

# Convergence Rate Analysis of a Stochastic Trust Region Method via Supermartingales

Jose Blanchet\*    Coralia Cartis<sup>†</sup>    Matt Menickelly<sup>‡</sup>    Katya Scheinberg<sup>§</sup>

October 19, 2018

## Abstract

We propose a novel framework for analyzing convergence rates of stochastic optimization algorithms with adaptive step sizes. This framework is based on analyzing properties of an underlying generic stochastic process, in particular by deriving a bound on the expected stopping time of this process. We utilize this framework to analyze the bounds on expected global convergence rates of a stochastic variant of a traditional trust region method, introduced in [8]. While traditional trust region methods rely on exact computations of the gradient, Hessian and values of the objective function, this method assumes that these values are available up to some dynamically adjusted accuracy. Moreover, this accuracy is assumed to hold only with some sufficiently large, but fixed, probability, without any additional restrictions on the variance of the errors. This setting applies, for example, to standard stochastic optimization and machine learning formulations. Improving upon the analysis in [8], we show that the stochastic process defined by the algorithm satisfies the assumptions of our proposed general framework, with the stopping time defined as reaching accuracy  $\|\nabla f(x)\| \leq \epsilon$ . The resulting bound for this stopping time is  $O(\epsilon^{-2})$ , under the assumption of sufficiently accurate stochastic gradient, and is the first global complexity bound for a stochastic trust-region method. Finally, we apply the same framework to derive second order complexity bound under some additional assumptions.

## 1 Introduction

In this paper we aim to solve a stochastic unconstrained, possibly nonconvex, optimization problem

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

---

\*Department of Industrial Engineering and Operations Research, Columbia University 500 West 120th street, New York, NY 10027, USA. [jose.blanchet@columbia.edu](mailto:jose.blanchet@columbia.edu).

<sup>†</sup>Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, United Kingdom. [cartis@maths.ox.ac.uk](mailto:cartis@maths.ox.ac.uk). This work was partially supported by the Oxford University EPSRC Platform Grant EP/I01893X/1.

<sup>‡</sup>Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439, USA. [mmenickelly@anl.gov](mailto:mmenickelly@anl.gov). The work of this author was partially supported by NSF Grant DMS 13-19356.

<sup>§</sup>Department of Industrial and Systems Engineering, Lehigh University, Harold S. Mohler Laboratory, 200 West Packer Avenue, Bethlehem, PA 18015-1582, USA. [katyas@lehigh.edu](mailto:katyas@lehigh.edu). The work of this author was partially supported by NSF Grants, DMS 13-19356, CCF 13-20137 and CCF 16-18717. It was partially performed while the author was visiting the University of Oxford, sponsored by the Visiting Professorship Grant VP1-2016-053 from the Leverhulme Trust.

where  $f(x)$  is a function which is assumed to be smooth and bounded from below, and whose value can only be computed with some noise. Let  $\tilde{f}(x, \xi)$  be the noisy computable version of  $f$ , where the noise  $\xi$  is a random variable. One of the most common settings is

$$f(x) = \mathbb{E}_{\xi}[\tilde{f}(x, \xi)].$$

Stochastic optimization methods, in particular stochastic gradient descent (SGD), have recently become the focus of much research in optimization, especially in applications to machine learning domains. This is because in machine learning the objective function of the optimization problem is typically a sum of a (possibly) very large number of terms, each term being the loss function evaluated using one data example. This objective function can also be viewed as an expected loss, in which case it cannot be accurately computed, but can only be evaluated approximately, given a subset of data samples. During the last decade significant theoretical and algorithmic advances were developed for convex optimization problems, such as logistic regression and support vector machines. However, with the recent practical success of deep neural networks and other nonlinear, nonconvex ML models, the focus has shifted to the analysis and development of methods for nonconvex optimization problems. While SGD remains the method of choice in the nonconvex setting for ML applications, theoretical results are weaker than those in the convex case. In particular, little has been achieved in terms of convergence rates. A notable paper [14] is the first to provide convergence rates guarantee of a sort for a randomized stochastic gradient method in nonconvex setting. This method, however, utilizes a carefully chosen step size and a randomized stopping scheme, which are quite different from what is used in practice.

As an alternative to the basic SGD several *variance reducing* stochastic methods have been proposed recently, such as SAGA [12], SVRG [18] and SARAH [21]. They enjoy much stronger convergence rates than SGD, and have been extended to nonconvex problems [25, 22]. However, these methods specifically exploit the structure of the ML problems, where the objective function is a sum over a deterministic (if large) set of data. SVRG requires the full gradient of the objective function to be computed on some (but not all) of the iterations. Hence, SVRG, essentially is a hybrid between SGD and the full gradient method applied to a finite sum. The overall convergence rate per number of data accesses for SVRG seems better than those for the other two methods. From the practical perspective, however, SGD has low per-iteration complexity and high number of iterations and thus is not effective in a distributed setting, while each iteration of a full gradient method can be efficiently distributed, reducing the overall wall-clock time. SVRG (as well as SARAH) being a hybrid does not easily fit with either setting because it alternates between cheap stochastic gradient computation, which have to be sequential and expensive full gradient computations, which can be distributed. In other words, the superior theoretical computational complexity of the SVRG does not necessarily reflect its practical performance. Moreover, the assumption that the data set is fixed (deterministic) contradicts the ultimate goal of learning, which is to obtain a solution with good generalization performance. The method we describe in this paper is applicable in the purely stochastic setting (without assuming that there is a fixed finite set of data) and as our theory shows, it relies on variance reduction that can simply be achieved by choosing adaptive sample sizes that tend to grow as the algorithm progresses to optimality. Such adaptive schemes have been proposed in the literature primarily for gradient descent methods and in a convex setting [5, 13].

With the rise of interest in nonconvex optimization, the ML community started to consider a classical alternative to gradient descent/line search methods - trust region methods [9, 17, 11].

Their usefulness is largely dictated by their ability to utilize negative curvature in Hessian approximations and hence, potentially, escape the neighborhoods of saddle points [11], which can significantly slow down or even trap a line search method. It is argued, that while saddle points are undesirable, the local minima are typically sufficient for the purposes of training nonconvex ML models, such as deep neural networks. There has been a number of recent works that propose trust-region methods that use stochastic gradient and Hessian estimates [16, 29], but they all assume that the objective function is deterministic. One of the original stochastic trust region methods for this stochastic optimization setting has been proposed in [7] and a more sophisticated adaptive method has been recently introduced in [26]. For both methods the convergence is achieved by repeatedly sampling the function values (and gradients, when applicable) so that eventually the estimates become asymptotically error-free with probability 1. No convergence rates have been derived for these algorithms, because the progress happens in the asymptotics. Fully stochastic versions of trust region methods with adaptive sampling, such as may be used in ML context, have not yet been explored to our knowledge. In addition, our analysis applies to the setting where the objective functions and gradient estimates may be biased.

The method, which we analyze in this paper and which we will refer to as STORM (Stochastic trust region method with Random Models), has been introduced in [8], where the authors prove the almost-sure convergence to a first order stationary point. It is a stochastic, variance reducing trust region method which is essentially a minor modification of a classical trust region framework. A very similar method has been analyzed in [19], under more restrictive conditions on  $\tilde{f}(x, \xi)$ . We believe that our convergence rate analysis framework can be easily applied to that method as well, but we choose to focus on STORM.

STORM uses adaptive trust region radii and is close to what is known to be efficient in practice, hence here we focus on the theoretical analysis of this method in the first and second order settings. We recover the convergence rates whose dependence on  $\epsilon$  is the same as of those for deterministic trust region method. Since the method is stochastic our convergence rates are derived in the form of the bound on the expected number of iterations the algorithms takes until achieving the  $\epsilon$ -accuracy. In contrast, convergence rate result for SGD in [14], for example, only bounds the expected sum of the norms of all the gradients up to iteration  $T$ , as a function of  $T$ . Other weaker types of convergence rates are established in [29] and [27]. In [29] a trust region and a cubic regularization methods based on sampled Hessian are considered. The number of samples is selected in such a way that the error in the Hessian approximation is smaller than  $\epsilon$  with large probability  $p$ . Then the deterministic convergence rate can be established under the assumption that the condition on the Hessian approximation holds at each iteration until  $\epsilon$ -accuracy is reached. Hence, the established bound on the number of iterations  $T$  holds with probability  $p^T$  and with probability  $1 - p^T$  no bound is known. The same type of complexity result is derived in [27] for a cubic regularization method, where gradients and Hessians are also sampled at a rate dictated by  $\epsilon$  and the resulting bound holds only with some probability.

Algorithms in [29] have some similarities with algorithms analyzed in [6] and [16]. In [6] and [16], the global rates of convergence of a trust region, a line search and an adaptive cubic regularization methods are analyzed under the assumption that first and second order information is inexact, but sufficiently accurate with some probability, however, the analysis in all of these papers relies heavily on the assumption that function values are computed accurately, in particular that no increasing steps are allowed. This implies that the results in [6] and [16] cannot be applied in a stochastic setting. [27], on the other hand, does not explicitly use function values,

because it does not utilize adaptive step sizes. This paper can be seen as an extension of [6] and [16] to the case of stochastic functions.

Unlike what is done in most of the literature on stochastic methods, we do not make the assumption that the function, gradient or Hessian estimates are unbiased. Instead it is assumed that at each iteration, the function values  $f(x)$ , the gradient  $\nabla f(x)$  and possibly the Hessian  $\nabla^2 f(x)$  can be approximated up to sufficient accuracy with a fixed, but sufficiently high probability  $p$ , conditioned on the past. This assumption, which we will make formal later, is very general and does not explicitly specify how such approximations can be obtained. In case when unbiased estimators are available, one can utilize sampling techniques described, for example in [29, 8]. On the other hand, in [8] that are examples of  $\tilde{f}(x, \xi)$  which is a biased estimator of  $f(x)$ , which is arbitrarily erroneous with some small fixed probability, and yet, the required approximations can be constructed and the trust region method converges. Note that while our condition on the approximations hold only with probability  $p$ , we provide complexity result in expectation, thus accounting for "failed" approximations.

The goal of our paper is twofold: First, we introduce a novel framework for bounding expected complexity of a stochastic optimization method. This framework is based on defining a renewal-reward process associated with the algorithm as well as its stopping time, which is the time when the algorithm reaches desired accuracy. Then, under certain assumptions, we derive a bound on the expected stopping time. This framework, in principal, can be used for analysis of convergence rates of a variety of algorithms - for instance it applies to all algorithms in [6] and [16]. In recent work [24] it has been applied to analyze a stochastic line-search method. In this paper, specifically, we use this general framework to derive a bound on the convergence rate of the STORM algorithm defined in [8], by proving that these assumptions are satisfied by this algorithm. In particular, we show that the expected number of iterations required to achieve  $\|\nabla f(x)\| \leq \epsilon$  is bounded by  $O(\epsilon^{-2}/(2p-1))$ , which is an improvement on the result in [14] and a similar one to those in [25, 22], in terms of dependence on  $\epsilon$ , but such that, in principal, it never requires computation of the true gradient. The result is a natural extension of the standard, best-known worst-case complexity of any first order method for nonconvex optimization [20]. In this paper we also make a significant improvement upon the results in [8] by relaxing a very restrictive condition on the size of the steps taken by the algorithm. By applying the general analytic framework again, we also provide a second order complexity analysis. We show that a second order STORM variant takes an expected number of iterations that is at most  $O(\epsilon^{-3}/(2p-1))$  to ensure  $\max\{\|\nabla f(x)\|, -\lambda_{\min}(\nabla^2 f(x))\} \leq \epsilon$ ; this result requires slightly stronger assumptions on the function estimates but provides generalization of results in [29, 16] to the stochastic case.

Our main complexity results does not yet provide a termination criterion that would guarantee that  $\|f(\bar{x})\| \leq \epsilon$ , where  $\bar{x}$  is the last iterate. However, the analysis provides a foundation for establishing such a criterion. In particular, while in this paper we simply bound the expected complexity, bounding the tail of the complexity distribution will follow from the analysis here.

