^\ell f}(x_k) - \nabla_x^\ell f(x_k)\| \leq \frac{\nu}{2\|\hat{d}_{k,j}\|^\ell} \overline{\Delta t}_{f,j}(x_k, \hat{d}_{k,j}),
\tag{2.22}
$$

instead of (2.18) is of course again *sufficient* to ensure the desired conclusions. These conditions are reminiscent of the conditions required in [7] for the STORM algorithm with $p = 2$, namely that, for some constant κ_ℓ and all y in the trust-region $\{y \in \mathbb{R}^n \mid \|y - x_k\| \leq r_k\}$,

$$
\|\overline{\nabla_x^\ell f}(y) - \nabla_x^\ell f(y)\| \leq \kappa_\ell \, r_k^{3-\ell} \quad (\ell \in \{0,1,2\}).
$$

This latter condition is however stronger than (2.17)–(2.18) because it insists on a uniform accuracy guarantee in the full-dimensional trust region.

Having considered the accuracy of the model, we now turn to that on the objective function's values. At the end of Step 3 of the k-th iteration, we define the event

$$
\mathcal{F}_k = \{|\Delta f(X_k, S_k) - \Delta F(X_k, S_k)| \leq 2\nu \Delta T_{f,j_k}(X_k, S_k)\}
\tag{2.23}
$$

where $\Delta f(X_k, S_k) \stackrel{\text{def}}{=} f(X_k) - f(X_k + S_k)$ and $\Delta F(X_k, S_k) \stackrel{\text{def}}{=} F(X_k) - F(X_k + S_k)$. This occurs when the difference in function values used in the course of iteration k is

reasonably accurate relative to the model decrease obtained in that iteration. Note that, because of the triangular inequality,

$$|\Delta f(X_k, S_k) - \Delta F(X_k, S_k)| = |(f(X_k) - f(X_k + S_k)) - (F(X_k) - F(X_k + S_k))|$$
$$\leq |f(X_k) - F(X_k)| + |f(X_k + S_k) - F(X_k + S_k)|$$

so that $\mathcal{F}_k$ must occur if both terms on the right-hand side are bounded above by $\nu \Delta T_{f,j_k}(X_k, S_k)$. Combining accuracy requests on model and function values, we define

$$\mathcal{E}_k \stackrel{\text{def}}{=} \mathcal{F}_k \cap \mathcal{M}_k \tag{2.24}$$

and say that iteration k is *accurate* if $\mathbb{1}_{\mathcal{E}_k} = \mathbb{1}_{\mathcal{F}_k} \mathbb{1}_{\mathcal{M}_k} = 1$ and the iteration k is *inaccurate* if $\mathbb{1}_{\mathcal{E}_k} = 0$. Moreover, we say that the iteration k has accurate model if $\mathbb{1}_{\mathcal{M}_k} = 1$ and that iteration k has accurate function estimates if $\mathbb{1}_{\mathcal{F}_k} = 1$. Finally we let

$$p_{\mathcal{M}_k} \stackrel{\text{def}}{=} \mathbb{P}\text{r}\Big[\mathcal{M}_k \mid \mathcal{A}_{k-1}\Big], \quad p_{\mathcal{F}_k} \stackrel{\text{def}}{=} \mathbb{P}\text{r}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-1}\Big].$$

We will verify in what follows that the TRqNE algorithm does progress towards an approximate minimizer satisfying (2.6) as long as the following holds.

AS.3   There exists $\alpha_*, \gamma_* \in (\frac{1}{2}, 1]$ such that $p_* = \alpha_* \gamma_* > \frac{1}{2}$,

$$p_{\mathcal{M}_k} \geq \alpha_*, \quad p_{\mathcal{F}_k} \geq \gamma_* \text{ and } \mathbb{E}\Big[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\Delta f(X_k, S_k) \mid \mathcal{A}_{k-1}\Big] \geq 0, \tag{2.25}$$

where $\mathcal{S}_k$ is the event $\mathcal{S}_k \stackrel{\text{def}}{=} \{\text{iteration } k \text{ is successful}\}$.

We notice that due to the tower property for conditional expectations

$$\mathbb{P}\text{r}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-1}\Big] = \mathbb{E}\Big[\mathbb{1}_{\mathcal{F}_k} \mid \mathcal{A}_{k-1}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{\mathcal{F}_k} \mid \mathcal{A}_{k-\frac{1}{2}}\Big] \mid \mathcal{A}_{k-1}\Big],$$

and hence that assuming, as in [16] and [7],

$$\mathbb{P}\text{r}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-\frac{1}{2}}\Big] > \gamma_*$$

is stronger than assuming $p_{\mathcal{F}_k} \geq \gamma_*$. Similarly,

$$\mathbb{E}\Big[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\Delta f(X_k, S_k) \mid \mathcal{A}_{k-\frac{1}{2}}\Big] \geq 0 \text{ implies } \mathbb{E}\Big[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\Delta f(X_k, S_k) \mid \mathcal{A}_{k-1}\Big] \geq 0. \tag{2.26}$$

Assuming AS.3 is not unreasonable as it merely requires that an accurate model and accurate functions "happen more often than not", and that the discrepancy between true and inexact function values at successful iterations does not, on average, prevent decrease

of the objective function. If either of these condition fails, it is easy to imagine that the TR*q*NE algorithm could be completely hampered by noise and/or diverge completely. Because the last condition in (2.25) is less intuitive, we now show that it can be realistic in the specific context where reasonable assumptions are made on the (possibly extended) cumulative distribution of the error on the function decreases (conditioned to $\mathcal{A}_{k-\frac{1}{2}}$).

Theorem 2.3. *Let $G_k : \mathbb{R}_+ \to [0,1]$ be a differentiable monotone increasing random function which is measurable for $\mathcal{A}_{k-1}$ and such that*

$$G_k(0) = 0 \text{ and } \lim_{\tau \to \infty} G_k(\tau) = 1, \tag{2.27}$$

$$\lim_{\tau \to \infty} \tau \left(1 - G_k(\tau)\right) = 0, \tag{2.28}$$

$$\int_0^\infty \left(1 - G_k(\tau)\right) d\tau < \infty, \tag{2.29}$$

and

$$\Pr\left[\Delta F(X_k, S_k) - \Delta f(X_k, S_k) > \tau \mid \mathcal{A}_{k-\frac{1}{2}}\right] \le 1 - G_k(\tau) \tag{2.30}$$

for $\tau > 0$. Then,

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})(f(X_k) - f(X_{k+1})) \mid \mathcal{A}_{k-1}\right] \ge 0 \tag{2.31}$$

for each k such that

$$\Delta T_{f,j_k}(X_k, S_k) \ge \frac{1}{\eta} \int_0^\infty \left(1 - G_k(\tau)\right) d\tau.$$

Proof. Consider $\omega \in \Omega$, an arbitrary realization of the stochastic process defined by the TR*q*NE algorithm. Suppose first that $\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k}) \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) = 0$. We then deduce that

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})(f(X_k) - f(X_{k+1})) \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) = 0. \tag{2.32}$$

Assume therefore that

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k}) \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) = \bar{p}_k \tag{2.33}$$

for some $\bar{p}_k > 0$. To further simplify notations, set

$$\Delta T_k \stackrel{\text{def}}{=} \Delta T_{f,j_k}(X_k, S_k) \text{ and } E_k = \Delta F(X_k, S_k) - \Delta f(X_k, S_k). \tag{2.34}$$

If we define $\mathcal{I} \stackrel{\text{def}}{=} \{E_k(\omega) > 0\}$, the definition of successful iterations, (2.14) and the triangular inequality then imply that, if $\mathbb{1}_{\mathcal{S}_k(\omega)} = 1$ then

$$\Delta f(X_k, S_k)(\omega) = \Delta F(X_k, S_k)(\omega) - E_k(\omega) \geq \eta \Delta T_k(\omega) - \mathbb{1}_{\mathcal{I}} E_k(\omega). \tag{2.35}$$

This in turn ensures that

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\Delta f(X_k, S_k) \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) \geq \eta \Delta T_k(\omega)\, \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k}) \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega)$$
$$- \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\mathbb{1}_{\mathcal{I}} E_k \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega). \tag{2.36}$$

Moreover, we have that

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\mathbb{1}_{\mathcal{I}} E_k \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega)$$
$$= \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\mathbb{1}_{\mathcal{I}} E_k \mid \mathcal{A}_{k-\frac{1}{2}}, \mathcal{S}_k \cap \mathcal{F}_k^c\right](\omega) \cdot \Pr[\mathcal{S}_k \cap \mathcal{F}_k^c \mid \mathcal{A}_{k-\frac{1}{2}}](\omega)$$
$$+ \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\mathbb{1}_{\mathcal{I}} E_k \mid \mathcal{A}_{k-\frac{1}{2}}, (\mathcal{S}_k \cap \mathcal{F}_k^c)^c\right](\omega) \cdot \Pr[(\mathcal{S}_k \cap \mathcal{F}_k^c)^c \mid \mathcal{A}_{k-\frac{1}{2}}](\omega)$$
$$= \bar{p}_k\, \mathbb{E}\left[\mathbb{1}_{\mathcal{I}} E_k \mid \mathcal{A}_{k-\frac{1}{2}}, \mathcal{S}_k \cap \mathcal{F}_k^c\right](\omega)$$
$$\leq \bar{p}_k\, \mathbb{E}\left[\mathbb{1}_{\mathcal{I}} E_k \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega),$$

where we used the fact that $[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})](\omega) = 0$ whenever $(\mathcal{S}_k^c \cup \mathcal{F}_k)(\omega)$ happens, (2.33) to derive the second equality, and the bound $\mathbb{1}_{\mathcal{I}} E_k(\omega) \geq 0$ to obtain the final inequality. Now, (2.30) implies that, for $\tau > 0$

$$\Pr\left[\mathbb{1}_{\mathcal{I}} E_k > \tau \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) = \Pr\left[E_k > \tau \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) \leq 1 - g_k(\tau)$$

where $g_k(\tau) \stackrel{\text{def}}{=} G_k(\omega)(\tau)$, and thus

$$\Pr\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\mathbb{1}_{\mathcal{I}} E_k > \tau \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) \leq (1 - g_k(\tau))\, \bar{p}_k = \bar{p}_k \int_\tau^\infty g_k'(t)\, dt.$$

Then, employing (2.27)–(2.30), and integrating by parts

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\mathbb{1}_{\mathcal{I}} E_k \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) \leq \bar{p}_k \int_0^\infty t\, g_k'(t)\, dt = \bar{p}_k \int_0^\infty (1 - g_k(t))\, dt < \infty.$$

Finally, using (2.36),

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})\Delta f(X_k, S_k) \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) \geq \bar{p}_k \left[\eta \Delta T_k(\omega) - \int_0^\infty (1 - g_k(t))\, dt\right]$$

and thus

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})(f(X_k) - f(X_{k+1})) \mid \mathcal{A}_{k-\frac{1}{2}}\right](\omega) \geq 0 \tag{2.37}$$

holds when

$$\Delta T_k(\omega) \geq \frac{1}{\eta} \int\limits_0^\infty (1 - g_k(t))\, dt.$$

Combining (2.32) and (2.37) and taking into account that ω is arbitrary give that

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}(1 - \mathbb{1}_{\mathcal{F}_k})(f(X_k) - f(X_{k+1})) \mid \mathcal{A}_{k-\frac{1}{2}}\right] \geq 0.$$

Thus, (2.26) yields (2.31).   □

Note that the assumptions of the theorem are for instance satisfied for the exponential case where $G_k(\tau) = e^{-T\tau}$ for $T > 0$ and measurable for $\mathcal{A}_{k-1}$. We will return to this result in Section 4 and discuss there the condition that $\Delta T_{f,j_k}(X_k, S_k)$ should be sufficiently large.