In the next section we present and analyze our generic framework, while the STORM algorithm is analyzed in Section 3.

The rest of the paper is organized as follows: we begin by introducing the stochastic framework and deriving the bound on its expected stopping time in Section 2. In Section 3 we provide the first order complexity analysis of the STORM algorithm by showing that it fits into the framework introduced in Section 2. The second order complexity analysis follows in Section 4.

**Notation** Throughout the paper we use  $\|\cdot\|$  to denote the Euclidian norm. Some of the bounding constants that are used are denoted by  $\kappa$  with a subscript that is meant to indicate the objects that the constant bounds. In particular we use the following constants.

$$\begin{aligned} \kappa_{ef} & \text{ “error in the function value”,} \\ \kappa_{eg} & \text{ “error in the gradient”,} \\ \kappa_{fcd} & \text{ “fraction of Cauchy decrease”,} \\ \kappa_{bhm} & \text{ “bound on the Hessian of the models”,} \end{aligned}$$

We will also use  $I(A)$  to denote the indicator of a random event  $A$  occurring.

## 2 A Renewal-Reward Martingale Process

In this section we consider a general random process and a stopping time  $T$ , which possesses certain properties. We analyze the behavior of this random process and derive a bound on the expected stopping time. These results will be used later in the paper in the specific setting of convergence of a stochastic trust region method to first order stationary points. We argue that the framework presented in this section can be used for convergence analysis of a variety of stochastic algorithms. We start by defining a stopping time of a discrete time stochastic process.

**Definition 2.1.** *Given a stochastic process  $\{X_k\} = \{X_k : k \geq 0\}$ , we say that  $T$  is a stopping time with respect to  $\{X_k\}$  if for each  $m \geq 0$  the occurrence of the event  $\{T = m\}$  is determined by observing  $X_1, \dots, X_m$ . That is,  $\{T = m\} \in \sigma(X_0, \dots, X_m)$ , the  $\sigma$ -field generated by  $X_1, \dots, X_m$ , for each  $m \geq 0$ .*

Now, let  $\{(\Phi_k, \Delta_k)\}$  be a random process such that  $\Phi_k \in [0, \infty)$  and  $\Delta_k \in [0, \infty)$  for  $k \geq 0$ . Let  $V_{k+1} = \Phi_{k+1} - \Phi_k$  for  $k \geq 0$ . We also assume the existence of a sequence  $\{W_k\}_{k=1}^{\infty}$ , defined on the same probability space as  $\{(\Phi_k, \Delta_k)\}$ , we introduce  $W_0 = 1$  and let  $\mathcal{F}_k$  denote the  $\sigma$ -algebra generated by  $\{(\Phi_0, \Delta_0, W_0), \dots, (\Phi_k, \Delta_k, W_k)\}$ . We assume that  $\{W_k\}_{k=1}^{\infty}$  satisfies

$$\begin{aligned} P(W_{k+1} = 1 | \mathcal{F}_k) &= p, \\ P(W_{k+1} = -1 | \mathcal{F}_k) &= 1 - p. \end{aligned} \tag{2}$$

Note that under the assumption (2) the  $W_k$ 's are independent and also independent of the sequence  $\{(\Phi_k, \Delta_k)\}$ .

Let  $\{T_\epsilon\}_{\epsilon > 0}$  be a family of stopping times with respect to  $\{\mathcal{F}_k\}_{k \geq 0}$ , parametrized by some quantity  $\epsilon > 0$ . The following assumptions will be imposed on  $\{(\Phi_k, \Delta_k)\}$  and  $T_\epsilon$ .

**Assumption 2.1.**

- (i) *There exists constants  $\lambda \in (0, \infty)$  and  $\Delta_{\max} = \Delta_0 e^{\lambda j_{\max}}$  (for some  $j_{\max} \in \mathbb{Z}$ ) such that  $\Delta_k \leq \Delta_{\max}$  for all  $k$ .*
- (ii) *There exists a constant  $\Delta_\epsilon = \Delta_0 e^{\lambda j_\epsilon}$  (for some  $j_\epsilon \in \mathbb{Z}$ ,  $j_\epsilon \leq 0$ ) such that, the following holds for each  $k \geq 0$ ,*

$$I(T_\epsilon > k) \Delta_{k+1} \geq I(T_\epsilon > k) \min(\Delta_k e^{\lambda W_{k+1}}, \Delta_\epsilon), \tag{3}$$

where  $W_{k+1}$  satisfies (2) with  $p > \frac{1}{2}$ .

(iii) There exists a nondecreasing function  $h(\cdot) : [0, \infty) \rightarrow (0, \infty)$  and a constant  $\Theta > 0$  such that

$$\mathbb{E}(V_{k+1} | \mathcal{F}_k) I(T_\epsilon > k) \leq -\Theta h(\Delta_k) I(T_\epsilon > k) \quad (4)$$

or, equivalently,

$$\mathbb{E}(\Phi_{k+1} | \mathcal{F}_k) I(T_\epsilon > k) \leq \Phi_k I(T_\epsilon > k) - \Theta h(\Delta_k) I(T_\epsilon > k). \quad (5)$$

In summary, Assumption 2.1 states that the nonnegative stochastic process  $\Phi_k$  gets reduced by at least  $\Theta h(\Delta_k)$  at each step, as long as  $T_\epsilon > k$ . Also,  $\Delta_k$  tends to increase whenever it is smaller than some threshold  $\Delta_\epsilon$ . Our goal is to bound  $\mathbb{E}(T_\epsilon)$  in terms of  $h(\Delta_\epsilon)$ . What we will show in this section is that, on average,  $\Delta_k \geq \Delta_\epsilon$  frequently, and hence,  $\mathbb{E}(\Phi_{k+1} - \Phi_k)$  can be bounded by a negative fixed value (dependent on  $\epsilon$ ), sufficiently frequently, which will allow us to apply Wald's identity (stated below) and derive the bound on  $\mathbb{E}(T_\epsilon)$ . In order to formalize this, we introduce a renewal process, where renewals occur at times when  $\Delta_k \geq \Delta_\epsilon$  and we consider the sum of  $V_j$ 's obtained between two renewals.

In order to define this renewal process we first introduce an auxiliary process. Define  $\{Z_k\}_{k=0}^\infty$  as follows. First, let  $Z_0 = j_\epsilon$  and set

$$Z_{k+1} = \min(Z_k + W_{k+1}, j_\epsilon).$$

Note that the process  $\{Z_k\}_{k=0}^\infty$  is a birth-death process on the set  $\{k : k \leq j_\epsilon\}$ . Then, define the renewal process,  $A_0 = 0$  and  $A_n = \inf\{m > A_{n-1} : Z_m = j_\epsilon\}$ . By Assumption (3) and using a simple inductive argument for the second inequality below we have that

$$I(T_\epsilon > k) \Delta_{k+1} \geq I(T_\epsilon > k) \min(\Delta_k e^{\lambda W_{k+1}}, \Delta_\epsilon) \geq I(T_\epsilon > k) \Delta_0 \exp(\lambda Z_{k+1}).$$

In other words, on  $T_\epsilon > k$ , the process  $A_n$  only counts the iterations for which  $\Delta_k$  has value at least  $\Delta_\epsilon$ . The interarrival times of this renewal process are defined for all  $k \geq 1$  by

$$\tau_n = A_n - A_{n-1},$$

As a final piece of notation, we define the counting process

$$N(k) = \max\{n : A_n \leq k\},$$

which is the number of renewals that occur before time  $k$ .

First, we have a lemma which relies on the simple structure of the process  $\{W_k\}$  to bound  $\mathbb{E}[\tau_n]$ .

**Lemma 2.1.** *Let  $\tau_n$  be defined as above. Then, for all  $n$*

$$\mathbb{E}[\tau_n] = p + \left(1 + \frac{1}{2p-1}\right) (1-p) = p/(2p-1).$$

*Proof.* Define the process  $\bar{Z}_{k+1} = \bar{Z}_k + W_{k+1}$ , which is a simple random walk. Suppose that  $\bar{Z}_0 = -1$  and define  $\bar{\tau} = \inf\{n \geq 0 : \bar{Z}_n = 0\}$  it is well known (in fact, this follows by Wald's identity) that

$$\mathbb{E}(\bar{\tau}) = \frac{1}{2p-1}.$$

On the other hand, by first step analysis (i.e. conditioning on  $W_1$ ) we have that

$$\mathbb{E}[\tau_1] = 1 \cdot P(W_1 = 1) + (1 + \mathbb{E}[\bar{\tau}])P(W_1 = -1).$$

The above identity follows because the distribution of  $\tau_1$  conditioned on  $Z_1 = j_\epsilon - 1$  is the same as the distribution of  $\bar{\tau}$ . So, we conclude that

$$\mathbb{E}[\tau_1] = p + \left(1 + \frac{1}{2p-1}\right)(1-p),$$

the last equality follows by simplifying the expression above.  $\square$

We now bound the number of renewals that can occur before the time  $T_\epsilon$ .

**Lemma 2.2.**

$$\mathbb{E}(N(T_\epsilon - 1) + 1) \leq \frac{\Phi_0}{\Theta h(\Delta_\epsilon)}.$$

*Proof.* For ease of notation, let  $k \wedge T_\epsilon = \min\{k, T_\epsilon\}$ . Consider the stochastic process defined via  $R_0 = \Phi_0$  and

$$R_k = \Phi_{k \wedge T_\epsilon} + \Theta \sum_{j=0}^{(k \wedge T_\epsilon) - 1} h(\Delta_j),$$

for  $k \geq 1$ , where  $\Theta$  is defined in (5). Observe that  $R_k$  is a non-negative supermartingale with respect to  $\{\mathcal{F}_k\}$ , to see this we first write

$$\mathbb{E}[R_{k+1} | \mathcal{F}_k] = \mathbb{E}[R_{k+1} I(T_\epsilon > k) | \mathcal{F}_k] + \mathbb{E}[R_{k+1} I(T_\epsilon \leq k) | \mathcal{F}_k].$$

Then,

$$\begin{aligned} \mathbb{E}[R_{k+1} I(T_\epsilon \leq k) | \mathcal{F}_k] &= \mathbb{E}[R_{k+1} I(T_\epsilon \leq k) | \mathcal{F}_k] \\ &= \mathbb{E} \left[ \left( \Phi_{T_\epsilon} + \Theta \sum_{j=0}^{T_\epsilon - 1} h(\Delta_j) \right) I(T_\epsilon \leq k) | \mathcal{F}_k \right] \\ &= \Phi_{T_\epsilon} I(T_\epsilon \leq k) + \Theta \sum_{j=0}^{T_\epsilon - 1} h(\Delta_j) I(T_\epsilon \leq k), \end{aligned} \quad (6)$$

where the last equality follows because  $T_\epsilon$  is a stopping time and therefore the random variable the expectation is  $\mathcal{F}_k$ -measurable.

On the other hand, since  $\{T_\epsilon \geq k+1\} = \{T_\epsilon > k\} = \{T_\epsilon \leq k\}^c \in \mathcal{F}_k$  we conclude, using (5), that

$$\begin{aligned} \mathbb{E}[R_{k+1} I(T_\epsilon > k) | \mathcal{F}_k] &= \mathbb{E}[R_{k+1} | \mathcal{F}_k] I(T_\epsilon > k) \\ &= \mathbb{E}[\Phi_{k+1} | \mathcal{F}_k] I(T_\epsilon > k) + \mathbb{E} \left[ \Theta \sum_{j=0}^k h(\Delta_j) | \mathcal{F}_k \right] I(T_\epsilon > k) \\ &\leq \left( \Phi_k - \Theta h(\Delta_k) + \Theta \sum_{j=0}^k h(\Delta_j) \right) I(T_\epsilon > k) \\ &= \left( \Phi_k + \Theta \sum_{j=0}^{k-1} h(\Delta_j) \right) I(T_\epsilon > k). \end{aligned} \quad (7)$$

Combining (6) and (7) we conclude that

$$\mathbb{E}[R_{k+1}|\mathcal{F}_k] \leq R_k$$

as claimed. We then conclude, since  $\Phi_k \geq 0$  for each  $k \geq 0$ , that

$$\Theta \mathbb{E} \left( \sum_{j=0}^{(k \wedge T_\epsilon)-1} h(\Delta_j) \right) = \mathbb{E}[R_k] \leq \mathbb{E}[R_0] = \Phi_0.$$

Now, since  $h(\cdot) \geq 0$ , observe that

$$0 \leq \sum_{j=0}^{(k \wedge T_\epsilon)-1} h(\Delta_j) \nearrow \sum_{j=0}^{T_\epsilon-1} h(\Delta_j)$$

as  $k \rightarrow \infty$ , note that this conclusion holds also on the event  $\{T_\epsilon = \infty\}$ . Therefore, by the Monotone Convergence Theorem

$$\Theta \mathbb{E} \left( \sum_{j=0}^{T_\epsilon-1} h(\Delta_j) \right) = \lim_{k \rightarrow \infty} \Theta \mathbb{E} \left( \sum_{j=0}^{(k \wedge T_\epsilon)-1} h(\Delta_j) \right) \leq \mathbb{E}[R_0] = \Phi_0. \quad (8)$$

Now, by the definition of the counting process  $N(\cdot)$ , since the renewal times  $A_n$  when  $\Delta_{A_n} \geq \Delta_\epsilon$ , are a subset of the iterations  $0, 1, \dots, T_\epsilon$ , and since  $h(\cdot)$  is nondecreasing, we have

$$\Theta \sum_{j=0}^{T_\epsilon-1} h(\Delta_j) \geq \Theta \sum_{j=0}^{T_\epsilon-1} h(\Delta_j) I(j \in \{A_i\}_{i=1}^\infty) = \Theta(N(T_\epsilon - 1) + 1) h(\Delta_\epsilon),$$

the term  $+1$  being added to  $N(T_\epsilon - 1)$  comes from the fact that  $A_0 = 0$ . Inserting this in (8),

$$\mathbb{E}((N(T_\epsilon - 1) + 1)) \leq \frac{\Phi_0}{\Theta h(\Delta_\epsilon)},$$

which concludes the proof.  $\square$

We now state and prove a well known theorem on expected stopping time, known as the Wald's Identity (non-negative increments case) (e.g., see Theorem 2.2.4 in [1]). The reason we provide a proof here, is that in the literature this result is typically shown under the assumption that the stopping time is finite a.s. Dropping this condition is particularly important in our framework, as this condition is equivalent to a convergence result for the optimization algorithm which generates the stochastic process. It is convenient and useful not to have to prove the convergence result before establishing the convergence rates bounds, since the convergence immediately follows from these bounds.

**Theorem 2.1. Wald's Identity (non-negative increments).** *Suppose that  $\{Y_i\}_{i=1}^n$  is a sequence of independent random variables such that  $Y_i \in [0, \infty]$  with probability one. Define  $E(Y_i) = \mu_i \in [0, \infty]$  and let  $N \in [0, \infty]$  be a stopping time with respect to the filtration generated by the  $Y_n$ 's. Define  $S_n = Y_1 + \dots + Y_n$ ,  $S_0 = 0$ ,  $s_n = \mu_1 + \dots + \mu_n$  and  $s_0 = 0$ . Then*

$$E(S_N) = E(s_N).$$

*Proof.* Let  $m > 0$  be an arbitrary integer and define  $Y_i(m) = \min(Y_i, m)$ ,  $N_m = \min(N, m)$ ,  $\mu_i(m) = E(Y_i(m))$ ,  $S_n(m) = Y_1(m) + \dots + Y_n(m)$  and  $s_n(m) = \mu_1(m) + \dots + \mu_n(m)$ . Note that all of these quantities are non-negative and non-decreasing in  $m$ . By the optional sampling theorem applied to the martingale  $M_n = S_n(m) - s_n(m)$ , we have that

$$E(S_{N_m}(m)) = E(\mu_1(m) + \dots + \mu_{N_m}(m)).$$

Because of monotonicity,

$$S_{N_m}(m) \nearrow S_N$$

as  $m \rightarrow \infty$ . For the case  $N = \infty$ , we interpret  $S_N = \sup_{n \geq 0} \sup_m S_n(m)$ . Similarly,

$$s_{N_m}(m) \nearrow s_N,$$

as  $m \rightarrow \infty$ . By the monotone convergence theorem we then conclude that

$$E(S_N) = E(s_N).$$

□

**Remark 2.1.** If  $\mu_i = \mu$ , then  $E(S_N) = \mu \cdot E(N)$ . If  $\mu = 0$ , then  $X_i = 0$  almost surely and  $S_N = 0$ . Therefore, if  $\mu = 0$ , we interpret  $\mu \cdot E(N) = 0$ , even if  $E(N) = \infty$ . This interpretation is consistent with the case in which  $N = 0$  almost surely as well, in this case  $0 = \mu \cdot E(N) = E(S_N)$ , even  $\mu = \infty$ .

We now apply this theorem to  $S_n = A_n = \sum_{i=0}^n \tau_i$  and obtain the main result of this section, which will be used in the following sections to establish the main complexity result.

**Theorem 2.2.** *Let Assumption 2.1 hold. Then*

$$\mathbb{E}[T_\epsilon] \leq \frac{p}{2p-1} \cdot \frac{\Phi_0}{\Theta h(\Delta_\epsilon)} + 1.$$

*Proof.* Define  $\mathcal{G}_n = \mathcal{F}_{A_n}$ , that is,

$$\mathcal{G}_n = \{A \in \sigma(\cup_{m=0}^\infty \mathcal{F}_m) : A \cap \{A_n \leq k\} \in \mathcal{F}_k \text{ for all } k\}.$$

Note that  $A_n$  is a stopping time with respect to  $\{\mathcal{F}_n\}_{n \geq 0}$ , so  $\mathcal{G}_n$  is well defined. We claim that  $N(T_\epsilon - 1) + 1$  is a stopping time with respect to  $\{\mathcal{G}_n\}_{n \geq 0}$ . To see this, note, since  $N(k) \leq k$

$$\begin{aligned} & \{N(T_\epsilon - 1) + 1 \leq n\} \\ &= \cup_{k=0}^{n-1} \{N(k) \leq n-1, T_\epsilon - 1 = k\} \\ &= \cup_{k=0}^{n-1} \{N(k) + 1 \leq n, T_\epsilon = k+1\} \subseteq \mathcal{F}_{A_n}, \end{aligned}$$

where the last inclusion follows because  $N(k) + 1$  is a stopping time with respect to  $\{\mathcal{F}_{A_n}\}_{n \geq 0}$  and because  $A_n \geq n$ , so  $\mathcal{F}_n \subseteq \mathcal{F}_{A_n}$ , which implies that  $T_\epsilon$  is also stopping time with respect to  $\{\mathcal{G}_n\}_{n \geq 0}$ .

Now, because of the independence assumption implied by (2) we have that

$$\mathbb{E}[\tau_{n+1}|\mathcal{G}_n] = \mathbb{E}[\tau_{n+1}] = \frac{p}{2p-1}.$$

Recalling that  $A_{N(T_\epsilon-1)+1} = \sum_{k=1}^{N(T_\epsilon-1)+1} \tau_k$ , we can invoke Wald's identity to conclude that

$$\mathbb{E}[A_{N(T_\epsilon-1)+1}] = \frac{p}{2p-1} \mathbb{E}[N(T_\epsilon-1)+1].$$

Since  $A_{N(T_\epsilon-1)+1} \geq T_\epsilon - 1$ , we have by Lemmas 2.1 and 2.2

$$\mathbb{E}[T_\epsilon - 1] \leq \mathbb{E}[\tau_1] \mathbb{E}[N(T_\epsilon - 1) + 1] \leq \frac{p}{2p-1} \left( \frac{\Phi_0}{\Theta h(\Delta_\epsilon)} \right).$$

The statement of the theorem follows from the last inequality.  $\square$

### 3 The first order STORM algorithm

We now state and analyze a stochastic trust region (TR) algorithm (Algorithm 1) which is essentially very similar to its deterministic counterpart [9]. This method uses the inexact (noisy) information about  $f$  and its derivatives, just as the deterministic method uses the exact information. This algorithm, as stated, and the assumptions on its steps that we will impose below aim at convergence to a first order stationary point. In this section we will analyze the global rate of convergence of this algorithm to such a point (while in Section 4, we extend Algorithm 1 to calculate second order critical points).

---

#### Algorithm 1 STOCHASTIC DFO WITH RANDOM MODELS, [8]

---

- 1: (Initialization): Choose constants  $\gamma > 1$ ,  $\eta_1 \in (0, 1)$ ,  $\eta_2 > 0$ . Choose an initial point  $x^0$  and an initial trust-region radius  $\delta_0 > 0$  and the maximum radius  $\delta_{\max} = \gamma^{j_{\max}} \delta_0$  for some  $j_{\max} \geq 0$ . Set  $k \leftarrow 0$ .
  - 2: (Model construction): Build a (random) model  $m_k(x_k + s) = f_k + g_k^\top s + \frac{1}{2} s^\top H_k s$  that approximates  $f(x)$  in the ball  $B(x_k, \delta_k)$  with  $s = x - x_k$ .
  - 3: (Step calculation) Compute  $s_k = \arg \min_{s: \|s\| \leq \delta_k} m_k(s)$  (approximately) so that it satisfies condition (9).
  - 4: (Estimates calculation) Obtain estimates  $f_k^0$  and  $f_k^s$  of  $f(x_k)$  and  $f(x_k + s_k)$ , respectively.
  - 5: (Acceptance of the trial point): Compute  $\rho_k = \frac{f_k^0 - f_k^s}{m_k(x_k) - m_k(x_k + s_k)}$ .  
If  $\rho_k \geq \eta_1$  and  $\|g_k\| \geq \eta_2 \delta_k$ , set  $x_{k+1} = x_k + s_k$ ; otherwise, set  $x_{k+1} = x_k$ .
  - 6: (Trust-region radius update): If  $\rho_k \geq \eta_1$  and  $\|g_k\| \geq \eta_2 \delta_k$ , set  $\delta_{k+1} = \min\{\gamma \delta_k, \delta_{\max}\}$ ; otherwise  $\delta_{k+1} = \gamma^{-1} \delta_k$ ;  $k \leftarrow k + 1$  and go to step 2.
- 

For every  $k$ , the step  $s_k$  is computed so that the well-known *Cauchy decrease* condition is satisfied,

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{fcd}}{2} \|g_k\| \min \left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\} \quad (9)$$

for some constant  $\kappa_{fcd} \in (0, 1]$ . This condition is standard for the TR methods, easy to enforce in practice and is discussed in detail in the literature [9, 23]. Iterations on which  $x_{k+1} = x_k + s_k$  occurs are called *successful*.

Algorithm 1 generates a random process. The source of randomness are the random models and random estimates constructed on each iteration, based on some random information obtained from the stochastic function  $f(x, \varepsilon)$ .  $M_k$  will denote a random model in the  $k$ -th iteration, while we will use the notation  $m_k = M_k(\omega)$  for its realizations. As a consequence of using random models, the iterates  $X_k$ , the trust-region radii  $\Delta_k$  and the steps  $S_k$  are also random quantities, and so  $x_k = X_k(\omega)$ ,  $\delta_k = \Delta_k(\omega)$ ,  $s_k = S_k(\omega)$  will denote their respective realizations. Similarly, let random quantities  $\{F_k^0, F_k^s\}$  denote the estimates of  $f(X_k)$  and  $f(X_k + S_k)$ , with their realizations denoted by  $f_k^0 = F_k^0(\omega)$  and  $f_k^s = F_k^s(\omega)$ . In other words, Algorithm 1 results in a stochastic process  $\{M_k, X_k, S_k, \Delta_k, F_k^0, F_k^s\}$ . Our goal is to show that under certain conditions on the sequences  $\{M_k\}$  and  $\{F_k\} = \{(F_k^0, F_k^s)\}$  the resulting stochastic process has desirable convergence rate. In particular, we will assume that models  $M_k$  and estimates  $F_k^0, F_k^s$  are sufficiently accurate with sufficiently high probability, conditioned on the past.