3. Worst-case evaluation complexity

We now turn to the evaluation complexity analysis for the TRqNE algorithm, whose aim is to derive a bound on the expected number of iterations for which optimality fails. This number is given by

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

We first state a crucial lower bound on the model decrease, in the spirit of [10, Lemma 3.4].

Lemma 3.1. *Consider any realization of the algorithm and assume that $\mathcal{M}_k$ occurs. Assume that (2.6) fails at iteration k. Then, there exists a $j_k \in \{1, \ldots, q\}$ such that $\overline{\Delta t}_{f,j_k}(x_k, d_{k,j_k}) > \varsigma\epsilon_{j_k}\delta_k^{j_k}/(j_k!(1+\nu))$ at Step 1 of the iteration. Moreover,*

$$\overline{\Delta t}_{f,j_k}(x_k, s_k) \geq \widehat{\phi}_{f,k} \frac{\delta_k^{j_k}}{j_k!} \tag{3.2}$$

where

$$\widehat{\phi}_{f,k} \stackrel{\text{def}}{=} \frac{j_k!\, \overline{\Delta t}_{f,j_k}(x_k, d_{k,j_k})}{\delta_k^{j_k}} > \frac{\varsigma\epsilon_{\min}}{1+\nu}. \tag{3.3}$$

Proof. We proceed by contradiction and assume that

$$\overline{\Delta t}_{f,j}(x_k, d_{k,j}) \leq \frac{\varsigma \epsilon_j}{1 + \nu} \frac{\delta_k^j}{j!}, \tag{3.4}$$

for all $j \in \{1, \ldots, q\}$. Since $\mathcal{M}_k$ occurs, we have that, for all $j \in \{1, \ldots, q\}$,

$$\phi_{f,j}^{\delta_k}(x_k) \leq \left(\frac{1 + \nu}{\varsigma}\right) \overline{\Delta t}_{f,j}(x_k, d_{k,j}) \leq \epsilon_j \frac{\delta_k^j}{j!}, \quad j \in \{1, ..., q\},$$

which contradicts the assumption that (2.6) does not hold for x_k and δ_k. The bound (3.2) directly results from

$$\overline{\Delta t}_{f,j_k}(x_k, s_k) \geq \overline{\Delta t}_{f,j_k}(x_k, d_{k,j_k}) = \widehat{\phi}_{f,k} \frac{\delta_k^{j_k}}{j_k!},$$

where we have used (2.13) to derive the first inequality and the definition (3.3) to obtain the equality. The rightmost inequality in (3.3) trivially follows from the negation of (3.4) and (2.10). $\quad\square$

We now show that, for any accurate iteration k there exists $\bar{r} > 0$ such that the iteration is successful whenever $r_k \leq \bar{r}$. For simplicity of notation, given $\vartheta_{f,j}$, $j \in \{1, \ldots, q\}$, as in (2.1), we define

$$\vartheta_f \stackrel{\text{def}}{=} \max[1, \max_{j \in \{1, \ldots, q\}} \vartheta_{f,j}]. \tag{3.5}$$

Lemma 3.2. *Suppose that AS.1 holds. Consider any realization of the algorithm and suppose that (2.6) does not hold for x_k and δ_k and that $\mathcal{E}_k$ occurs. If*

$$r_k \leq \bar{r} \stackrel{\text{def}}{=} \min\left\{\theta, \frac{\varsigma(1 - \eta)}{4(1 + \nu)\vartheta_f} \epsilon_{\min}\right\} = \frac{\varsigma(1 - \eta)}{4(1 + \nu)\vartheta_f} \epsilon_{\min} \stackrel{\text{def}}{=} \kappa_r \epsilon_{\min}, \tag{3.6}$$

$\kappa_r \in (0, 1)$, holds, then iteration k is successful.

Proof. First, note that the minimum in (3.6) is attained at $\kappa_r \epsilon_{\min}$ since $\theta \geq \epsilon_{\min}$ and $\kappa_r \in (0, 1)$. Suppose now that (3.6) holds, which implies that $\delta_k = \min[\theta, r_k] = r_k$. Let j_k be as in Lemma 3.1, and denote $\Delta f(x_k, s_k) = f(x_k) - f(x_k + s_k)$, $\overline{\Delta f}(x_k, s_k) = \overline{f}(x_k) - \overline{f}(x_k + s_k)$.

Using (2.14), the triangle inequality and $\mathbb{1}_{\mathcal{E}_k} = \mathbb{1}_{\mathcal{M}_k}\mathbb{1}_{\mathcal{F}_k} = 1$, we obtain

$$|\rho_k - 1| = \left| \frac{\overline{\Delta f}(x_k, s_k) - \overline{\Delta t}_{f,j_k}(x_k, s_k)}{\overline{\Delta t}_{f,j_k}(x_k, s_k)} \right|$$

$$\leq \frac{|\Delta f(x_k, s_k) - \Delta t_{f,j_k}(x_k, s_k)|}{\overline{\Delta t}_{f,j_k}(x_k, s_k)} + \frac{|\Delta t_{f,j_k}(x_k, s_k) - \overline{\Delta t}_{f,j_k}(x_k, s_k)|}{\overline{\Delta t}_{f,j_k}(x_k, s_k)}$$

$$+ \frac{|\overline{\Delta f}(x_k, s_k) - \Delta f(x_k, s_k)|}{\overline{\Delta t}_{f,j_k}(x_k, s_k)}$$

$$\leq \frac{|f(x_k + s_k) - t_{f,j_k}(x_k, s_k)|}{\overline{\Delta t}_{f,j_k}(x_k, s_k)} + \frac{3\nu \overline{\Delta t}_{f,j_k}(x_k, s_k)}{\overline{\Delta t}_{f,j_k}(x_k, s_k)}.$$

Invoking (2.3), the bound $\|s_k\| \leq r_k = \delta_k$, (3.5), (3.2) and $\nu \leq \frac{1}{4}(1 - \eta)$ we get

$$|\rho_k - 1| < \frac{\vartheta_f r_k}{\hat{\phi}_{f,k}} + \frac{3}{4}(1 - \eta).$$

Using (3.3) and (3.6) we deduce that

$$|\rho_k - 1| \leq \frac{(1 + \nu)\vartheta_f r_k}{\varsigma \epsilon_{\min}} + \frac{3}{4}(1 - \eta) \leq 1 - \eta. \tag{3.7}$$

Thus, $\rho_k \geq \eta$ and the iteration k is successful. $\square$

We now provide a crucial lower bound, depending on a suitable power of $\epsilon_{\min}$, for realizations of $\Delta T_{f,j_k}(X_k, S_k)$ at any accurate iteration k such that $r_k > \bar{r}$.

Lemma 3.3. *Suppose that AS.1 holds. Consider any realization of the algorithm and suppose that (2.6) does not hold for x_k and δ_k, $\mathcal{E}_k$ occurs, and that $r_k \geq \bar{r}$ with $\bar{r}_k$ defined in (3.6). Then*

$$\overline{\Delta t}_{f,j_k}(x_k, s_k) = \bar{t}_{f,j_k}(x_k, 0) - \bar{t}_{f,j_k}(x_k, s_k) > \frac{\varsigma}{q!}(\kappa_\delta \epsilon_{\min})^{q+1}, \tag{3.8}$$

where $\kappa_\delta \in (0, 1)$ is defined by

$$\kappa_\delta \stackrel{\text{def}}{=} \frac{\kappa_r}{1 + \nu} \tag{3.9}$$

with κ_r defined in (3.6).

Proof. Let j_k be as in Lemma 3.1. By (3.2), (3.3) we obtain

$$\overline{\Delta t}_{f,j_k}(x_k, s_k) > \frac{\varsigma \epsilon_{\min}}{1 + \nu} \frac{\delta_k^q}{q!}.$$

If $r_k > \theta$ then $\delta_k = \theta$ and the bound $\theta \geq \epsilon_{\min}$ implies

$$\overline{\Delta t}_{f,j}(x_k, s_k) > \frac{\varsigma \epsilon_{\min}^{q+1}}{q!(1+\nu)}.$$

Thus (3.8) holds by definition of κ_δ and the fact that $\kappa_r \in (0,1)$. If $\bar{r} < r_k \leq \theta$, then $\delta_k = r_k$. The proof is completed by noting that the form of $\bar{r}$ in (3.6) gives that $r_k > \kappa_r \epsilon_{\min}$. $\square$

3.1. Bounding the expected number of steps with $R_k \leq \bar{r}$

We now return to the general stochastic process generated by the TRqNE algorithm and bound the expected number of steps in N_ϵ from above. For this purpose, let us define, for all $0 \leq k \leq N_\epsilon - 1$, the events

$$\Lambda_k \overset{\text{def}}{=} \{R_k > \bar{r}\}, \qquad \Lambda_k^c \overset{\text{def}}{=} \{R_k \leq \bar{r}\},$$

where $\bar{r}$ is given by (3.6), and let

$$N_\Lambda \overset{\text{def}}{=} \sum_{k=0}^{N_\epsilon - 1} \mathbb{1}_{\Lambda_k}, \qquad N_{\Lambda^c} \overset{\text{def}}{=} \sum_{k=0}^{N_\epsilon - 1} \mathbb{1}_{\Lambda_k^c}, \tag{3.10}$$

be the number of steps, in the stochastic process induced by the TRqNE algorithm and before N_ϵ, such that $R_k > \bar{r}$ or $R_k \leq \bar{r}$, respectively. In what follows we suppose that AS.1–AS.2 hold.

An upper bound on $\mathbb{E}[N_{\Lambda^c}]$ can be derived as follows.

(i) We apply [12, Lemma 2.2] to deduce that, for any $\ell \in \{0, \ldots, N_\epsilon - 1\}$ and for all realizations of Algorithm 2.1, one has that

$$\sum_{k=0}^{\ell} \mathbb{1}_{\Lambda_k^c} \mathbb{1}_{\mathcal{S}_k} \leq \frac{\ell + 1}{2}. \tag{3.11}$$

(ii) Both $\sigma(\mathbb{1}_{\Lambda_k})$ and $\sigma(\mathbb{1}_{\Lambda_k^c})$ belong to $\mathcal{A}_{k-1}$, because the random variable Λ_k is fully determined by the first $k-1$ iterations of the TRqNE algorithm. Setting $\ell = N_\epsilon - 1$ we can rely on [12, Lemma 2.1] (with $W_k = \mathbb{1}_{\Lambda_k^c}$), whose proof is detailed in Appendix A, to deduce that

$$\mathbb{E}\left[\sum_{k=0}^{N_\epsilon - 1} \mathbb{1}_{\Lambda_k^c} \mathbb{1}_{\mathcal{E}_k}\right] \geq p_* \mathbb{E}\left[\sum_{k=0}^{N_\epsilon - 1} \mathbb{1}_{\Lambda_k^c}\right]. \tag{3.12}$$

(iii) As a consequence, given that Lemma 3.2 ensures that each iteration k where $\mathcal{E}_k$ occurs and $r_k \leq \bar{r}$ is successful, we have that

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

in which the last inequality follows from (3.11), with $\ell = N_\epsilon - 1$. Taking expectation in the above inequality, using (3.12) and recalling the rightmost definition in (3.10), we obtain, as in [12, Lemma 2.3], that

$$\mathbb{E}[N_{\Lambda^c}] \le \frac{1}{2p_*}\mathbb{E}[N_\epsilon]. \tag{3.13}$$

3.2. Bounding the expected number of steps with $R_k > \overline{r}$

For analyzing $\mathbb{E}[N_\Lambda]$, where N_Λ is defined in (3.10), we now introduce the following variables.