The key to the analysis lies in the assumption that the accuracy improves in coordination with the perceived progress of the algorithm. The main challenge of the analysis lies in the fact that, while in the deterministic case the function  $f(x)$  never increases from one iteration to another, this can easily happen in the stochastic case. The analysis is based on properties of supermartingales where the increments of a supermartingale depend on the function change between iterates (which as we will show, *tend* to decrease). To make the analysis simpler we need a technical assumption that these increments are bounded from above. Hence, overall we make the following assumptions on  $f$ :

**Assumption 3.1.** *We assume that all iterates  $x_k$  generated by Algorithm 1 the gradient  $\nabla f$  is  $L$ -Lipschitz continuous and*

$$f(x) \geq 0$$

The assumptions of Lipschitz continuity of  $\nabla f$  and boundedness of  $f$  from below are standard. Here for simplicity and w.l.o.g. we assume that the lower bound on  $f$  is nonnegative.

### 3.1 Assumptions on the first order STORM algorithm

Let  $\mathcal{F}_{k-1}^{M \cdot F}$  denote the  $\sigma$ -algebra generated by  $M_0, \dots, M_{k-1}$  and  $F_0, \dots, F_{k-1}$  and let  $\mathcal{F}_{k-1/2}^{M \cdot F}$  denote the  $\sigma$ -algebra generated by  $M_0, \dots, M_k$  and  $F_0, \dots, F_{k-1}$ .

**Definition 3.1.** *1) A function  $m_k$  is a  $\kappa$ -fully linear model of  $f$  on  $B(x_k, \delta_k)$  provided, for  $\kappa = (\kappa_{ef}, \kappa_{eg})$  and  $\forall y \in B(x_k, \delta_k)$ ,*

$$\begin{aligned} \|\nabla f(x_k) - g_k\| &\leq \kappa_{eg} \delta_k, \\ |f(y) - m_k(y)| &\leq \kappa_{ef} \delta_k^2. \end{aligned} \tag{10}$$

*2) The estimates  $f_k^0$  and  $f_k^s$  are said to be  $\epsilon_F$ -accurate estimates of  $f(x_k)$  and  $f(x_k + s_k)$ , respectively, for a given  $\delta_k$  if*

$$|f_k^0 - f(x_k)| \leq \epsilon_F \delta_k^2 \text{ and } |f_k^s - f(x_k + s_k)| \leq \epsilon_F \delta_k^2. \tag{11}$$

**Definition 3.2.** A sequence of random models  $\{M_k\}$  is said to be  $\alpha$ -probabilistically  $\kappa$ -fully linear with respect to the corresponding sequence  $\{B(X_k, \Delta_k)\}$  if the events

$$I_k = \mathbb{1}\{M_k \text{ is a } \kappa\text{-fully linear model of } f \text{ on } B(X_k, \Delta_k)\} \quad (12)$$

satisfy the condition

$$P(I_k = 1 | \mathcal{F}_{k-1}^{M \cdot F}) \geq \alpha.$$

**Definition 3.3.** A sequence of random estimates  $\{F_k\}$  is said to be  $\beta$ -probabilistically  $\epsilon_F$ -accurate with respect to the corresponding sequence  $\{X_k, \Delta_k, S_k\}$  if the events

$$J_k = \mathbb{1}\{F_k^0, F_k^s \text{ are } \epsilon_F\text{-accurate estimates of } f(x_k) \text{ and } f(x_k + s_k), \text{ respectively, for } \Delta_k\} \quad (13)$$

satisfy the condition

$$P(J_k = 1 | \mathcal{F}_{k-1/2}^{M \cdot F}) \geq \beta,$$

where  $\epsilon_F$  is a fixed constant.

Next is the key assumption on the nature of the stochastic (and deterministic) information used by our algorithm.

**Assumption 3.2.** The following hold for the quantities used in Algorithm 1

- (a) The model Hessians satisfy  $\|H_k\|_2 \leq \kappa_{bhm}$  for some  $\kappa_{bhm} \geq 1$ , for all  $k$ , deterministically.
- (b) The sequence of random models  $M_k$ , generated by Algorithm 1, is  $\alpha$ -probabilistically  $\kappa$ -fully linear, for some  $\kappa = (\kappa_{ef}, \kappa_{eg})$  and for a sufficiently large  $\alpha \in (0, 1)$ .
- (c) The sequence of random estimates  $\{F_k\}$  generated by Algorithm 1 is  $\beta$ -probabilistically  $\epsilon_F$ -accurate for  $\epsilon_F \leq \kappa_{ef}$  and  $\epsilon_F < \frac{1}{4}\eta_1\eta_2\kappa_{fcd} \min\left\{\frac{\eta_2}{\kappa_{bhm}}, 1\right\}$ , and for a sufficiently large  $\beta \in (0, 1)$ .

**Remark 3.1.** In [8] the analysis of the algorithm requires an additional assumption that  $\eta_2 \geq \kappa_{ef}$  and for simplicity it is further assumed that  $\eta_2 \geq \kappa_{bhm}$ . This assumption is undesirable since it restricts the size of the steps that can be taken by the trust region algorithm. In this paper we manage to improve the analysis and drop this assumption, hence allowing  $\eta_2$  to be set to a small value. Note that small values of  $\eta_2$  imply small  $\epsilon_F$  because of Assumption 3.2(c), so there is a potential trade-off in choosing  $\eta_2$ . On the other hand, this relationship indicates that when  $\epsilon_F = 0$ , that is when there is no error in the function estimates, then  $\eta_2$  can be arbitrarily small.

Under Assumption 3.2,  $P\{I_k J_k = 1 | \mathcal{F}_{k-1}^{M \cdot F}\} \geq \alpha\beta$  and  $P\{I_k + J_k = 0 | \mathcal{F}_{k-1}^{M \cdot F}\} \leq (1 - \alpha)(1 - \beta)$ . At iteration  $k$ , if  $I_k J_k = 1$  then the behavior of the algorithm reduces to that of an (inexact) deterministic algorithm; while if  $I_k + J_k = 0$ , then not only may the algorithm produce a bad step (that is a step which increases the objective function), but it also may accept this bad step by mistaking it for an improving step (that is a step that decreases the function value). In the cases when only one of  $I_k = 0$  and  $J_k = 0$  holds, then either the model is good but the estimates are faulty, or the estimates are good and the model is faulty. In this case an improving step is still possible, but a bad step is not. In the worst case, no step is taken and the trust region radius is reduced. The main idea of our framework is to choose probabilities of  $I_k J_k = 1$  and  $I_k + J_k = 0$  occurring according to the possible corresponding decrease and increase in  $f(x)$ , so that in expectation,  $f(x)$  is sufficiently reduced.

### 3.2 Useful existing results

Algorithm 1 is analyzed in [8] and the following almost-sure stationarity result is shown: there exists a selection of  $\alpha$  and  $\beta$  such that under Assumption 3.2 with additional requirement that  $\eta_2 \geq \kappa_{ef}$ , the sequence of random iterates generated by Algorithm 1,  $\{X_k\}$ , almost surely satisfies  $\lim_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0$ . The important observation is that  $\alpha$  and  $\beta$  do not have to increase as the algorithm progresses. Hence with the same, constant but small enough, probabilities our models and estimates can be arbitrarily erroneous.

Our primary goal in this paper is to bound the expected number of steps that the algorithm takes until  $\|\nabla f(X_k)\| \leq \epsilon$  occurs and the secondary goal is to relax the assumption  $\eta_2 \geq \kappa_{ef}$ . We will modify the analysis that led to the above stationarity result in [8]. First, we state (without proof) several auxiliary lemmas from [8].

**Lemma 3.1.** [Good model  $\Rightarrow$  function reduction in  $\|g_k\|$ ] *Suppose that a model  $m_k$  is a  $(\kappa_{ef}, \kappa_{eg})$ -fully linear model of  $f$  on  $B(x_k, \delta_k)$ . If*

$$\delta_k \leq \min \left\{ \frac{1}{\kappa_{bhm}}, \frac{\kappa_{fcd}}{8\kappa_{ef}} \right\} \|g_k\|,$$

*then the trial step  $s_k$  leads to an improvement in  $f(x_k + s_k)$  such that*

$$f(x_k + s_k) - f(x_k) \leq -\frac{\kappa_{fcd}}{4} \|g_k\| \delta_k. \quad (14)$$

**Lemma 3.2.** [Good model  $\Rightarrow$  function reduction in  $\|\nabla f(x_k)\|$ ] *Under Assumption 3.2(a), suppose that a model is  $(\kappa_{ef}, \kappa_{eg})$ -fully linear on  $B(x_k, \delta_k)$ . If*

$$\delta_k \leq \min \left\{ \frac{1}{\kappa_{bhm} + \kappa_{eg}}, \frac{1}{\frac{8\kappa_{ef}}{\kappa_{fcd}} + \kappa_{eg}} \right\} \|\nabla f(x_k)\|, \quad (15)$$

*then the trial step  $s_k$  leads to an improvement in  $f(x_k + s_k)$  such that*

$$f(x_k + s_k) - f(x_k) \leq -C_1 \|\nabla f(x_k)\| \delta_k, \quad (16)$$

*for any  $C_1 \leq \frac{\kappa_{fcd}}{4} \cdot \max \left\{ \frac{\kappa_{bhm}}{\kappa_{bhm} + \kappa_{eg}}, \frac{8\kappa_{ef}}{8\kappa_{ef} + \kappa_{fcd}\kappa_{eg}} \right\}$ .*

**Lemma 3.3.** [Good model + good estimates  $\Rightarrow$  successful step] *Under Assumption 3.2(a), suppose that  $m_k$  is  $(\kappa_{ef}, \kappa_{eg})$ -fully linear on  $B(x_k, \delta_k)$  and the estimates  $\{f_k^0, f_k^s\}$  are  $\epsilon_F$ -accurate with  $\epsilon_F \leq \kappa_{ef}$ . If*

$$\delta_k \leq \min \left\{ \frac{1}{\kappa_{bhm}}, \frac{1}{\eta_2}, \frac{\kappa_{fcd}(1 - \eta_1)}{8\kappa_{ef}} \right\} \|g_k\|, \quad (17)$$

*then the  $k$ -th iteration is successful.*

**Lemma 3.4.** [Good estimates + successful step  $\Rightarrow$  function reduction in  $\delta_k^2$ ] *Under Assumption 3.2(a), suppose that the estimates  $\{f_k^0, f_k^s\}$  are  $\epsilon_F$ -accurate with  $\epsilon_F < \frac{1}{4}\eta_1\eta_2\kappa_{fcd} \min \left\{ \frac{\eta_2}{\kappa_{bhm}}, 1 \right\}$ . If a trial step  $s_k$  is accepted (a successful iteration occurs), then the improvement in  $f$  is bounded below as follows*

$$f(x_{k+1}) - f(x_k) \leq -C_2 \delta_k^2, \quad (18)$$

*where*

$$C_2 = \frac{1}{2}\eta_1\eta_2\kappa_{fcd} \min \left\{ \frac{\eta_2}{\kappa_{bhm}}, 1 \right\} - 2\epsilon_F > 0. \quad (19)$$

**Choosing constants** We now explain briefly the role of the constants  $\eta_2, \epsilon_F, \alpha$ , and  $\beta$  and their expected magnitude. First, we note that constants  $\kappa_{ef}, \kappa_{eg}$ , and  $\kappa_{bhm}$  can be chosen arbitrarily large, but ideally should be chosen as small as possible while guaranteeing Assumption 3.2. Let us assume that  $\kappa_{ef}, \kappa_{eg}$  and  $\kappa_{bhm}$  can be all chosen as  $\Theta(L)^1$ , where  $L$  is the Lipschitz constant of  $\nabla f(x)$  in  $\mathcal{X}$ , even though it may not be explicitly known (see [10] for construction of fully-linear models in the case of unavailable derivative estimates). Once these constants are chosen,  $\epsilon_F$  is chosen so that it satisfies the conditions in Assumption 3.2(c). Note that if  $\eta_2$  is chosen to equal  $L$ , this means that Algorithm 1 only takes steps when  $\delta_k \leq \frac{\|g_k\|}{L}$ , which is similar to constraining step size by  $\frac{1}{L}$  in gradient descent. In this case,  $\epsilon_F$  can be chosen relatively large and thus the estimates need to be slightly more accurate than the models, but the order of required accuracy is similar. On the other hand, since in trust region methods, step sizes are meant to be chosen adaptively, it is desirable to allow larger steps, which can be done by setting  $\eta_2$  to be small. This, in turn, implies that  $\epsilon_F$  has to be chosen small and hence the estimates will have to be a lot more accurate than the models. Another trade-off when choosing a small value for  $\eta_2$  will become apparent in our main complexity results, as we will see that the expected improvement per iteration may depend on  $\eta_2$ . But reasonable values for  $\eta_2$  allow the removal of this dependency.

To simplify expressions for various constants we will assume that  $\eta_1 = 0.1, \gamma = 2$  and  $\kappa_{fcd} = 0.5$  which are typical values for these constants. We also assume, w.l.o.g., that  $\kappa_{bhm} \leq 12\kappa_{ef}$  and  $\eta_2 \leq \kappa_{eg}$ . To simplify expressions further we will consider  $\kappa_{ef} = \kappa_{eg}$ . It is clear that if  $\kappa_{ef}$  or  $\kappa_{eg}$  happen to be smaller, somewhat better bounds than the ones we derive here will result, because the models give tighter approximations of the true function. We are interested in deriving bounds for the case when  $\kappa_{ef}$  or  $\kappa_{eg}$  may be large. The analysis can be performed for any other values of the above constants, hence the choice here is done merely for convenience and simplicity.

The conditions on  $\alpha$  and  $\beta$  under the above choice of constants will be shown in our results below.

### 3.3 Defining and analyzing the process $\{\Phi_k, \Delta_k\}$ .

We consider a random process  $\{\Phi_k, \Delta_k\}$  derived from the process generated by Algorithm 1, with  $\Delta_k$  - the trust region radius and

$$\Phi_k = \nu f(X_k) + (1 - \nu)\Delta_k^2 \quad (20)$$

where  $\nu \in (0, 1)$  is a deterministic, large enough constant, which will be defined later. Clearly  $\Phi_k \geq 0$ . Recall the notation  $\phi_k$  for realizations of  $\Phi_k$ . Here we defined  $\mathcal{F}_k$  as  $\mathcal{F}_k^{M \cdot F}$ .

Define a random time

$$T_\epsilon = \inf\{k \geq 0 : \|\nabla f(X_k)\| \leq \epsilon\}. \quad (21)$$

It is easy to see that  $T_\epsilon$  is a stopping time for the stochastic process defined by Algorithm 1 and hence for  $\{\Phi_k, \Delta_k\}$ .

As stated our goal is to bound the expected stopping time  $\mathbb{E}(T_\epsilon)$ . We will do so by showing that Assumption 2.1 is satisfied for  $\{\Phi_k, \Delta_k\}$  and applying results of Section 2.

Let us show that Assumptions 2.1(i)-(ii) hold with the following  $\Delta_\epsilon$

---

<sup>1</sup>Note that it is possible to have  $\kappa_{ef}$  and  $\kappa_{eg}$  of different magnitudes, namely when  $\kappa_{eg}$  is small as we have sufficiently accurate gradients, but  $\kappa_{ef}$  remains as  $\Theta(L)$ . Our analysis and results apply then as well.

$$\Delta_\epsilon = \frac{\epsilon}{\zeta}, \quad \text{for } \zeta \geq \kappa_{eg} + \max\left\{\eta_2, \kappa_{bhm}, \frac{8\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)}\right\}. \quad (22)$$

Note that with our choice of algorithmic parameters the above is satisfied by  $\zeta = 20\kappa_{eg}$ .

For simplicity of the presentation and without loss of generality, we assume that  $\Delta_\epsilon = \gamma^i \delta_0$ , for some integer  $i \leq 0$ . If not, we can always choose  $\zeta$  within a factor of  $\gamma$  of its lower bound in (22). It follows that for any  $k$ ,  $\Delta_k = \gamma^{i_k} \Delta_\epsilon$ , for some integer  $i_k$ . Choosing  $\lambda$  in Assumption 2.1(i)-(ii) so that  $e^\lambda = \gamma$  Assumption 2.1(i) follows immediately from the definition of  $\{\Phi_k, \Delta_k\}$ , and the choice of  $\delta_{max}$  imposed by Algorithm 1 and for Assumption 2.1(ii) we now only need to show that the dynamics (3) hold for  $\Delta_k$ .

**Lemma 3.5.** *Let Assumptions 3.1 and 3.2 hold. Let  $\alpha$  and  $\beta$  be such that  $\alpha\beta \geq 1/2$ , then Assumption 2.1(ii) is satisfied for  $W_k = 2(I_k J_k - \frac{1}{2})$ ,  $\lambda = \log(\gamma)$  and  $p = \alpha\beta$ .*

*Proof.* Clearly inequality (3) holds when  $I(T_\epsilon > k) = 0$ . We will show that conditioned on  $T_\epsilon > k$  (i.e.  $I(T_\epsilon > k) = 1$ ) we have

$$\Delta_{k+1} \geq \min\{\Delta_\epsilon, \min\{\Delta_{max}, \gamma\Delta_k\}I_k J_k + \gamma^{-1}\Delta_k(1 - I_k J_k)\}. \quad (23)$$

First we note that for each realization when  $\delta_k > \Delta_\epsilon$ , we have  $\delta_k \geq \gamma\Delta_\epsilon$  and hence  $\delta_{k+1} \geq \Delta_\epsilon$ . Now, assume that  $\delta_k \leq \Delta_\epsilon$ , then, because  $T_\epsilon > k$ , we have  $\|\nabla f(x_k)\| > \epsilon$  and hence, from the definition of  $\zeta$ , we know that

$$\|\nabla f(x_k)\| \geq \left( \kappa_{eg} + \max\left\{\eta_2, \kappa_{bhm}, \frac{8\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)}\right\} \right) \delta_k.$$

Assume that  $I_k = 1$  and  $J_k = 1$ , i.e., both the model and the estimates are good on iteration  $k$ . Since the model  $m_k$  is  $\kappa$ -fully linear and

$$\|g_k\| \geq \|\nabla f(x_k)\| - \kappa_{eg}\delta_k \geq (\zeta - \kappa_{eg})\delta_k \geq \max\left\{\eta_2, \kappa_{bhm}, \frac{8\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)}\right\} \delta_k,$$

and the estimates  $\{f_k^0, f_k^s\}$  are  $\epsilon_F$ -accurate, with  $\epsilon_F \leq \kappa_{ef}$ , condition (17) in Lemma 3.3 holds. Hence, iteration  $k$  is successful, i.e.  $x_{k+1} = x_k + s_k$  and  $\delta_{k+1} = \max\{\delta_{max}, \gamma\delta_k\}$ . If  $I_k J_k = 0$ , then  $\delta_{k+1} \geq \gamma^{-1}\delta_k$  simply by the dynamics of Algorithm 1.

Finally, observing that  $P\{I_k J_k = 1 | \mathcal{F}_{k-1}^{M.F}\} \geq p = \alpha\beta$  we conclude that (23) implies Assumption 2.1(ii).  $\square$

We now show that Assumption 2.1(iii) holds, which is the key theorem in this section and is similar to Theorem 4.11 in [8], while dropping the restrictive conditions on  $\eta_2$  and simplifying the proof. We will omit the parts of the proof that are identical to those of Theorem 4.11 in [8].

**Theorem 3.1.** *There exist probabilities  $\alpha$  and  $\beta$  such that under Assumptions 3.1 and 3.2 there exists a constant  $\Theta > 0$  such that, conditioned on  $T_\epsilon > k$*

$$1(T_\epsilon > k)\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}] \leq -1(T_\epsilon > k)\Theta\Delta_k^2. \quad (24)$$

Moreover, under the particular choice of constants described in the last section, let  $\alpha$  and  $\beta$  satisfy

$$\frac{(\alpha\beta - \frac{1}{2})}{(1-\alpha)(1-\beta)} \geq 10 + \frac{30L}{40\kappa_{eg}},$$

and

$$\beta \geq \frac{\kappa_{eg} + 0.064L + 4 \cdot 10^{-4}\eta_2}{\kappa_{eg} + 0.064L + 4.5 \cdot 10^{-4}\eta_2}.$$

Then<sup>2</sup>,  $\Theta = \frac{1}{1800} \min \{ \eta_2 \beta, \kappa_{eg}^{-1} \}$ .

*Proof.* Since (24) holds trivially if  $T_\epsilon \leq k$ , we assume henceforth in this proof that  $\nabla f(X_k) > \epsilon$ . We will consider two possible cases:  $\|\nabla f(x_k)\| \geq \zeta \delta_k$  and  $\|\nabla f(x_k)\| < \zeta \delta_k$ . We will show that (24) holds in both cases and hence it holds on every iteration. Let  $\nu \in (0, 1)$  be such that

$$\frac{\nu}{1-\nu} > \max \left\{ \frac{4\gamma^2}{\zeta C_1}, \frac{4\gamma^2}{\eta_1 \eta_2 \kappa_{fcd}}, \frac{\gamma^2}{\kappa_{ef}} \right\}, \quad (25)$$

with  $C_1$  defined as in Lemma 3.2.

Let  $x_k, \delta_k, s_k, g_k$ , and  $\phi_k$  denote realizations of random quantities  $X_k, \Delta_k, S_k, G_k$ , and  $\Phi_k$ , respectively. Let us consider some realization of Algorithm 1. Note that on all successful iterations,  $x_{k+1} = x_k + s_k$  and  $\delta_{k+1} = \min\{\gamma \delta_k, \delta_{max}\}$  with  $\gamma > 1$ , hence

$$\phi_{k+1} - \phi_k \leq \nu(f(x_{k+1}) - f(x_k)) + (1-\nu)(\gamma^2 - 1)\delta_k^2. \quad (26)$$

On all unsuccessful iterations,  $x_{k+1} = x_k$  and  $\delta_{k+1} = \frac{1}{\gamma} \delta_k$ , i.e.

$$\phi_{k+1} - \phi_k = (1-\nu)\left(\frac{1}{\gamma^2} - 1\right)\delta_k^2 \equiv b_1 < 0. \quad (27)$$

**Case 1:**  $\|\nabla f(x_k)\| \geq \zeta \delta_k$  **with  $\zeta$  satisfying (22).** Let  $\alpha$  and  $\beta$  satisfy

$$\frac{(\alpha\beta - \frac{1}{2})}{(1-\alpha)(1-\beta)} \geq \frac{C_3}{C_1}, \quad (28)$$

with  $C_1$  defined in Lemma 3.2 and  $C_3 = 1 + \frac{3L}{2\zeta}$ .

- a.  $I_k = 1$  and  $J_k = 1$ , i.e., both the model and the estimates are good on iteration  $k$ . The proof is almost identical to that in Theorem 4.11 [8], but with small modification due to the different definition of  $\zeta$  because we no longer assume that  $\eta_2 \geq \kappa_{bhm}$ .

By observing that Lemma 3.2 and Lemma 3.3 hold we can derive

$$\phi_{k+1} - \phi_k \leq -\nu C_1 \|\nabla f(x_k)\| \delta_k + (1-\nu)(\gamma^2 - 1)\delta_k^2 \equiv b_2, \quad (29)$$

for  $\nu \in (0, 1)$  satisfying (25).

- b.  $I_k = 1$  and  $J_k = 0$ , i.e., we have a good model and bad estimates on iteration  $k$ . The proof is identical to that in Theorem 4.11 [8] where it is shown that (27) holds.
- c.  $I_k = 0$  and  $J_k = 1$ , i.e., we have a bad model and good estimates on iteration  $k$ . Again (27) holds, as is shown in Theorem 4.11 [8].