Definition 1. Consider the random process (2.15) generated by the TRqNE algorithm and define:

- $\overline{\Lambda}_k$ $= \{\text{iteration } k \text{ is such that } R_k \ge \overline{r}\};$

- N_I $= \displaystyle\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{E}_k^c}$: the number of inaccurate iterations with $R_k \ge \overline{r}$;

- N_A $= \displaystyle\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{E}_k}$: the number of accurate iterations with $R_k \ge \overline{r}$;

- N_{AS} $= \displaystyle\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{E}_k}\mathbb{1}_{\mathcal{S}_k}$: the number of accurate successful iterations with $R_k \ge \overline{r}$;

- N_{AU} $= \displaystyle\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\Lambda_k}\mathbb{1}_{\mathcal{E}_k}\mathbb{1}_{\mathcal{S}_k^c}$: the number of accurate unsuccessful iterations with $R_k > \overline{r}$;

- N_{IS} $= \displaystyle\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{E}_k^c}\mathbb{1}_{\mathcal{S}_k}$: the number of inaccurate successful iterations with $R_k \ge \overline{r}$;

- N_S $= \displaystyle\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{S}_k}$: the number of successful iterations with $R_k \ge \overline{r}$;

- N_U $= \displaystyle\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\Lambda_k}\mathbb{1}_{\mathcal{S}_k^c}$: the number of unsuccessful iterations with $R_k > \overline{r}$.

$$\tag{3.14}$$

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

$$\mathbb{E}[N_\Lambda] \le \mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\right] = \mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{E}_k^c} + \sum_{k=0}^{N_\epsilon-1} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{E}_k}\right] = \mathbb{E}[N_I] + \mathbb{E}[N_A]. \quad (3.15)$$

Using again [12, 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, whose proof is detailed in Appendix A. This is the first result of a series of lemmas needed within this section to prove the final complexity result stated in Theorem 3.7.

Lemma 3.4 *([12, Lemma 2.6]). Suppose that AS.1-AS.3 hold and let N_I, N_A be defined as in Definition 1 in the context of the stochastic process (2.15) generated by the TRqNE algorithm. Then,*

$$\mathbb{E}[N_I] \le \frac{1 - p_*}{p_*} \mathbb{E}[N_A]. \quad (3.16)$$

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

$$\mathbb{E}[N_A] = \mathbb{E}[N_{AS}] + \mathbb{E}[N_{AU}] \le \mathbb{E}[N_{AS}] + \mathbb{E}[N_U]. \quad (3.17)$$

Hence, bounding $\mathbb{E}[N_A]$ 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 TRqNE algorithm ensures that R_k is increased by a factor γ on successful steps and decreased by the same factor on unsuccessful ones. Consequently, based on [12, Lemma 2.5], we obtain the following upper bound on the number of unsuccessful iterations with $R_k > \overline{r}$.

Lemma 3.5. *For any $\ell \in \{0, ..., N_\epsilon - 1\}$ and for all realizations of Algorithm 2.1, we have that*

$$\sum_{k=0}^{\ell} \mathbb{1}_{\Lambda_k}\mathbb{1}_{\mathcal{S}_k^c} \le \sum_{k=0}^{\ell} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{S}_k} + \left\lceil\left|\log_{\gamma^{-1}}\left(\frac{\overline{r}}{r_0}\right)\right|\right\rceil = \sum_{k=0}^{\ell} \mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\mathcal{S}_k} + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil.$$

From the inequality stated in the previous lemma with $\ell = N_\epsilon - 1$, recalling Definition 1 and taking expectations, we therefore obtain that

$$\mathbb{E}[N_U] \le \mathbb{E}[N_S] + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil = \mathbb{E}[N_{AS}] + \mathbb{E}[N_{IS}] + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil. \quad (3.18)$$

An upper bound on $\mathbb{E}[N_{AS}]$, in terms of a suitable power of $\epsilon_{\min}$, is given by the following lemma.

Lemma 3.6. *Suppose that AS.1, AS.2 and AS.3 hold. Then we have that*

$$\mathbb{E}[N_{AS}] \leq \frac{2q!(f_0 - f_{\text{low}})}{\varsigma\eta\,(\kappa_\delta\epsilon_{\min})^{q+1}} + 1,\tag{3.19}$$

where κ_δ is defined in (3.9).

Proof. Consider any realization of the TRqNE algorithm.

i) If iteration k is successful and the functions are accurate (i.e., $\mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{F}_k} = 1$) then
(2.14), (2.10) and (2.23) imply that

$$\begin{aligned}
f(x_k) - f(x_{k+1}) &\geq [\overline{f}(x_k) - \overline{f}(x_{k+1})] - 2\nu\overline{\Delta t}_{f,j_k}(x_k, s_k)\\
&\geq \eta\overline{\Delta t}_{f,j_k}(x_k, s_k) - 2\nu\overline{\Delta t}_{f,j_k}(x_k, s_k)\\
&= (\eta - 2\nu)\overline{\Delta t}_{f,j_k}(x_k, s_k)\\
&\geq \frac{1}{2}\eta\overline{\Delta t}_{f,j_k}(x_k, s_k) \geq 0.
\end{aligned}\tag{3.20}$$

Thus

$$\mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{F}_k} = \mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{F}_k}\mathbb{1}_{\{\Delta F_{k,+}\geq 0\}},\tag{3.21}$$

where $\Delta F_{k,+} = f(X_k) - f(X_{k+1})$. Moreover, if $\mathcal{M}_k$ also occurs with $r_k \geq \bar{r}$ (i.e., if
$\mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{E}_k}\mathbb{1}_{\overline{\Lambda}_k} = 1$) and (2.6) fails for x_k and δ_k, we may then use (3.8) to deduce from
(3.20) that

$$f(x_k) - f(x_{k+1}) \geq \frac{\varsigma\eta}{2q!}\,(\kappa_\delta\epsilon_{\min})^{q+1} > 0,\tag{3.22}$$

which implies that, as long as (2.6) fails,

$$\mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{E}_k}\mathbb{1}_{\overline{\Lambda}_k} = \mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{E}_k}\mathbb{1}_{\overline{\Lambda}_k}\mathbb{1}_{\{\Delta F_{k,+}>0\}}.\tag{3.23}$$

ii) If iteration k is unsuccessful, the mechanism of the TRqNE algorithm guarantees that
$x_k = x_{k+1}$ and, hence, that $f(x_{k+1}) = f(x_k)$, giving that $(1 - \mathbb{1}_{\mathcal{S}_k})\Delta F_{k,+} = 0$.

Setting $f_0 \stackrel{\text{def}}{=} f(X_0)$ and using this last relation and AS.2, we have that, for any $\ell \in
\{0, ..., N_\epsilon - 1\}$,

$$f_0 - f_{\text{low}} \geq f_0 - f(X_{\ell+1}) = \sum_{k=0}^{\ell} \Delta F_{k,+} = \sum_{k=0}^{\ell} \mathbb{1}_{\mathcal{S}_k}\Delta F_{k,+}.\tag{3.24}$$

Remembering that X_0 and thus f_0 are deterministic and taking the total expectation on both sides of (3.24) then gives that

$$f_0 - f_{\text{low}} = \mathbb{E}[f_0 - f_{\text{low}}] \geq \sum_{k=0}^{\ell} \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \Delta F_{k,+}\right] = \sum_{k=0}^{\ell} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \Delta F_{k,+} \mid \mathcal{A}_{k-1}\right]\right]. \quad (3.25)$$

Now, for $k \in \{0, \ldots, \ell\}$,

$$\mathbb{1}_{\mathcal{S}_k} \Delta F_{k,+} = \mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} \Delta F_{k,+} + \mathbb{1}_{\mathcal{S}_k} (1 - \mathbb{1}_{\mathcal{F}_k}) \Delta F_{k,+}$$

and so, using the second part of (2.25),

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \Delta F_{k,+} \mid \mathcal{A}_{k-1}\right] = \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} \Delta F_{k,+} \mid \mathcal{A}_{k-1}\right] + \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} (1 - \mathbb{1}_{\mathcal{F}_k}) \Delta F_{k,+} \mid \mathcal{A}_{k-1}\right]$$
$$\geq \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} \Delta F_{k,+} \mid \mathcal{A}_{k-1}\right]. \quad (3.26)$$

Thus, again using the law of total expectations, (3.25) yields that

$$f_0 - f_{\text{low}} \geq \sum_{k=0}^{\ell} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} \Delta F_{k,+} \mid \mathcal{A}_{k-1}\right]\right] = \sum_{k=0}^{\ell} \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} \Delta F_{k,+}\right]. \quad (3.27)$$

Moreover, successively using (3.21), (2.24), (3.23) and (3.22),

$$\mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} \Delta F_{k,+} = \mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} \mathbb{1}_{\{\Delta F_{k,+} > 0\}} \Delta F_{k,+}$$
$$= \mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} \mathbb{1}_{\mathcal{M}_k} \mathbb{1}_{\{\Delta F_{k,+} > 0\}} \Delta F_{k,+} + \mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{F}_k} (1 - \mathbb{1}_{\mathcal{M}_k}) \mathbb{1}_{\{\Delta F_{k,+} > 0\}} \Delta F_{k,+}$$
$$\geq \mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{E}_k} \mathbb{1}_{\{\Delta F_{k,+} > 0\}} \Delta F_{k,+}$$
$$\geq \mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{E}_k} \mathbb{1}_{\overline{\Lambda}_k} \mathbb{1}_{\{\Delta F_{k,+} > 0\}} \Delta F_{k,+}$$
$$\geq \frac{\varsigma\eta}{2q!} \left(\kappa_\delta \epsilon_{\min}\right)^{q+1} \left(\mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{E}_k} \mathbb{1}_{\overline{\Lambda}_k}\right). \quad (3.28)$$

Substituting this bound in (3.27) then gives that, as long as (2.6) fails for iterations $\{1, \ldots, \ell\}$,

$$f_0 - f_{\text{low}} \geq \frac{\varsigma\eta}{2q!} \left(\kappa_\delta \epsilon_{\min}\right)^{q+1} \sum_{k=0}^{\ell} \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{E}_k} \mathbb{1}_{\overline{\Lambda}_k}\right]. \quad (3.29)$$

We now notice that, by Definition 1,

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

and therefore

$$\mathbb{E}[N_{AS} - 1] \leq \sum_{k=0}^{N_\epsilon - 2} \mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k} \mathbb{1}_{\mathcal{E}_k} \mathbb{1}_{\overline{\Lambda}_k}\right]. \quad (3.30)$$