---

<sup>2</sup>Note that  $\beta > \frac{1}{2}$  and so  $\Theta = \frac{1}{1800} \kappa_{eg}^{-1}$ , independently of  $\eta_2$ , provided  $\eta_2 \geq 2\kappa_{eg}^{-1}$ ; the latter implies that small values are allowed for  $\eta_2$  as  $\kappa_{eg}$  values of interest are large.

d.  $I_k = 0$  and  $J_k = 0$ , i.e., both the model and the estimates are bad on iteration  $k$ . The proof of Theorem 4.11 [8] applies, where it is shown that

$$\phi_{k+1} - \phi_k \leq \nu C_3 \|\nabla f(x_k)\| \delta_k + (1 - \nu)(\gamma^2 - 1) \delta_k^2 \equiv b_3. \quad (30)$$

holds with  $C_3 = 1 + \frac{3L}{2\zeta}$ .

Next, following the proof of Case 1 of Theorem 4.11 in [8] we combine the four outcomes to obtain that under condition (28), we have

$$\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\|\nabla f(X_k)\| \geq \zeta \Delta_k\}] \leq -\frac{1}{4} C_1 \nu \|\nabla f(X_k)\| \Delta_k \leq -\frac{1}{4} \frac{C_1 \nu}{\zeta} \Delta_k^2.$$

where last inequality is due to  $\|\nabla f(X_k)\| \geq \zeta \Delta_k$ .

We now derive the bounds on the expectation of  $\Phi_{k+1} - \Phi_k$  in the remaining case. The proof of this case is different than that of Case 2 of Theorem 4.11 [8], because of the dropped bound on  $\eta_2$ .

**Case 2:**  $\|\nabla f(x_k)\| < \zeta \delta_k$  **with  $\zeta$  satisfying (22).** First we note that if  $\|g_k\| < \eta_2 \delta_k$ , then we have an unsuccessful step and (27) holds. Hence, we now assume that  $\|g_k\| \geq \eta_2 \delta_k$ . Here we consider only two outcomes, in particular, we will show that when the estimates are good, (27) holds. Otherwise, because  $\|\nabla f(x_k)\| < \zeta \delta_k$ , the increase in  $\phi_k$  can be bounded from above by a multiple of  $\delta_k^2$ . Hence by selecting appropriate value for probability  $\beta$  we will be able to establish the bound on expected decrease in  $\Phi_k$  as in Case 1.

a.  $J_k = 1$ , i.e., the estimates are good on iteration  $k$ , while the model might be good or bad.

The iteration may or may not be successful. On successful iterations, the good estimates ensure reduction in  $f$ , while on unsuccessful iterations,  $\delta_k$  is reduces. Applying the same argument as in the Case 1(c) we establish that (27) always holds.

b.  $J_k = 0$ , i.e., the estimates are bad on iteration  $k$ , while the model might be good or bad.

Here, as in Case 1, we bound the maximum possible increase in  $\phi_k$ . Using the Taylor expansion, the Lipschitz continuity of  $\nabla f(x)$  and taking into account the bound  $\|\nabla f(x_k)\| < \zeta \delta_k$  we have

$$f(x_k + s_k) - f(x_k) \leq \|\nabla f(x_k)\| \delta_k + \frac{1}{2} L \delta_k^2 < C_3 \zeta \delta_k^2.$$

Hence, the change in function  $\phi$  is

$$\phi_{k+1} - \phi_k \leq [\nu C_3 \zeta + (1 - \nu)(\gamma^2 - 1)] \delta_k^2. \quad (31)$$

We are now ready to bound the expectation of  $\phi_{k+1} - \phi_k$  as we did in Case 1 except that now we simply combine (31), which holds with probability at most  $(1 - \beta)$ , and (27), which holds otherwise:

$$\begin{aligned} & \mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\|\nabla f(X_k)\| < \zeta \Delta_k\}] \\ & \leq \beta (1 - \nu) \left(\frac{1}{\gamma^2} - 1\right) \Delta_k^2 \\ & \quad + (1 - \beta) [\nu C_3 \zeta + (1 - \nu)(\gamma^2 - 1)] \Delta_k^2 \end{aligned} \quad (32)$$

If we choose probability  $0 < \beta \leq 1$  so that the following holds,

$$\frac{\beta}{1-\beta} \geq \frac{2\nu\gamma^2 C_3 \zeta}{(1-\nu)(\gamma^2-1)} + 2\gamma^2, \quad (33)$$

then the first term in (32), which is negative, is at least twice as large in absolute value as the second term, which is positive. We thus have

$$\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\|\nabla f(X_k)\| < \zeta \Delta_k\}] \leq \frac{1}{2} \beta (1-\nu) \left(\frac{1}{\gamma^2} - 1\right) \Delta_k^2. \quad (34)$$

To complete the proof of the lemma it remains to substitute the appropriate constants in the above expressions. In particular, because of our assumptions that  $\kappa_{bhm} \leq 12\kappa_{ef}$  and  $\kappa_{ef} = \kappa_{eg}$ , we can choose  $C_1 = \frac{1}{10}$ , and, recalling the choice of  $\zeta = 20\kappa_{eg}$ ,  $\gamma = 2$ ,  $\eta_1 = 0.1$ ,  $\kappa_{fcd} = 0.5$  and  $\eta_2 \leq \kappa_{eg}$ , (25) reduces to

$$\frac{\nu}{1-\nu} \geq \frac{4\gamma^2}{\eta_1 \eta_2 \kappa_{fcd}} \geq \frac{320}{\eta_2}, \quad (35)$$

which holds if

$$\nu \geq \frac{320}{320 + \eta_2}.$$

We can assume that  $\nu > \frac{1}{2}$  without loss of generality.

**Case 1:** For the probabilities  $\alpha$  and  $\beta$  to satisfy (28) with  $C_3 = 1 + \frac{3L}{2\zeta}$ , it is sufficient that

$$\frac{(\alpha\beta - \frac{1}{2})}{(1-\alpha)(1-\beta)} \geq 10 + \frac{30L}{40\kappa_{eg}}.$$

Then, using  $\nu > \frac{1}{2}$  (31) implies

$$\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\|\nabla f(X_k)\| < \zeta \Delta_k\}] \leq -\frac{1}{1600\kappa_{eg}} \Delta_k^2.$$

**Case 2:** Recalling the expression for  $C_3$  and the values for constant  $\nu$ ,  $\zeta$  and  $\gamma = 2$ , and choosing  $\nu$  so that (35) is satisfied with equality, we see that (33) is satisfied if

$$\frac{\beta}{(1-\beta)} \geq \frac{4 \times 320(40\kappa_{eg} + 3L)}{3\eta_2} + 8,$$

which is satisfied if

$$\frac{\beta}{(1-\beta)} \geq \frac{1280(14\kappa_{eg} + L)}{\eta_2} + 8, \quad (36)$$

which in turn is satisfied by

$$\beta \geq \frac{2 \cdot 10^4 \kappa_{eg} + 1280L + 8\eta_2}{2 \cdot 10^4 \kappa_{eg} + 1280L + 9\eta_2} = \frac{\kappa_{eg} + 0.064L + 4 \cdot 10^{-4}\eta_2}{\kappa_{eg} + 0.064L + 4.5 \cdot 10^{-4}\eta_2}.$$

Then, observing that  $\nu$  is chosen so that  $1-\nu = \frac{\eta_2}{320+\eta_2}$ , from (34) and  $\eta_2 < 320$  (as  $\nu > \frac{1}{2}$ ),

$$\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\|\nabla f(X_k)\| < \zeta \Delta_k\}] \leq -\frac{3\eta_2}{8(320 + \eta_2)} \beta \Delta_k^2 \leq -\frac{1}{1800} \eta_2 \beta \Delta_k^2.$$

Hence we conclude that

$$\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}] \leq -\Theta \Delta_k^2$$

for  $\Theta = \frac{1}{1800} \min \{\eta_2 \beta, \kappa_{eg}^{-1}\}$ , which completes the proof.  $\square$

The almost-sure stationarity result follows immediately from Theorem 3.1 with the same proof as in [8], but this time without the assumption  $\eta_2 \geq \kappa_{ef}$ .

**Theorem 3.2.** *Let Assumptions 3.1 and 3.2 hold, and let  $\alpha$  and  $\beta$  satisfy conditions of Theorem 3.1, then the sequence of random iterates generated by Algorithm 1,  $\{X_k\}$ , almost surely satisfies*

$$\lim_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0.$$

The validity of the Assumption 2.1(iii) follows from Theorem 3.1. We state the result below for completeness and convenience of reference.

**Lemma 3.6.** *Let the assumptions of Theorem 3.1 hold. Then Assumption 2.1(iii) is satisfied, with  $\Theta = \frac{1}{1800} \min \{\eta_2 \beta, \kappa_{eg}^{-1}\}$  for the process  $\{\Phi_k, \Delta_k\}$ , where  $\Phi_k$  is defined as in (20) with  $\nu$  satisfying (25) and  $h(\delta) = \delta^2$ .*

### 3.4 Complexity result for first order STORM algorithm

**Theorem 3.3.** *Consider Algorithm 1 and the corresponding stochastic process. Let  $T_\epsilon$  be defined as in (21). Then, under the assumptions of Theorem 3.1,*

$$\mathbb{E}[T_\epsilon] \leq \frac{\alpha\beta}{2\alpha\beta - 1} \left( \frac{20\Phi_0\kappa_{eg}}{\Theta\epsilon^2} + 1 \right),$$

where  $\Theta = \frac{1}{1800} \min \{\eta_2 \beta, \kappa_{eg}^{-1}\}$ ,  $\Phi_0$  defined as in (20) with  $k = 0$ , with  $\nu$  satisfying (25).

### 3.5 Example of models and estimates satisfying Assumption 3.2

While the assumption 3.2, which allows us to develop the general complexity analysis, is fairly general it is easy to satisfy in practice in the classical stochastic optimization setting by taking a sufficient number of samples of the function, gradient and Hessian estimates. A number of recent papers rely on this technique, for producing sufficiently accurate gradient and Hessian approximations. For example Lemma 4 in [27] uses matrix concentration results from [28] to show that given the bound on the variance of the gradient

$$\mathbb{E}[\|\nabla \tilde{f}(x, \xi) - \nabla f(x)\|] \leq \sigma_g^2,$$

the average of  $\tilde{\mathcal{O}}(\frac{\sigma_g^2}{\epsilon^2})$  gradient samples  $\nabla \tilde{f}(x, \xi)$ ,  $g$ , satisfies

$$\|g - \nabla \tilde{f}(x, \xi)\| \leq \epsilon$$

with probability  $p$ , where  $\tilde{\mathcal{O}}$  hides the term dependent on  $-\log(1-p)$ . Similar result is established for the Hessian sample average approximation. The result for the function estimates, given the

variance  $\sigma_f$  is a simpler version of the same inequalities and can be derived using Chebychev inequality.

Using these results, we can obtain  $\alpha$ -probabilistically fully-linear models as follows. Compute  $f_k$  by averaging  $\mathcal{O}(\frac{\sigma_f^2}{\Delta_k^4} \log(\frac{1}{1-\sqrt{\alpha}}))$  samples  $\tilde{f}(x_k, \xi)$ , then compute  $g_k$  as an average of  $\tilde{\mathcal{O}}(\frac{\sigma_g^2}{\kappa_{eg}^2 \Delta_k^2} \log(\frac{1}{1-\sqrt{\alpha}}))$  gradient samples  $\nabla \tilde{f}(x_k, \xi)$ . This ensures that  $\|g_k - \nabla f(x_k)\| \leq \kappa_{eg} \Delta_k$  and  $|f_k - f(x_k)| \leq \Delta_k^2$  with probability at least  $\alpha$ . Condition  $|m(y) - f(y)| \leq \kappa_{ef} \Delta_k^2$  follows automatically with appropriately chosen  $\kappa_{ef}$ .

We want to note that all sample sizes are determined by quantities that are generally either chosen or known by the algorithm or can be correctly estimated.

Similarly, we can obtain  $\beta$ -probabilistically  $\epsilon_F$ -accurate estimates  $f_k^0$  and  $f_k^s$  by averaging  $\tilde{\mathcal{O}}(\frac{\sigma_f^2}{\epsilon_F^2 \Delta_k^4} \log(\frac{1}{1-\sqrt{\beta}}))$  samples of  $\tilde{f}(x_k, \xi)$  and  $\tilde{f}(x_k + s_k, \xi)$  for each., respectively.