Hence, letting $\ell = N_\epsilon - 2$, substituting (3.30) in (3.29), we deduce that

$$\mathbb{E}[N_{AS} - 1]\frac{\varsigma\eta}{2q!}\left(\kappa_\delta\epsilon_{\min}\right)^{q+1} \leq f_0 - f_{\text{low}}$$

and (3.19) follows. □

While inequalities (3.18) and (3.19) provide upper bounds on $\mathbb{E}[N_{AS}]$ and $\mathbb{E}[N_U]$, as desired, the first still depends on $\mathbb{E}[N_{IS}]$, which has to be bounded from above as well. This can be done by following [12] once more: Definition 1, (3.16) and (3.17) 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

$$\mathbb{E}[N_{IS}] \leq \frac{1 - p_*}{2p_* - 1}\left(2\mathbb{E}[N_{AS}] + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil\right) \tag{3.32}$$

follows from (3.18) (remember that $\frac{1}{2} < p_* \leq 1$). Thus, the right-hand side in (3.16) is in turn bounded above because of (3.17), (3.18), (3.32) and (3.19), giving

$$\begin{aligned}
\mathbb{E}[N_A] &\leq \mathbb{E}[N_{AS}] + \mathbb{E}[N_U] \leq 2\mathbb{E}[N_{AS}] + \mathbb{E}[N_{IS}] + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil \\
&\leq \left(\frac{1 - p_*}{2p_* - 1} + 1\right)\left(2\mathbb{E}[N_{AS}] + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil\right) \\
&\leq \frac{p_*}{2p_* - 1}\left[\frac{4q!(f_0 - f_{\text{low}})}{\varsigma\eta\left(\kappa_\delta\epsilon_{\min}\right)^{q+1}} + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil + 2\right].
\end{aligned} \tag{3.33}$$

This inequality, together with (3.15) and (3.16), finally gives the desired bound

$$\mathbb{E}[N_\Lambda] \leq \frac{1}{p_*}\mathbb{E}[N_A] \leq \frac{1}{2p_* - 1}\left[\frac{4q!(f_0 - f_{\text{low}})}{\varsigma\eta\left(\kappa_\delta\epsilon_{\min}\right)^{q+1}} + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil + 2\right]. \tag{3.34}$$

We can now express our final complexity result in full.

Theorem 3.7. *Suppose that AS.1–AS.3 hold, then*

$$\mathbb{E}[N_\epsilon] \leq \frac{2p_*}{(2p_* - 1)^2}\left[\frac{4q!(f_0 - f_{\text{low}})}{\varsigma\eta\left(\kappa_\delta\epsilon_{\min}\right)^{q+1}} + \left\lceil\log_\gamma\left(\frac{r_0}{\overline{r}}\right)\right\rceil + 2\right], \tag{3.35}$$

with N_ϵ, $\overline{r}$ and κ_δ defined as in (3.1), (3.6) and (3.9), respectively.

Proof. Recalling the definitions (3.10) and the bound (3.13), 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

$$\frac{2p_* - 1}{2p_*}\mathbb{E}[N_\epsilon] \leq \frac{1}{2p_* - 1}\left[\frac{4q!(f_0 - f_{\text{low}})}{\varsigma\eta\left(\kappa_\delta\epsilon_{\text{min}}\right)^{q+1}} + \left\lceil\log_\gamma\left(\frac{r_0}{\bar{r}}\right)\right\rceil + 2\right].$$

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

We note that the $\mathcal{O}\left(\epsilon_{\text{min}}^{-(q+1)}\right)$ evaluation bound given by (3.35) is known to be sharp in order of ϵ_{min} for trust-region methods using exact evaluations of functions and derivatives (see [11, Theorem 12.2.6]), which implies that Theorem 3.7 is also sharp in order.

We conclude this section by noting that alternatives to the second part of (2.25) do exist. For instance, we could assume that

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}\Delta F_{k,+} \mid \mathcal{A}_{k-1}\right] \geq \mu\mathbb{E}\left[\mathbb{1}_{\mathcal{S}_k}\mathbb{1}_{\mathcal{F}_k}\Delta F_{k,+} \mid \mathcal{A}_{k-1}\right]$$

for some $\mu > 0$. This condition can be used to replace the second part of (2.25) to ensure (3.26) in the proof of Lemma 3.6 and all subsequent arguments.

4. The impact of noise for first-order minimization

While the above theory covers inexact evaluations of the objective function and its derivatives, it does rely on AS.3. Thus, as long as the inexactness/noise on these values remains small enough for this assumption to hold, the TRqNE algorithm iterates ultimately produce an approximate local minimizer. There are however practical applications, such as minimization of finite sum using sampling strategies (discussed in more detail below), where AS.3 may be unrealistic because of noise intrinsic to the application. We already saw that, under the assumptions of Theorem 2.3, a large enough value of $\Delta T_{f,j_k}(X_k, S_k)$ is sufficient for ensuring the third condition in AS.3, but we also know from (2.23),(2.34), (2.35) and the definition of ν that, at successful iterations,

$$\Delta f(X_k, S_k) \geq \eta\Delta T_{f,j_k}(X_k, S_k) - E_k \geq (\eta - 2\nu)\Delta T_{f,j_k} \geq \tfrac{1}{2}\eta\Delta T_{f,j_k}$$

whenever $\mathcal{F}_k$ holds. Thus a large $\Delta T_{f,j_k}(X_k, S_k)$ is only possible if either $\Delta f(X_k, S_k)$ is large or $\mathcal{F}_k$ fails. But a large $\Delta f(X_k, S_k)$ is impossible close to a (global) minimizer, and thus either $\mathcal{F}_k$ (and AS.3) fails, or the guarantee that the third condition of AS.3 holds vanishes when approaching a minimum.

Clearly, the above theory does not say anything about what happens for the algorithm once AS.3 fails due to intrinsic noise. Of course, this does not mean it will not proceed

meaningfully, but we can't guarantee it. In order to improve our understanding of what can happen, we need to consider particular realizations of the iterative process where AS.3 fails. This is the objective of this section where we focus on the instantiation TR1NE of TRqNE for first-order optimality. Fortunately, limiting one's ambition to first order results in substantial simplifications in the TRqNE algorithm. We first note that the mechanism of Step 1 of TRqNE (whose purpose is to determine j_k) is no longer necessary since j_k must be equal to one if only (approximate) gradients are available, so we can implicitly set

$$D_{k,1} = \frac{\overline{\nabla_x^1 f}(X_k)}{\|\overline{\nabla_x^1 f}(X_k)\|}\Delta_k \ \ \text{and} \ \ \Delta T_{f,1}(X_k, D_{k,1}) = -\overline{\nabla_x^1 f}(X_k)^T D_{k,1} = \|\overline{\nabla_x^1 f}(X_k)\|\Delta_k$$

and immediately branch to the step computation. This in turn simplifies to

$$S_k = \frac{\overline{\nabla_x^1 f}(X_k)}{\|\overline{\nabla_x^1 f}(X_k)\|}R_k \ \ \text{and} \ \ \Delta T_{f,1}(X_k, S_k) = \|\overline{\nabla_x^1 f}(X_k)\| \, R_k$$

irrespective of the value of θ, and (2.13) automatically holds. We thus observe that the simplified algorithm no longer needs $\delta_{k,j}$ (since neither $\overline{\phi}_{f,j}^{\delta_k}(x_k)$ or $\overline{\Delta t}_{f,j}(x_k, d_{k,j})$ needs to be effectively calculated), that the computed step s_k is the global minimizer within the trust-region and that the constant θ (used in Step 1 and the start of Step 2 of the TRqNE algorithm) is no longer necessary. The resulting streamlined TR1NE algorithm is stated as Algorithm 4.1.

Algorithm 4.1. The TR1NE algorithm.

Step 0: Initialization. A starting point x_0, a constant $\gamma > 1$, a maximum radius $r_{\max} > 0$ and an accuracy level $\epsilon \in (0,1)$ are given. The initial trust-region radius $r_0 \in (\epsilon, r_{\max}]$ is also given. For a given constant $\eta \in (0,1)$, define $\nu \stackrel{\text{def}}{=} \min\left[\frac{1}{2}\eta, \frac{1}{4}(1-\eta)\right]$. Set $k = 0$.

Step 1: Derivatives estimation. Compute the derivative estimate $\overline{\nabla_x f}(x_k)$.

Step 2: Step computation. Set $s_k = -\dfrac{\overline{\nabla_x^1 f}(x_k)}{\|\overline{\nabla_x^1 f}(x_k)\|}r_k$.

Step 3: Function decrease estimation. Compute the estimate $\overline{f}(x_k) - \overline{f}(x_k + s_k)$ of $f(x_k) - f(x_k + s_k)$.

Step 4: Test of acceptance. Compute $\rho_k = \dfrac{\overline{f}(x_k) - \overline{f}(x_k + s_k)}{\|\overline{\nabla_x^1 f}(x_k)\| r_k}$.

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

Step 5: Trust-region radius update. Set

$$r_{k+1} = \begin{cases} \frac{1}{\gamma}r_k, & \text{if } \rho_k < \eta \\ \min[r_{\max}, \gamma r_k], & \text{if } \rho_k \geq \eta. \end{cases}$$

Increment k by one and go to Step 1.

The definition of the event $\mathcal{M}_k$ in (2.16) ensures that $\mathcal{M}_k^{(2)}$ implies $\mathcal{M}_k^{(1)}$ when first-order models are considered, and thus, using also (2.23), that $\mathcal{M}_k$ and $\mathcal{F}_k$ then reduce to

$$\mathcal{M}_k = \{|\|\nabla_x^1 f(X_k)\| - \|\overline{\nabla_x^1 f}(X_k)\|| \leq \nu\|\overline{\nabla_x^1 f}(X_k)\|\}$$

and

$$\mathcal{F}_k = \{|\Delta F(X_k, S_k) - \Delta f(X_k, S_k)| \leq 2\nu\|\overline{\nabla_x^1 f}(X_k)\|R_k\},$$

respectively. Observe now that, because of the triangle inequality, $\mathcal{M}_k$ is true whenever the event

$$\widetilde{\mathcal{M}}_k \overset{\text{def}}{=} \{\|\nabla_x^1 f(X_k) - \overline{\nabla_x^1 f}(X_k)\| \leq \nu\|\overline{\nabla_x^1 f}(X_k)\|\} \tag{4.1}$$

holds, and, since $\nu\|\overline{\nabla_x^1 f}(X_k)\|\min\{1, R_k\} \leq \nu\|\overline{\nabla_x^1 f}(X_k)\|$, it also follows that $\mathcal{F}_k$ is true whenever the event

$$\widetilde{\mathcal{F}}_k \overset{\text{def}}{=} \{|\Delta F(X_k, S_k) - \Delta f(X_k, S_k)| \leq 2\nu\|\overline{\nabla_x^1 f}(X_k)\|\min\{1, R_k\}\} \tag{4.2}$$

holds. As a consequence,

$$\Pr\left[\mathcal{M}_k \mid \mathcal{A}_{k-1}\right] \geq \Pr\left[\widetilde{\mathcal{M}}_k \mid \mathcal{A}_{k-1}\right] \text{ and } \Pr\left[\mathcal{F}_k \mid \mathcal{A}_{k-1}\right] \geq \Pr\left[\widetilde{\mathcal{F}}_k \mid \mathcal{A}_{k-1}\right]. \tag{4.3}$$

Our analysis of the impact of noise on the TR1NE algorithm starts by considering a relatively general form for error distributions (as we did in Theorem 2.3) and we then specialize our arguments to the particular case of finite sum minimization with subsampling.