In the case of simulation optimization, when  $\nabla f(x, \xi)$  is not available,  $\kappa$ -fully-linear models  $m_k$  can be constructed via polynomial interpolation [10], and  $\alpha$ -probabilistically  $\kappa$ -fully-linear models are similarly obtained by combining interpolation and sufficiently accurate function value estimates (see, e.g. [26]).

Another setting that is explored in [8] is when  $f(x)$  and, possibly,  $\nabla \tilde{f}(x)$  are computed (accurately) via some procedure, which may fail with some small, but fixed, probability. In this case  $\tilde{f}(x, \xi)$  and  $\nabla \tilde{f}(x, \xi)$  are the true values of the function and the gradients, or some arbitrarily corrupted values. If the probability of failure is sufficiently small, conditioned on the past, the STORM algorithm still converges almost surely. See [8] for details.

## 4 The second order STORM algorithm

We now introduce a variant of Algorithm 1 that attempts to achieve second order criticality in the stochastic setting; we use the same notation as in Algorithm 1. Firstly, the model minimization may need to provide more than just the Cauchy decrease (9), namely, we require that on each iteration  $k$  and for all model realizations  $m_k$  (as defined in Step 2) of  $M_k$ , we are able to compute a step  $s_k$ , so that the following level of second order improvement is achieved,

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{scd}}{2} \max \left\{ \|g_k\| \min \left[ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right], \max\{-\lambda_{\min}(H_k), 0\} \delta_k^2 \right\}. \quad (37)$$

for some constant  $\kappa_{scd} \in (0, 1]$ . A step satisfying this (typical second order) assumption is given, for instance, by computing both the Cauchy step and, in the presence of negative curvature in the model, the *eigenstep*, and by choosing the one that provides the largest reduction in the model<sup>3</sup> [9].

In our analysis (not in the algorithm), we will use — instead of just the true gradient of  $f$  — the following measure of proximity to a second order stationary point for the objective  $f$ ,

$$\tau(x) = \max \{ \|\nabla f(x)\|, -\lambda_{\min}(\nabla^2 f(x)) \}. \quad (38)$$

The corresponding optimality measure for the model  $m_k$  is defined slightly differently than

---

<sup>3</sup>The eigenstep is the minimizer of the quadratic model in the trust region along an eigenvector corresponding to the smallest (negative) eigenvalue of  $H_k$ .

above, following [3],

$$\tau^m(x) = \max \left\{ \min \left[ \|\nabla m(x)\|, \frac{\|\nabla m(x)\|}{\|\nabla^2 m(x)\|} \right], -\lambda_{\min}(\nabla^2 m(x)) \right\}. \quad (39)$$

The additional term in (39) compared to (38) is needed because there is no longer a bound on the model Hessians on all iterations, as in the first order case. We will only use (39) at the iterate  $x_k$ , in which case, it becomes

$$\tau_k^m = \max \left\{ \min \left[ \|g_k\|, \frac{\|g_k\|}{\|H_k\|} \right], -\lambda_{\min}(H_k) \right\}. \quad (40)$$

We are now ready to present our second order STORM algorithm, by modifying the first order STORM algorithm.

---

**Algorithm 2** SECOND ORDER STOCHASTIC DFO WITH RANDOM MODELS

---

Perform the STORM Algorithm 1, with the following modifications to Steps 3, 5 and 6:

- 3: (Step calculation) Compute  $s_k = \arg \min_{s: \|s\| \leq \delta_k} m_k(s)$  (approximately) so that it satisfies condition (37).
  - 5: (Acceptance of the trial point): If  $\rho_k \geq \eta_1$  and  $\tau_k^m \geq \eta_2 \delta_k$ , set  $x_{k+1} = x_k + s_k$ ; otherwise, set  $x_{k+1} = x_k$ .
  - 6: (Trust-region radius update): If  $\rho_k \geq \eta_1$  and  $\tau_k^m \geq \eta_2 \delta_k$ , set  $\delta_{k+1} = \min\{\gamma \delta_k, \delta_{\max}\}$ ; otherwise  $\delta_{k+1} = \gamma^{-1} \delta_k$ ;  $k \leftarrow k + 1$  and go to step 2.
- 

The analysis for the second order STORM variant will again use the framework proposed in Section 2, thus serving as another illustration of the applicability of our generic set up. Before applying this framework we need to describe our assumptions required for a second order analysis.

In terms of problem assumptions, we will need one more order of smoothness compared to first order ones (Assumption 3.1).

**Assumption 4.1.** *Assume that  $f$  satisfies Assumption 3.1 and that it is twice continuously differentiable on  $\mathcal{X}$ , and also that the Hessian  $\nabla^2 f$  is  $L_H$ -Lipschitz continuous.*

#### 4.1 Assumptions on the second order STORM algorithm

Let us now introduce a measure of second order quality or accuracy of the models  $m_k$  (see [10, 4, 19] for more details).

**Definition 4.1.** *1) A function  $m_k$  is a  $\kappa$ -fully quadratic model of  $f$  on  $B(x_k, \delta_k)$  provided, for  $\kappa = (\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$  and  $\forall y \in B(x_k, \delta_k)$ ,*

$$\begin{aligned} \|\nabla^2 f(x_k) - H_k\| &\leq \kappa_{eh} \delta_k, \\ \|\nabla f(y) - \nabla m_k(y)\| &\leq \kappa_{eg} \delta_k^2, \\ \|f(y) - m(y)\| &\leq \kappa_{ef} \delta_k^3. \end{aligned}$$

2) The estimates  $f_k^0$  and  $f_k^s$  are said to be  $\epsilon_F$ -s.o.-accurate (s.o. for "second order") estimates of  $f(x_k)$  and  $f(x_k + s_k)$ , respectively, for a given  $\delta_k$  if

$$|f_k^0 - f(x_k)| \leq \epsilon_F \delta_k^3 \text{ and } |f_k^s - f(x_k + s_k)| \leq \epsilon_F \delta_k^3. \quad (41)$$

**Definition 4.2.** [2] A sequence of random models  $\{M_k\}$  is said to be  $\alpha$ -**probabilistically  $\kappa$ -fully quadratic** with respect to the corresponding sequence  $\{B(X_k, \Delta_k)\}$  if the events

$$I_k = \mathbb{1}\{M_k \text{ is a } \kappa\text{-fully quadratic model of } f \text{ on } B(X_k, \Delta_k)\} \quad (42)$$

satisfy the condition

$$P(I_k = 1 | \mathcal{F}_{k-1}^{M \cdot F}) \geq \alpha.$$

**Definition 4.3.** A sequence of random estimates  $\{F_k^0, F_k^s\}$  is said to be  $\beta$ -**probabilistically  $\epsilon_F$ -s.o.-accurate** with respect to the corresponding sequence  $\{X_k, \Delta_k, S_k\}$  if the events

$$J_k = \mathbb{1}\{F_k^0, F_k^s \text{ are } \epsilon_F\text{-s.o. accurate estimates of } f(x_k) \text{ and } f(x_k + s_k), \text{ respectively, for } \Delta_k\} \quad (43)$$

satisfy the condition

$$P(J_k = 1 | \mathcal{F}_{k-1/2}^{M \cdot F}) \geq \beta,$$

where  $\epsilon_F$  is a fixed constant.

We will no longer assume that the Hessian  $H_k$  of the models is bounded in norm, since we cannot simply disregard large Hessian model values without possibly affecting the chances of the model being fully quadratic. However, a simple analysis can show that  $\|H_k\|$  is uniformly bounded from above for any fully quadratic model  $m_k$  (although we may not know what this bound is and hence may not be able to use it in an algorithm).

**Lemma 4.1.** [3] Let Assumption 4.1 hold. Given constants  $\kappa_{eh}$ ,  $\kappa_{eg}$ ,  $\kappa_{ef}$ , and  $\delta_{\max}$ , there exists a constant  $\kappa_{bhm} \geq 1$  such that for every  $k$  and every realization  $m_k$  of  $M_k$  which is a  $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic model of  $f$  on  $B(x_k, \delta_k)$  with  $x_k \in \mathcal{X}$  and  $\delta_k \leq \delta_{\max}$  we have

$$\|H_k\| \leq \kappa_{bhm}.$$

*Proof.* The proof follows trivially from the definition of fully quadratic models and the assumption that  $\|\nabla^2 f\| \leq L$  is bounded above on  $\mathcal{X}$ , which follows from the gradient of  $f$  being Lipschitz continuous with constant  $L$ . Then we can let  $\kappa_{bhm} := \delta_{\max} \kappa_{eh} + L$ .

For our convergence analysis we again need to impose conditions on the nature of the stochastic (and deterministic) information used by our algorithm.

**Assumption 4.2.** The following hold for the quantities used in Algorithm 2

- (a) The sequence of random models  $M_k$ , generated by Algorithm 2, is  $\alpha$ -probabilistically  $\kappa$ -fully quadratic, for some  $\kappa = (\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$  and for a sufficiently large  $\alpha \in (0, 1)$ .
- (b) The sequence of random estimates  $\{F_k^0, F_k^s\}$  generated by Algorithm 2 is  $\beta$ -probabilistically  $\epsilon_F$ -s.o.-accurate for  $\epsilon_F \leq \kappa_{ef}$  and  $\epsilon_F < \frac{1}{4} \eta_1 \eta_2 \kappa_{scd} \min\{\eta_2, 1\}$ , and for a sufficiently large  $\beta \in (0, 1)$ .

Note that as in the first order case, we are able to allow for unrestricted values of  $\eta_2$  in Algorithm 2, with a potential trade-off of increased accuracy on the function estimates.

## 4.2 Useful preliminary results for second order STORM analysis

The analysis of Algorithm 2 is similar to that of the first order STORM described in Section 3. However, there are more cases to consider and the convergence rate to the second order stationary point is different, as it is in the deterministic case. There will also be another significant difference, such as a requirement for an additional assumption on function estimates, to be detailed in the next section. First, we state and prove the analogues of Lemmas 3.1–3.4 for the function decrease in terms of first and second order optimality. The first three lemmas are almost identical to Lemmas 3.1–3.3, except that the models are assumed to be fully quadratic, instead of fully linear; the model decrease condition (37) is now used and condition  $\|H_k\| \leq \kappa_{bhm}$  is only valid when the model  $m_k$  is fully-quadratic according to Lemma 4.1. For completeness, we have delegated the proofs of Lemmas 4.2–Lemmas 4.4 to the Appendix.

**Lemma 4.2.** [Good quadratic model  $\Rightarrow$  function reduction in  $\|g_k\|$ ] *Let Assumption 4.1 hold. Suppose that a model  $m_k$  is a  $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic model of  $f$  on  $B(x_k, \delta_k)$ . If  $\delta_k \leq 1$  and*

$$\delta_k \leq \min \left\{ \frac{1}{\kappa_{bhm}}, \frac{\kappa_{scd}}{8\kappa_{ef}} \right\} \|g_k\|,$$

*then the trial step  $s_k$  leads to an improvement in  $f(x_k + s_k)$  such that*

$$f(x_k + s_k) - f(x_k) \leq -\frac{\kappa_{scd}}{4} \|g_k\| \delta_k.$$

**Lemma 4.3.** [Good quadratic model  $\Rightarrow$  function reduction in  $\|\nabla f(x_k)\|$ ] *Let Assumption 4.1 hold. Suppose that a model is  $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic on  $B(x_k, \delta_k)$ . If  $\delta_k \leq 1$  and*

$$\delta_k \leq \min \left\{ \frac{1}{\kappa_{bhm} + \kappa_{eg}}, \frac{1}{\frac{8\kappa_{ef}}{\kappa_{scd}} + \kappa_{eg}} \right\} \|\nabla f(x_k)\|, \quad (44)$$

*then the trial step  $s_k$  leads to an improvement in  $f(x_k + s_k)$  such that*

$$f(x_k + s_k) - f(x_k) \leq -C_1 \|\nabla f(x_k)\| \delta_k, \quad (45)$$

*for any  $C_1 \leq \frac{\kappa_{scd}}{4} \cdot \max \left\{ \frac{\kappa_{bhm}}{\kappa_{bhm} + \kappa_{eg}}, \frac{8\kappa_{ef}}{8\kappa_{ef} + \kappa_{scd}\kappa_{eg}} \right\}$ .*

**Lemma 4.4.** [Good quadratic model + good s.o. estimates  $\Rightarrow$  successful step] *Let Assumption 4.1 hold. Suppose that  $m_k$  is  $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic on  $B(x_k, \delta_k)$  and the estimates  $\{f_k^0, f_k^s\}$  are  $\epsilon_F$ -s.o. accurate with  $\epsilon_F \leq \kappa_{ef}$ . If  $\delta_k \leq 1$  and*

$$\delta_k \leq \min \left\{ \frac{1}{\kappa_{bhm}}, \frac{1}{\eta_2 \kappa_{bhm}}, \frac{\kappa_{scd}(1 - \eta_1)}{8\kappa_{ef}} \right\} \|g_k\|, \quad (46)$$

*then the  $k$ -th iteration is successful.*

The remaining lemmas address the case of negative curvature in the model and that of second order accurate estimates.