4.1. Failure of AS.3 for general error distributions

At a generic iteration k, suppose that $H_{0,k}$ and $H_{1,k}$, are continuous and increasing random functions from $\mathbb{R}_+$ to $[0, 1]$ which are measurable for $\mathcal{A}_{k-1}$ and such that $H_{0,k}(0) = H_{1,k}(0) = 0$, $\lim_{\tau \to +\infty} H_{0,k}(\tau) = \lim_{\tau \to +\infty} H_{1,k}(\tau) = 1$ and,

$$\Pr\left[|\Delta F(X_k, S_k) - \Delta f(X_k, S_k)| < \tau | \mathcal{A}_{k-1}\right] \geq H_{0,k}(\tau)$$

$$\Pr\left[\|\nabla_x^1 f(X_k) - \overline{\nabla_x^1 f}(X_k)\| < \tau | \mathcal{A}_{k-1}\right] \geq H_{1,k}(\tau) \tag{4.4}$$

For sake of simplicity, assume $\alpha_* = \gamma_* \geq \sqrt{\frac{1}{2}}$ in AS.3 and let B_0 and B_1 such that $H_{0,k}(B_0) = \sqrt{\alpha_*}$ and $H_{1,k}(B_1) = \sqrt{\alpha_*}$, and $B = \max[B_0, B_1]$. Then,

$$\Pr\left[|\Delta F(X_k, S_k) - \Delta f(X_k, S_k)| < \tau \mid \mathcal{A}_{k-1}, \tau \geq B\right] \geq \sqrt{\alpha_*}, \tag{4.5}$$

$$\Pr\left[\|\nabla_x^1 f(X_k) - \overline{\nabla_x^1 f}(X_k)\| < \tau \mid \mathcal{A}_{k-1}, \tau \geq B\right] \geq \sqrt{\alpha_*}. \tag{4.6}$$

Define

$$\bar{B} = \frac{B}{\nu \min\{1, R_k\}} \geq \frac{B}{\nu} > B, \tag{4.7}$$

and note that $\bar{B}$ is measurable for $\mathcal{A}_{k-1}$. Then (4.5) and (4.6) ensure that

$$\Pr\left[|\Delta F(X_k, S_k) - \Delta f(X_k, S_k)| \leq \bar{B} \mid \mathcal{A}_{k-1}\right] \geq \sqrt{\alpha_*} \tag{4.8}$$

$$\Pr\left[\|\nabla_x^1 f(X_k) - \overline{\nabla_x^1 f}(X_k)\| \leq \bar{B} \mid \mathcal{A}_{k-1}\right] \geq \sqrt{\alpha_*}. \tag{4.9}$$

Finally, define the events

$$\mathcal{G}_k \overset{\text{def}}{=} \{\|\nabla_x^1 f(X_k)\| \geq 2\bar{B}\}, \tag{4.10}$$

$$\bar{\mathcal{G}}_k \overset{\text{def}}{=} \{\|\overline{\nabla_x^1 f}(X_k)\| \geq \bar{B}\}, \tag{4.11}$$

$$\mathcal{V}_k \overset{\text{def}}{=} \{\|\nabla_x^1 f(X_k) - \overline{\nabla_x^1 f}(X_k)\| < \frac{B}{\nu}\}, \tag{4.12}$$

and observe that (4.6) implies that

$$\Pr\left[\mathcal{V}_k \mid \mathcal{A}_{k-1}\right] \geq \sqrt{\alpha_*}. \tag{4.13}$$

The next theorem provides a lower bound on the probability of $\bar{\mathcal{G}}_k$ to occur, conditioned to the trace σ-algebra $\{\mathcal{A}_{k-1}, \mathcal{G}_k\}$.

Theorem 4.1. *Let $\bar{B}$ as in* (4.7). *Then, for each iteration k of the* TR1NE *algorithm,*

$$\Pr\left[\bar{\mathcal{G}}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k\right] \geq \sqrt{\alpha_*}. \tag{4.14}$$

Proof. For any realization of the TR1NE algorithm we have that

$$\|\overline{\nabla_x^1 f}(x_k)\| \geq \left|\|\nabla_x^1 f(x_k)\| - \|\nabla_x^1 f(x_k) - \overline{\nabla_x^1 f}(x_k)\|\right|.$$

Therefore, $\|\nabla_x^1 f(x_k)\| \geq 2\bar{\beta}$ (where $\bar{\beta}$ is the realization of $\bar{B}$) and $\|\nabla_x^1 f(x_k) - \overline{\nabla_x^1 f}(x_k)\| \leq \beta/\nu < \bar{\beta}$ ensure that $\|\overline{\nabla_x^1 f}(x_k)\| \geq \bar{\beta}$. Then, $\mathcal{G}_k \cap \mathcal{V}_k$ implies $\bar{\mathcal{G}}_k$, where the events $\mathcal{G}_k, \mathcal{V}_k$ and $\bar{\mathcal{G}}_k$ are defined in (4.10)-(4.12), and $\Pr\left[\bar{\mathcal{G}}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k, \mathcal{V}_k\right] = 1$. We therefore have that

$$\mathbb{E}\left[\mathbb{1}_{\bar{\mathcal{G}}_k} \mid \mathcal{A}_{k-1}, \mathcal{G}_k\right] \geq \mathbb{E}\left[\mathbb{1}_{\bar{\mathcal{G}}_k} \mid \mathcal{A}_{k-1}, \mathcal{G}_k, \mathcal{V}_k\right] \Pr\left[\mathcal{V}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k\right] \geq 1 \cdot \sqrt{\alpha_*},$$

where we have used (4.13) and the fact that

$$\mathbb{E}\Big[\mathbb{1}_{\mathcal{V}_k} \mid \mathcal{A}_{k-1}, \mathcal{G}_k\Big] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}_{\mathcal{V}_k} \mid \mathcal{A}_{k-1}\big] \mid \mathcal{A}_{k-1}, \mathcal{G}_k\Big]$$

to derive the last inequality. The conclusion (4.14) then follows. $\quad\square$

We are now in the position to show that the first requirement of (2.25) in AS.3 holds provided that the norm of the gradient is sufficiently large. Moreover, we prove that if the first requirement of (2.25) does not hold for a specific realization ω, then the norm of the gradient at x_k is smaller than $2\bar{B}(\omega)$.

Theorem 4.2. *Let $\bar{B}$ be defined by* (4.7). *Then, for each iteration k of the* TR1NE *algorithm,*

$$\Pr\Big[\mathcal{M}_k \mid \mathcal{A}_{k-1}, \big\{\|\nabla_x^1 f(X_k)\| \geq 2\bar{B}\big\}\Big] \geq \alpha_*. \qquad (4.15)$$

Moreover, if ω is a realization for which $\Pr\Big[\mathcal{M}_k \mid \mathcal{A}_{k-1}\Big](\omega) < \alpha_*$, then

$$\|\nabla_x^1 f(x_k)\| < 2\bar{B}(\omega). \qquad (4.16)$$

Proof. We obtain from (4.3) and (4.14) that

$$\begin{aligned}
\Pr\Big[\mathcal{M}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k\Big] &\geq \Pr\Big[\widetilde{\mathcal{M}}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k\Big] = \mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{M}}_k} \mid \mathcal{A}_{k-1}, \mathcal{G}_k\Big] \\
&\geq \mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{M}}_k} \mid \mathcal{A}_{k-1}, \mathcal{G}_k, \bar{\mathcal{G}}_k\Big]\Pr\Big[\bar{\mathcal{G}}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k\Big] \\
&\geq \sqrt{\alpha_*}\,\mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{M}}_k} \mid \mathcal{A}_{k-1}, \mathcal{G}_k, \bar{\mathcal{G}}_k\Big]. \qquad (4.17)
\end{aligned}$$

If $\bar{\mathcal{G}}_k$ is true, then it follows from (4.7) and (4.11) that $\nu\|\overline{\nabla_x^1 f}(X_k)\| \geq B$. Then, (4.1) and (4.6) yield

$$\mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{M}}_k} \mid \mathcal{A}_{k-1}, \bar{\mathcal{G}}_k\Big] \geq \sqrt{\alpha_*}. \qquad (4.18)$$

Because the trace σ-algebra $\{\mathcal{A}_{k-1}, \bar{\mathcal{G}}_k\}$ contains the trace σ- algebra $\{\mathcal{A}_{k-1}, \mathcal{G}_k, \bar{\mathcal{G}}_k\}$, the tower property and (4.18) then imply that

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{M}}_k} \mid \mathcal{A}_{k-1}, \mathcal{G}_k, \bar{\mathcal{G}}_k\Big] &= \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}_{\widetilde{\mathcal{M}}_k} \mid \mathcal{A}_{k-1}, \bar{\mathcal{G}}_k\big] \mid \mathcal{A}_{k-1}, \mathcal{G}_k, \bar{\mathcal{G}}_k\Big] \\
&\geq \mathbb{E}\Big[\sqrt{\alpha_*} \mid \mathcal{A}_{k-1}, \mathcal{G}_k, \bar{\mathcal{G}}_k\Big] = \sqrt{\alpha_*}
\end{aligned}$$

which, together with (4.17) gives (4.15). Since $\mathcal{G}_k$ is measurable for $\mathcal{A}_{k-1}$ we have

$$\Pr\Big[\mathcal{M}_k \mid \mathcal{A}_{k-1}\Big] \geq \Pr\Big[\mathcal{M}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k\Big]\mathbb{E}\Big[\mathbb{1}_{\mathcal{G}_k} \mid \mathcal{A}_{k-1}\Big] = \Pr\Big[\mathcal{M}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k\Big]\mathbb{1}_{\mathcal{G}_k}.$$

If we now consider a realization ω such that $\Pr\left[\mathcal{M}_k \mid \mathcal{A}_{k-1}\right](\omega) < \alpha_*$, we therefore obtain, using (4.15) taken for the realization ω, that

$$\alpha_* > \Pr\left[\mathcal{M}_k \mid \mathcal{A}_{k-1}, \mathcal{G}_k\right](\omega)\,\mathbb{1}_{\mathcal{G}_k(\omega)} \geq \alpha_* \,\mathbb{1}_{\mathcal{G}_k(\omega)},$$

which implies that $\mathbb{1}_{\mathcal{G}_k(\omega)} = 0$, and thus that (4.16) holds. $\square$

Similarly, but with reference to the second requirement in (2.25), the following theorem can be proved.

Theorem 4.3. *Let $\bar{B}$ be defined by (4.7). Then, for each iteration k of the TR1NE algorithm,*

$$\Pr\left[\mathcal{F}_k \mid \mathcal{A}_{k-1}, \left\{\|\nabla_x^1 f(X_k)\| \geq 2\bar{B}\right\}\right] \geq \alpha_*. \tag{4.19}$$

Moreover, if ω is a realization for which $\Pr\left[\mathcal{F}_k \mid \mathcal{A}_{k-1}\right](\omega) < \alpha_*$, then

$$\|\nabla_x^1 f(x_k)\| < 2\bar{B}(\omega). \tag{4.20}$$

Proof. The proof is similar to that of Theorem 4.2, and is given in Appendix A for completeness. $\square$

Theorems 4.2 and 4.3 indicate that the assumptions made in AS.3 about $\mathcal{M}_k$ and $\mathcal{F}_k$ are likely to be satisfied as long as the gradients remain sufficiently large, allowing the TR1NE algorithm to iterate meaningfully. Conversely, they show that, should these assumptions fail for a particular realization of the algorithm because of a high level of intrinsic noise, "degraded" versions of first-order optimality conditions given by (4.16) and (4.20) nevertheless hold when this failure occurs.

4.2. A subsampling example

We finally illustrate how intrinsic noise might affect our probabilistic framework on an example. Suppose that

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x), \tag{4.21}$$

where the f_i are functions from $\mathbb{R}$ to $\mathbb{R}$ having Lipschitz continuous gradients and where m is so large that computing the complete value of $f(x)$ or its derivatives is impractical. Such a situation occurs for instance in algorithms for deep-learning, an application of

growing importance. A well-known strategy to obtain approximations of the desired values at an iterate x_k is to sample the $f_i(x_k)$ and compute the sample averages, that is

$$\overline{f}(x_k) = \frac{1}{|\mathfrak{b}_0(x_k)|} \sum_{i \in \mathfrak{b}_0(x_k)} f_i(x_k), \quad \overline{\nabla_x^1 f}(x_k) = \frac{1}{|\mathfrak{b}_1(x_k)|} \sum_{i \in \mathfrak{b}_1(x_k)} \nabla_x^1 f_i(x_k), \qquad (4.22)$$

where $\mathfrak{b}_0(x)$ and $\mathfrak{b}_1(x)$ are realizations of random "batches", that is randomly selected[4] subsets of $\{1, \ldots, m\}$. Observe that Step 3 of the TRINE algorithm computes the estimate $\overline{f}(x_k) - \overline{f}(x_k + s_k)$, which we assume, in the context of (4.21), to be

$$\overline{f}(x_k) - \overline{f}(x_k + s_k) = \frac{1}{|\mathfrak{b}_0(x_k)|} \sum_{i \in \mathfrak{b}_0(x)} (f_i(x_k) - f_i(x_k + s_k)), \qquad (4.23)$$

(using a single batch for both the function estimates). Observe that our choice to make $\mathfrak{b}_0$ and $\mathfrak{b}_1$ dependent on x_k implies that their random counterparts $\mathfrak{B}_0(X_k)$ and $\mathfrak{B}_1(X_k)$ are measurable for $\mathcal{A}_{k-1}$ (clearly we could have chosen a more complicated dependence on the past of the random process). The mean-value theorem then yields that, for some $\{y_i\}_{i \in \mathfrak{b}_0(x_k)}$ in the segment $[x_k, x_k + s_k]$,

$$|\overline{f}(x_k) - \overline{f}(x_k + s_k)| \leq \left(\frac{1}{|\mathfrak{b}_0(x_k)|} \sum_{i \in \mathfrak{b}_0(x_k)} \nabla_x^1 f_i(y_i) \right) \|s_k\| \leq r_k \max_{\substack{y \in [x_k, x_k + s_k] \\ i \in \mathfrak{b}_0(x_k)}} \|\nabla_x^1 f_i(y)\|$$

Note that one expects the right-hand side of this inequality to be quite small when the trust-region radius is small or when convergence to a local minimizer occurs. To simplify our illustration, we assume, for the rest of this section, that there exists a constant κ_f such that for any $y \in \mathbb{R}^n$, for every realization $\mathfrak{b}_0(x_k)$,

$$r_k \max_{i \in \{1, \ldots, n\}} \|\nabla_x^1 f_i(y)\| \leq \kappa_f.$$

Returning to the random process and using the Bernstein concentration inequality, it results from [5, Relation (7.8)] that, for any k and deterministic $\tau > 0$,

$$\Pr\left[\Delta F(X_k, S_k) - \Delta f(X_k, S_k) > \tau\right] \leq e^{-W_0(\tau)} \text{ where } W_0(\tau) = \frac{\tau^2 |\mathfrak{B}_0(X_k)|}{4\kappa_f(2\kappa_f + \frac{1}{3}\tau)}. \quad (4.24)$$

Similarly,

$$\Pr\left[\|\nabla_x^1 F(X_k, S_k) - \nabla_x^1 f(X_k, S_k)\| > \tau\right] \leq \min\left[1, (n+1)e^{-W_1(\tau)}\right],$$

$$W_1(\tau) = \frac{\tau^2 |\mathfrak{B}_1(X_k)|}{4\kappa_g(2\kappa_g + \frac{1}{3}\tau)}, \qquad (4.25)$$

---

[4] With uniform distribution.

for some constant $\kappa_g > 0$. One also checks that, since $\mathfrak{B}_0(X_k)$ and $\mathfrak{B}_1(X_k)$ are measurable for $\mathcal{A}_{k-1}$, so are W_0 and W_1. One then easily verifies that $W_0(\tau)$ is an increasing function of τ, and hence $e^{-W_0(\tau)}$ is decreasing. Letting $G_k(\tau) = 1 - e^{-W_0(\tau)}$, we immediately obtain that conditions (2.27) and (2.28) hold. Let us now analyze condition (2.29) and consider any realization ω, where $w_0(\tau) \overset{\text{def}}{=} W_0(\omega)(\tau)$. Note that $w_0(\tau) \geq |\mathfrak{b}_0(x_k)|^{\frac{1}{2}}\tau$ when

$$\tau \geq \tau_* \overset{\text{def}}{=} \frac{8\kappa_f^2}{|\mathfrak{b}_0(x_k)|^{\frac{1}{2}} - \frac{4}{3}\kappa_f} \tag{4.26}$$

and $\frac{|\mathfrak{b}_0(x_k)|^{\frac{1}{2}}}{\kappa_f}$ is large enough so that

$$\frac{|\mathfrak{b}_0(x_k)|^{\frac{1}{2}}}{\kappa_f} > \frac{4}{3}. \tag{4.27}$$

Hence $e^{-w_0(\tau)} \leq e^{-|\mathfrak{b}_0(x_k)|^{\frac{1}{2}}\tau}$ for all $\tau \geq \tau_*$ and

$$\int_0^\infty e^{-w_0(\tau)}\, d\tau \leq \int_0^{\tau_*} e^{-w_0(\tau)}\, d\tau + \int_{\tau_*}^\infty e^{-|\mathfrak{b}_0(x_k)|^{\frac{1}{2}}\tau}\, d\tau.$$

In addition, since $e^{-w_0(\tau)}$ is decreasing and non-negative, we have that

$$\int_0^{\tau_*} e^{-w_0(\tau)}\, d\tau \leq \tau_* e^{-w_0(0)} = \tau_*.$$

This bound and (4.26) then imply that

$$\int_0^\infty (1 - g_k(\tau))\, d\tau = \int_0^\infty e^{-w_0(\tau)}\, d\tau \leq \tau_* + \frac{e^{-|\mathfrak{b}_0(x_k)|^{\frac{1}{2}}\tau_*}}{|\mathfrak{b}_0(x_k)|^{\frac{1}{2}}} < +\infty, \tag{4.28}$$

proving that (2.29) also holds for the arbitrary realization ω. We may therefore apply Theorem 2.3 provided $\mathfrak{B}_0(X_k)$ is sufficiently large[5] to ensure (4.27) and (4.26), and conclude that, under these conditions, (2.31) holds whenever

$$\Delta t_{f,1}(x_k, s_k) \geq \frac{1}{\eta} \int_0^\infty (1 - g_k(\tau))\, d\tau.$$

---

[5] While the bound given by (4.28) is adequate for our proof, this inequality can be pessimistic. For instance, if we set $\kappa_f = 1$ and $|\mathfrak{b}_0(X_k)| = 2056$, the numerically computed value of the left-hand side is 0.0556 while that of the right-hand side is 0.1818.

We can also apply the analysis in Section 4.1 with

$$H_{0,k}(\tau) = 1 - e^{-W_0(\tau)} \quad \text{and} \quad H_{1,k}(\tau) = \max\left[0, 1 - (n+1)e^{-W_1(\tau)}\right].$$

A short calculation shows that $B_0 = \mathcal{O}(\kappa_f|\mathfrak{B}_0|^{-\frac{1}{2}})$ and $B_1 = \mathcal{O}(\kappa_g|\mathfrak{B}_1|^{-\frac{1}{2}})$, where B_0 and B_1 are defined below (4.4). Then, Theorems 4.2 and 4.3 hold with $\bar{B} = \mathcal{O}\left(\frac{\max\{\kappa_f|\mathfrak{B}_0|^{-\frac{1}{2}}, \kappa_g|\mathfrak{B}_1|^{-\frac{1}{2}}\}}{\nu \max\{1, R_k\}}\right)$.

We finally illustrate the impact of intrinsic noise on the (admittedly ad-hoc) problem of minimizing

$$f(x) = \frac{1}{2m}\sum_{i \in \mathcal{Z}}\left[\tfrac{1}{2}x^2 + \tfrac{1}{2}\alpha\,\text{sgn}(i)\,e^{-x^2}\right] \overset{\text{def}}{=} \frac{1}{2m}\sum_{i \in \mathcal{Z}} f_i(x), \tag{4.29}$$

where $\alpha > 0$ is a noise level and where $\mathcal{Z} = \{-m, \ldots, m\} \setminus \{0\}$ for some large integer m. Suppose furthermore that the $\{f_i(x)\}_{i \in \mathcal{Z}}$ and $\{\nabla_x^1 f_i(x)\}_{i \in \mathcal{Z}}$ are computed by black-box routines, therefore hiding their relationships. Consider an iterate x_k at the start of iteration k of an arbitrary realization of the TRINE algorithm[6] applied to this problem. We verify that, for $i \in \mathcal{Z}$,

$$\nabla_x^1 f_i(x_k) = x_k\left(1 - \alpha\,\text{sgn}(i)\,e^{-x_k^2}\right)$$

and thus $\nabla_x^1 f(x_k) = x_k$ and, in view of (2.5),

$$\phi_{f,1}^{\delta_k}(x_k) = |x_k|\,\delta_k \tag{4.30}$$

for all x_k. As a consequence, $x = 0$ is the unique global minimizer of $f(x)$. Suppose, for the rest of this section, that $\mathfrak{B}_{0,k} \overset{\text{def}}{=} \mathfrak{B}_0(x_k) = \mathfrak{B}_0(x_k + s_k)$, $\mathfrak{B}_{1,k} \overset{\text{def}}{=} \mathfrak{B}_1(x_k)$, and that $n_{0,k}^{\mathfrak{B}}$ and $n_{1,k}^{\mathfrak{B}}$, the cardinalities of these two sets are known parameters. We deduce from (4.22) that

$$\overline{\nabla_x^1 f}(x_k) = x_k\left(1 - \alpha\Psi(\mathfrak{B}_{1,k})e^{-x_k^2}\right) \quad \text{where} \quad \Psi(\mathfrak{B}) \overset{\text{def}}{=} \frac{1}{|\mathfrak{B}|}\sum_{i \in \mathfrak{B}} \text{sgn}(i). \tag{4.31}$$

Thus $\Psi(\mathfrak{B})$ is a zero-mean random variable with values in $[-1, 1]$, depending on the randomly chosen batch $\mathfrak{B} \subseteq \mathcal{Z}$ of size $|\mathfrak{B}|$. Using the hypergeometric distribution, it is possible to show that $|\Psi(\mathfrak{B})|$ is (in probability) a decreasing function of $|\mathfrak{B}|$.

Moreover, the use of standard tail bounds [14] reveals that, for any $t \in (0, 1)$,

$$\Pr\left[|\Psi(\mathfrak{B}_{1,k})| \le t\right] = \Pr\left[\Psi(\mathfrak{B}_{1,k}) \le t\right]^2 = \left(1 - \Pr\left[\Psi(\mathfrak{B}_{1,k}) > t\right]\right)^2 \ge (1 - e^{-\frac{1}{2}t^2 n_{1,k}^{\mathfrak{B}}})^2, \tag{4.32}$$

---

[6] With given ν and $\varsigma = 1$.

in turn indicating that $\mathbb{P}r\left[|\Psi(\mathfrak{B}_{1,k})| \leq t\right] > \frac{1}{2}$ whenever

$$n_{1,k}^{\mathfrak{B}} \geq \frac{2}{t^2} \left|\log\left(1 - \frac{1}{\sqrt{2}}\right)\right| \approx \frac{2.4559}{t^2}.$$

Occurrence of $\mathcal{M}_k^{(1)}$ and $\mathcal{M}_k^{(2)}$. Let us now examine at what conditions the events $\mathcal{M}_k^{(1)}$ and $\mathcal{M}_k^{(2)}$ do occur for a specific realization $\mathfrak{b}_{1,k}$ of $\mathfrak{B}_{1,k}$, and consider the occurrence of $\mathcal{M}_k^{(1)}$ first. Because the minimum of first-order models in a ball of radius δ_k must occur on the boundary, we choose $d_{k,1} = -\text{sgn}(\overline{\nabla_x^1 f}(x_k))\delta_k$ so that

$$\overline{\Delta t}_{f,1}(x_k, d_{k,1}) = |\overline{\nabla_x^1 f}(x_k)| \, \delta_k.$$

Using (4.31), we then have that

$$\overline{\Delta t}_{f,1}(x_k, d_{k,1}) = |x_k| \, \delta_k \, \left|1 - \alpha\Psi(\mathfrak{b}_{1,k}) \, e^{-x_k^2}\right|. \tag{4.33}$$

Thus the quantity $1 - \alpha\Psi(\mathfrak{b}_{1,k}) \, e^{-x_k^2}$ may be interpreted as the local noise relative to the model decrease.

Taking (4.30) and (4.33) into account, $\mathcal{M}_k^{(1)}$ occurs, in any realization, whenever

$$\left|1 - \alpha\Psi(\mathfrak{b}_{1,k}) \, e^{-x_k^2}\right| \geq \frac{1}{1+\nu}, \tag{4.34}$$

that is

$$\Psi(\mathfrak{b}_{1,k}) \leq \frac{e^{x_k^2}}{\alpha} \left(1 - \frac{1}{1+\nu}\right) \quad \text{or} \quad \Psi(\mathfrak{b}_{1,k}) \geq \frac{e^{x_k^2}}{\alpha} \left(1 + \frac{1}{1+\nu}\right).$$

This condition may be quite weak, as shown in left picture in Fig. 4.1, where the shape of the left-hand side of (4.34) is shown for increasing values[7] of the local noise level $\alpha\exp(-x^2)$ as a function of $\Psi(\mathfrak{b}_{1,k})$, and where the lower bound $1/(1+\nu)$ is shown as a red horizontal dashed line. The corresponding ranges of acceptable values of $\Psi(\mathfrak{b}_{1,k})$ are shown below the horizontal axis (in matching colors). The one-sided nature of the inequality defining $\mathcal{M}_k^{(1)}$ is apparent in the picture, where restrictions on the acceptable values of $\Psi(\mathfrak{b}_{1,k})$ only occur for positive values. This reflects the fact that the model may be quite inaccurate and yet produce a decrease which is large enough for the condition to hold.

The constraints on $\Psi(\mathfrak{b}_{1,k})$, and thus on $n_{1,k}^{\mathfrak{B}}$, become more stringent when considering the occurrence of $\mathcal{M}_k^{(2)}$. Since, for any realization, $s_k = -\text{sgn}(\overline{\nabla_x^1 f}(x_k))r_k$, we deduce from (4.31) that

---

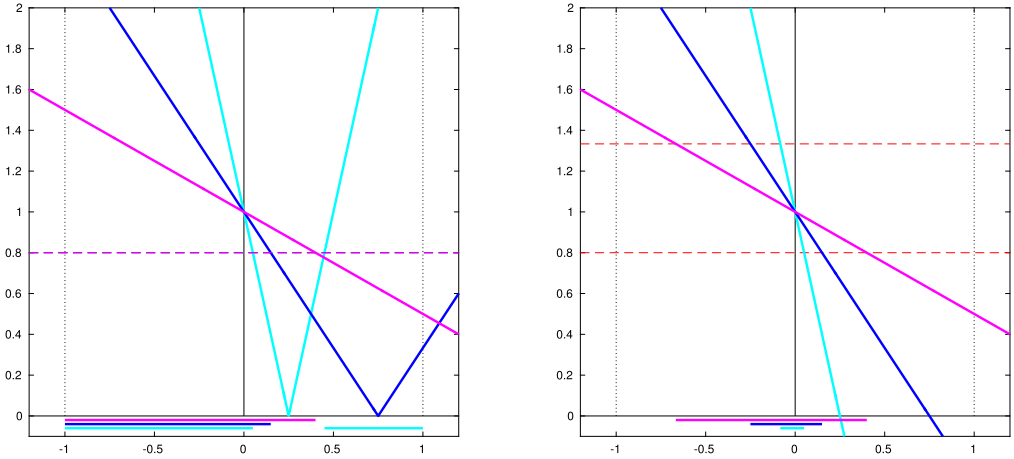[7] Magenta for 0.5, blue for 4/3 and cyan for 4.

Fig. 4.1. An illustration of conditions (4.34) (left) and (4.36) (right) as a function of $\Psi(\mathfrak{b}_{1,k})$, for $\nu = \frac{1}{4}$ and local relative noise levels $\alpha e^{-x_k^2} = \frac{1}{2}$ (magenta), $\frac{4}{3}$ (blue) and 4 (cyan). Acceptable ranges for $\Psi(\mathfrak{b}_{1,k})$ are shown below the horizontal axis in matching colors.

$$\Delta t_{f,1}(x_k, s_k) = -x_k s_k = -x_k \left[-\text{sgn}(x_k) \, \text{sgn}\left(1 - \alpha \, \Psi(\mathfrak{b}_{1,k}) \, e^{-x_k^2}\right) r_k \right]$$

$$= |x_k| \, r_k \, \text{sgn}\left(1 - \alpha \, \Psi(\mathfrak{b}_{1,k}) \, e^{-x_k^2}\right)$$

and

$$\overline{\Delta t}_{f,1}(x_k, s_k) = |x_k| \, r_k \left| 1 - \alpha \Psi(\mathfrak{b}_{1,k}) \, e^{-x_k^2} \right|. \tag{4.35}$$

One then verifies that $\mathcal{M}_k^{(2)}$ occurs whenever

$$\frac{1}{1+\nu} \leq 1 - \alpha \Psi(\mathfrak{b}_{1,k}) \, e^{-x_k^2} \leq \frac{1}{1-\nu}. \tag{4.36}$$

The acceptable values for $\Psi(\mathfrak{b}_{1,k})$ are illustrated in the right picture of Fig. 4.1, which shows the shape of the central term in (4.36) using the same conventions than for the left picture except that now the acceptable part of the curves lies between the lower and upper bounds resulting from (4.36) (again shown as dashed red lines). A short calculation reveals that (4.36) is equivalent to requiring

$$\frac{e^{x_k^2}}{\alpha}\left(1 - \frac{1}{1-\nu}\right) \leq \Psi(\mathfrak{b}_{1,k}) \leq \frac{e^{x_k^2}}{\alpha}\left(1 - \frac{1}{1+\nu}\right).$$

This therefore defines intervals around the origin, whose widths clearly decrease with the local relative noise level. Because $|\Psi(\mathfrak{B}_{1,k})|$ is (in probability) a decreasing function of $n_{1,k}^{\mathfrak{B}}$, this indicates that $n_{1,k}^{\mathfrak{B}}$ must increase with $\alpha e^{-x_k^2}$, that is when the local relative noise is large.

Occurrence of $\mathcal{F}_k$. A similar reasoning holds when considering the event $\mathcal{F}_k$. Given (4.23), we have that

$$|\Delta f(x_k, s_k) - \overline{\Delta f}(x_k, s_k)| = \tfrac{1}{2}\alpha\,|\Psi(\mathfrak{b}_{0,k})|\,\left|e^{-x_k^2} - e^{-(x_k+s_k)^2}\right| \tag{4.37}$$

and, in view of (4.35), $\mathcal{F}_k$ thus occurs whenever

$$\tfrac{1}{2}\alpha\,|\Psi(\mathfrak{b}_{0,k})|\,e^{-x_k^2}\,\left|1 - e^{-(2x_k s_k + s_k^2)}\right| \le 2\nu|x_k|\,r_k\,\left|1 - \alpha\Psi(\mathfrak{b}_{1,k})\,e^{-x_k^2}\right|. \tag{4.38}$$

Thus, if $|x_k|$ is small (e.g., if the optimum is close) then satisfying (4.38) requires the left-hand side of this inequality to be small, putting a high request on $n_{0,k}^{\mathfrak{B}}$, while the inequality is more easily satisfied if $|x_k|$ is large, irrespective of the batch sizes. Note that, in the first case (i.e., when $|x_k|$ is small), the request on $n_{0,k}^{\mathfrak{B}}$ is stronger for smaller $n_{1,k}^{\mathfrak{B}}$.

Occurrence of (2.31). Given (4.32) and (4.35), we see from Theorem 2.3 that (2.31) holds whenever

$$\overline{\Delta t}_{f,1}(x_k, s_k) = |x_k|\,r_k\,\left|1 - \alpha\Psi(\mathfrak{b}_{1,k})\,e^{-x_k^2}\right| \ge \frac{1}{\eta}\int_0^\infty (1 - g_k(\tau))\,d\tau = \frac{1}{\eta}\sqrt{\frac{\pi}{2n_{1,k}^{\mathfrak{B}}}}$$

where we have used the definition of the erf function to derive the last equality. Thus, as $|\Psi(\mathfrak{b}_{1,k})| \le 1$, guaranteeing (2.31) requires a larger $n_{1,k}^{\mathfrak{B}}$ for small value of x_k, that is when the optimum is approached.

5. Conclusions and perspectives

We have considered a trust-region method for unconstrained minimization inspired by [10] which is adapted to handle randomly perturbed function and derivatives values and is capable of finding approximate minimizers of arbitrary order. Exploiting ideas of [12,7], we have shown that its evaluation complexity is (in expectation) of the same order in the requested accuracy as that known for related deterministic methods [7,10].

In [4], the authors have considered the effect of intrinsic noise on complexity of a deterministic, noise tolerant variant of the trust-region algorithm. This important question is handled here by considering specific realizations of the algorithm under reasonable assumptions on the cumulative distribution of errors in the evaluations of the objective function and its derivatives. We have shown that, for such realizations, a first-order version of our trust-region algorithms still provides "degraded" optimality guarantees, should intrinsic noise cause the assumptions used for the complexity analysis to fail. We have specialized and illustrated those results in the case of sampling-based finite-sum minimization, a context of particular interest in deep-learning applications.

We have so far developed and analyzed "noise-aware" deterministic and stochastic algorithms for unconstrained optimization. Several research perspectives are natural extensions of the type of analysis presented here. Firstly, a challenging investigation is

to design adaptive regularization algorithms for smooth unconstrained functions whose values and derivatives are subject to random noise. Such a study would improve upon [3] where derivatives are subject to random noise while the objective function values are subject to a deterministically controlled error. While results are available in this reference for the first-order version of adaptive regularization algorithms since it is basically the same scheme as Algorithm TR1NE on page 24, deriving results for an arbitrary order q remains an open problem. Secondly, it is relevant to extend the study conducted in Section 4 and analyze the trust-region procedure under non-vanishing noise and with an arbitrary values of q. Finally, the constrained case deserves study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Additional proofs

Proof of (3.12). Since $\sigma(\mathbb{1}_{\Lambda_k^c})$ belong to $\mathcal{A}_{k-1}$, because the random variable Λ_k is fully determined, assuming $Pr(\mathbb{1}_{\Lambda_k^c}) > 0$, the tower property yields:

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_k} \mid \mathbb{1}_{\Lambda_k^c}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_k} \mid \mathcal{A}_{k-1}\right] \mid \mathbb{1}_{\Lambda_k^c}\right] \geq \mathbb{E}\left[p_* \mid \mathbb{1}_{\Lambda_k^c}\right] = p_*.$$

Then, by the total expectation law we have

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_k}\mathbb{1}_{\Lambda_k^c}\right] = \mathbb{E}\left[\mathbb{1}_{\Lambda_k^c}\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_k} \mid \mathbb{1}_{\Lambda_k^c}\right]\right] \geq p_*\mathbb{E}\left[\mathbb{1}_{\Lambda_k^c}\right].$$

Similarly,

$$\mathbb{E}\left[\mathbb{1}_{\{k<N_\epsilon\}}\mathbb{1}_{\mathcal{E}_k}\mathbb{1}_{\Lambda_k^c}\right] \geq p_*\mathbb{E}\left[\mathbb{1}_{\{k<N_\epsilon\}}\mathbb{1}_{\Lambda_k^c}\right],$$

as $\mathbb{1}_{\{k<N_\epsilon\}}$ is also determined by $\mathcal{A}_{k-1}$. In case $Pr(\mathbb{1}_{\Lambda_k^c}) = 0$, the above inequality holds trivially. Then

$$\mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1}\mathbb{1}_{\Lambda_k^c}\mathbb{1}_{\mathcal{E}_k}\right] = \mathbb{E}\left[\sum_{k=0}^{\infty}\mathbb{1}_{\{k<N_\epsilon\}}\mathbb{1}_{\Lambda_k^c}\mathbb{1}_{\mathcal{E}_k}\right] \geq p_M\mathbb{E}\left[\sum_{k=0}^{\infty}\mathbb{1}_{\{k<N_\epsilon\}}\mathbb{1}_{\Lambda_k^c}\right]$$

$$= p_*\,\mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1}\mathbb{1}_{\Lambda_k^c}\right],$$

and (3.12) follows. □

Proof of Lemma 3.4. Proceeding as in the proof of (3.12) with $\mathbb{1}_{\Lambda_k}$ in place of $\mathbb{1}_{\Lambda_k^c}$, we obtain:

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

Moreover, proceeding again as in the proof of (3.12) and substituting $\mathbb{1}_{\mathcal{E}_k}$ with $\mathbb{1}_{\mathcal{E}_k^c}$ we obtain

$$\mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1}\mathbb{1}_{\Lambda_k^c}\mathbb{1}_{\mathcal{E}_k^c}\right] \leq (1-p_*)\,\mathbb{E}\left[\sum_{k=0}^{N_\epsilon-1}\mathbb{1}_{\Lambda_k}\right].$$

Using the above inequalities we obtain (3.16). $\square$

Proof of Theorem 4.3. Because of (4.3) and (4.14), we have that

$$\begin{aligned}
\mathrm{Pr}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-1},\mathcal{G}_k\Big] &\geq \mathrm{Pr}\Big[\widetilde{\mathcal{F}}_k \mid \mathcal{A}_{k-1},\mathcal{G}_k\Big] = \mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{F}}_k} \mid \mathcal{A}_{k-1},\mathcal{G}_k\Big] \\
&\geq \mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{F}}_k} \mid \mathcal{A}_{k-1},\mathcal{G}_k,\bar{\mathcal{G}}_k\Big]\,\mathrm{Pr}\Big[\bar{\mathcal{G}}_k \mid \mathcal{A}_{k-1},\mathcal{G}_k\Big] \\
&\geq \sqrt{\alpha_*}\,\mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{F}}_k}\mid\mathcal{A}_{k-1},\mathcal{G}_k,\bar{\mathcal{G}}_k\Big].
\end{aligned} \tag{A.1}$$

If $\bar{\mathcal{G}}_k$ is true then by (4.7) it follows

$$2\nu\|\overline{\nabla_x^1 f}(X_k)\|\min\{1,R_k\} \geq \bar{B}\nu\min\{1,R_k\} > B.$$

Then, (4.2) and (4.5) yield

$$\mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{F}}_k} \mid \mathcal{A}_{k-\frac{1}{2}},\bar{\mathcal{G}}_k\Big] \geq \sqrt{\alpha_*}. \tag{A.2}$$

Because the trace σ-algebra $\{\mathcal{A}_{k-1},\bar{\mathcal{G}}_k\}$ contains the trace σ- algebra $\{\mathcal{A}_{k-1},\mathcal{G}_k,\bar{\mathcal{G}}_k\}$, the tower property and (A.2) then imply that

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{F}}_k} \mid \mathcal{A}_{k-1},\mathcal{G}_k,\bar{\mathcal{G}}_k\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{\widetilde{\mathcal{F}}_k} \mid \mathcal{A}_{k-1},\bar{\mathcal{G}}_k\Big] \mid \mathcal{A}_{k-1},\mathcal{G}_k,\bar{\mathcal{G}}_k\Big] \\
&\geq \mathbb{E}\Big[\sqrt{\alpha_*} \mid \mathcal{A}_{k-1},\mathcal{G}_k,\bar{\mathcal{G}}_k\Big] = \sqrt{\alpha_*}
\end{aligned}$$

which, together with (A.1), implies (4.19). Since $\mathcal{G}_k$ is measurable for $\mathcal{A}_{k-1}$ we have that

$$\mathrm{Pr}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-1}\Big] \geq \mathrm{Pr}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-1},\mathcal{G}_k\Big]\,\mathbb{E}\Big[\mathbb{1}_{\mathcal{G}_k} \mid \mathcal{A}_{k-1}\Big] = \mathrm{Pr}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-1},\mathcal{G}_k\Big]\mathbb{1}_{\mathcal{G}_k}.$$

Considering now a realization ω such that $\mathrm{Pr}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-1}\Big](\omega) < \alpha_*$, we therefore obtain, using (4.19) taken for this realization, that

$$\alpha_* > \mathrm{Pr}\Big[\mathcal{F}_k \mid \mathcal{A}_{k-1},\mathcal{G}_k\Big](\omega)\,\mathbb{1}_{\mathcal{G}_k(\omega)} \geq \alpha_*\,\mathbb{1}_{\mathcal{G}_k(\omega)},$$

which implies that $\mathbb{1}_{\mathcal{G}_k(\omega)} = 0$, in turn yielding (4.20). $\square$

References

[1] A.S. Bandeira, K. Scheinberg, L.N. Vicente, Convergence of trust-region methods based on probabilistic models, SIAM J. Optim. 24 (3) (2014) 1238–1264.

[2] S. Bellavia, G. Gurioli, Complexity analysis of a stochastic cubic regularisation method under inexact gradient evaluations and dynamic Hessian accuracy, Optimization 71 (2022) 227–261.

[3] S. Bellavia, G. Gurioli, B. Morini, Ph.L. Toint, Adaptive regularization for nonconvex optimization using inexact function values and randomly perturbed derivatives, J. Complex. 68 (2022) 101591.

[4] S. Bellavia, G. Gurioli, B. Morini, Ph.L. Toint, The impact of noise on evaluation complexity: the deterministic trust-region case, arXiv:2104.02519, 2021.

[5] S. Bellavia, G. Gurioli, B. Morini, Ph.L. Toint, Adaptive regularization algorithms with inexact evaluations for nonconvex optimization, SIAM J. Optim. 29 (2019) 2881–2915.

[6] A. Berahas, L. Cao, K. Scheinberg, Global convergence rate analysis of a generic line search algorithm with noise, SIAM J. Optim. 31 (2) (2021) 1489–1518.

[7] J. Blanchet, C. Cartis, M. Menickelly, K. Scheinberg, Convergence rate analysis of a stochastic trust region method via supermartingales, INFORMS J. Optim. 1 (2) (2019) 92–119.

[8] C. Cartis, N.I.M. Gould, Ph.L. Toint, Second-order optimality and beyond: characterization and evaluation complexity in convexly constrained nonlinear optimization, Found. Comput. Math. 18 (2020) 1073–1107.

[9] C. Cartis, N.I.M. Gould, Ph.L. Toint, Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints, SIAM J. Optim. 30 (1) (2020) 513–541.

[10] C. Cartis, N.I.M. Gould, Ph.L. Toint, Strong evaluation complexity of an inexact trust-region algorithm for arbitrary-order unconstrained nonconvex optimization, arXiv:2011.00854, 2020.

[11] C. Cartis, N.I.M. Gould, Ph.L. Toint, Evaluation Complexity of Algorithms for Nonconvex Optimization, MOS-SIAM Series on Optimization, vol. 30, SIAM, Philadelphia, USA, 2022.

[12] C. Cartis, K. Scheinberg, Global convergence rate analysis of unconstrained optimization methods based on probabilistic models, Math. Program., Ser. A 159 (2) (2018) 337–375.

[13] R. Chen, M. Menickelly, K. Scheinberg, Stochastic optimization using a trust-region method and random models, Math. Program., Ser. A 169 (2) (2018) 447–487.

[14] V. Chvátal, The tail of the hypergeometric distribution, Discrete Math. 25 (1979) 285–287.

[15] A.R. Conn, N.I.M. Gould, Ph.L. Toint, Trust-Region Methods, MOS-SIAM Series on Optimization 1, SIAM, Philadelphia, USA, 2000.

[16] C. Paquette, K. Scheinberg, A stochastic line search method with convergence rate analysis, SIAM J. Optim. 30 (1) (2020) 349–376.

[17] Y. Yuan, Recent advances in trust region algorithms, Math. Program., Ser. A 151 (1) (2015) 249–281.