**Lemma 4.5.** [Good quadratic model  $\Rightarrow$  function reduction in  $\lambda_{\min}(H_k)$ ] *Let Assumption 4.1 hold. Suppose that a model  $m_k$  is a  $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic model of  $f$  on  $B(x_k, \delta_k)$ . If*

$$\delta_k \leq \frac{\kappa_{scd}}{8\kappa_{ef}} (-\lambda_{\min}(H_k)), \quad (47)$$

then the trial step  $s_k$  leads to an improvement in  $f(x_k + s_k)$  such that

$$f(x_k + s_k) - f(x_k) \leq -\frac{\kappa_{scd}}{4}(-\lambda_{\min}(H_k))\delta_k^2. \quad (48)$$

*Proof.* Whenever  $\lambda_{\min}(H_k) < 0$ , the optimal decrease condition (37) ensures that

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{scd}}{2}(-\lambda_{\min}(H_k))\delta_k^2.$$

Since the model is  $\kappa$ -fully quadratic, the improvement in  $f$  achieved by  $s_k$  is

$$\begin{aligned} & f(x_k + s_k) - f(x_k) \\ &= f(x_k + s_k) - m(x_k + s_k) + m(x_k + s_k) - m(x_k) + m(x_k) - f(x_k) \\ &\leq 2\kappa_{ef}\delta_k^3 - \frac{\kappa_{scd}}{2}(-\lambda_{\min}(H_k))\delta_k^2 \\ &\leq -\frac{\kappa_{scd}}{4}(-\lambda_{\min}(H_k))\delta_k^2 \end{aligned}$$

where the last inequality is implied by (47).  $\square$

**Lemma 4.6.** [Good quadratic model  $\Rightarrow$  function reduction in  $\lambda_{\min}(\nabla^2 f(x_k))$ ] *Let Assumption 4.1 hold. Suppose that a model is  $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic on  $B(x_k, \delta_k)$ . If*

$$\delta_k \leq \frac{1}{\frac{8\kappa_{ef}}{\kappa_{scd}} + \kappa_{eh}}(-\lambda_{\min}(\nabla^2 f(x_k))), \quad (49)$$

then the trial step  $s_k$  leads to an improvement in  $f(x_k + s_k)$  such that

$$f(x_k + s_k) - f(x_k) \leq -C_4(-\lambda_{\min}(\nabla^2 f(x_k)))\delta_k^2, \quad (50)$$

for any  $C_4 \leq \frac{\kappa_{scd}}{4} \cdot \frac{8\kappa_{ef}}{8\kappa_{ef} + \kappa_{scd}\kappa_{eh}}$ .

*Proof.* The definition of a  $\kappa$ -fully-quadratic model, by Corollary 8.5.6 from [15] yield that

$$-\lambda_{\min}(H_k) \geq (-\lambda_{\min}(\nabla^2 f(x_k))) - \kappa_{eh}\delta_k. \quad (51)$$

Since condition (49) implies that  $-\lambda_{\min}(\nabla^2 f(x_k)) \geq (\frac{8\kappa_{ef}}{\kappa_{scd}} + \kappa_{eh})\delta_k$ , we have

$$-\lambda_{\min}(H_k) \geq \frac{8\kappa_{ef}}{\kappa_{scd}}\delta_k.$$

Hence, the conditions of Lemma 4.5 hold and we have

$$f(x_k + s_k) - f(x_k) \leq -\frac{\kappa_{scd}}{4}(-\lambda_{\min}(H_k))\delta_k^2. \quad (52)$$

From (51) and (49), we also have

$$-\lambda_{\min}(H_k) \geq \frac{8\kappa_{ef}}{8\kappa_{ef} + \kappa_{scd}\kappa_{eg}}(-\lambda_{\min}(\nabla^2 f(x_k))) \quad (53)$$

Combining (52) and (53) yields (50).  $\square$

**Lemma 4.7.** [Good quadratic model + good s.o. estimates  $\Rightarrow$  successful step] *Let Assumption 4.1 hold. Suppose that  $m_k$  is  $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic on  $B(x_k, \delta_k)$  and the estimates  $\{f_k^0, f_k^s\}$  are  $\epsilon_F$ -s.o. accurate with  $\epsilon_F \leq \kappa_{ef}$ . If*

$$\delta_k \leq \min \left\{ \frac{1}{\eta_2}, \frac{\kappa_{scd}(1-\eta_1)}{8\kappa_{ef}} \right\} (-\lambda_{\min}(H_k)), \quad (54)$$

then the  $k$ -th iteration is successful.

*Proof.* From the model decrease condition (37)

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{scd}}{2} (-\lambda_{\min}(H_k)) \delta_k^2. \quad (55)$$

The model  $m_k$  being  $(\kappa_{ef}, \kappa_{eg})$ -fully quadratic implies that

$$|f(x_k) - m_k(x_k)| \leq \kappa_{ef} \delta_k^3, \text{ and} \quad (56)$$

$$|f(x_k + s_k) - m_k(x_k + s_k)| \leq \kappa_{ef} \delta_k^3. \quad (57)$$

Since the estimates are  $\epsilon_F$ -s.o. accurate with  $\epsilon_F \leq \kappa_{ef}$ , we obtain

$$|f_k^0 - f(x_k)| \leq \kappa_{ef} \delta_k^3, \text{ and } |f_k^s - f(x_k + s_k)| \leq \kappa_{ef} \delta_k^3. \quad (58)$$

We have

$$\begin{aligned} \rho_k &= \frac{f_k^0 - f_k^s}{m_k(x_k) - m_k(x_k + s_k)} \\ &= \frac{f_k^0 - f(x_k)}{m_k(x_k) - m_k(x_k + s_k)} + \frac{f(x_k) - m_k(x_k)}{m_k(x_k) - m_k(x_k + s_k)} + \frac{m_k(x_k) - m_k(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} \\ &\quad + \frac{m_k(x_k + s_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} + \frac{f(x_k + s_k) - f_k^s}{m_k(x_k) - m_k(x_k + s_k)}, \end{aligned}$$

which, combined with (55)-(58), implies

$$|\rho_k - 1| \leq \frac{8\kappa_{ef}\delta_k^3}{\kappa_{scd}(-\lambda_{\min}(H_k))\delta_k^2} \leq 1 - \eta_1,$$

where we have used the assumptions  $\delta_k \leq \frac{\kappa_{scd}(1-\eta_1)}{8\kappa_{ef}}(-\lambda_{\min}(H_k))$  to deduce the last inequality. Hence,  $\rho_k \geq \eta_1$ . Moreover, the first term in (54) and (40) imply  $\tau_k^m \geq (-\lambda_{\min}(H_k)) \geq \eta_2 \delta_k$ . Thus the  $k$ -th iteration is successful.  $\square$

**Lemma 4.8.** [Good s.o. estimates + successful step  $\Rightarrow$  function decrease in  $\delta_k^3$ ] *Assume the estimates  $\{f_k^0, f_k^s\}$  are  $\epsilon_F$ -s.o. accurate with  $\epsilon_F < \frac{1}{4}\eta_1\eta_2 \min\{1, \eta_2\}\kappa_{scd}$ . If  $\delta_k \leq 1$  and a trial step  $s_k$  is accepted (a successful iteration occurs), then the improvement in  $f$  is bounded below as follows*

$$f(x_{k+1}) - f(x_k) \leq -C_2 \delta_k^3, \quad (59)$$

where

$$C_2 = \frac{1}{2}\eta_1\eta_2 \min\{1, \eta_2\}\kappa_{scd} - 2\epsilon_F > 0. \quad (60)$$

*Proof.* An iteration being successful indicates that  $\rho \geq \eta_1$  and either  $\min \left\{ \|g_k\|, \frac{\|g_k\|}{\|H_k\|} \right\} \geq \eta_2 \delta_k$  or  $-\lambda_{\min}(H_k) \geq \eta_2 \delta_k$ . First let us assume that  $\min \left\{ \|g_k\|, \frac{\|g_k\|}{\|H_k\|} \right\} \geq \eta_2 \delta_k$ ; then

$$\begin{aligned} f_k^0 - f_k^s &\geq \eta_1 (m_k(x_k) - m_k(x_k + s_k)) \\ &\geq \eta_1 \frac{\kappa_{scd}}{2} \|g_k\| \min \left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\} \\ &\geq \frac{1}{2} \eta_1 \kappa_{scd} \eta_2 \min\{1, \eta_2\} \delta_k^2 \\ &\geq \frac{1}{2} \eta_1 \kappa_{scd} \eta_2 \min\{1, \eta_2\} \delta_k^3, \end{aligned}$$

where we also used  $\delta_k \leq 1$ .

Let us now assume that  $-\lambda_{\min}(H_k) \geq \eta_2 \delta_k$ , thus,

$$\begin{aligned} f_k^0 - f_k^s &\geq \eta_1 (m_k(x_k) - m_k(x_k + s_k)) \\ &\geq \eta_1 \frac{\kappa_{scd}}{2} (-\lambda_{\min}(H_k)) \delta_k^2 \\ &\geq \frac{1}{2} \eta_1 \eta_2 \kappa_{scd} \delta_k^3 \\ &\geq \frac{1}{2} \eta_1 \kappa_{scd} \eta_2 \min\{1, \eta_2\} \delta_k^3. \end{aligned}$$

Thus in both cases, using the fact that the estimates are  $\epsilon_F$ -s.o. accurate, we have

$$f(x_k + s_k) - f(x_k) = f(x_k + s_k) - f_k^s + f_k^s - f_k^0 + f_k^0 - f(x_k) \leq -C_2 \delta_k^3,$$

here  $C_2$  is defined in (60). □

**Choosing constants** To simplify our calculations, just like for the first order case, we particularize our choices of constants, but we will clearly state when we use these choices. We let  $\kappa_{scd} = 0.5$ ,  $\eta_1 = 0.1$ ,  $\gamma = 2$ ,  $\delta_{\max} = 1$  and  $\kappa_{ef} = \kappa_{eg} = \kappa_{eh} = \Theta(\bar{L})$ , where  $\bar{L} = \max\{L, L_H\}$ . To satisfy Assumption 4.2, we let  $\epsilon_F = \frac{1}{160} \eta_2 \min\{1, \eta_2\} \leq \kappa_{eh}$  and  $\eta_2 \leq 18$ . Note that we cannot impose upper bounds on  $\kappa_{bhm}$  as the latter cannot be chosen freely, namely, from Lemma 4.1, we have  $\kappa_{bhm} = \kappa_{eh} + L \leq 2 \max\{\kappa_{eh}, \bar{L}\}$ .

### 4.3 Defining and analysing the process $\{\Phi_k, \Delta_k\}$ for the second order case

As the order of the function decrease that can be guaranteed on good iterations of Algorithm 2 changes from the first order  $\delta_k^2$  to  $\delta_k^3$  due to second order terms, we must modify the process  $\Phi_k$  accordingly. Namely, we let  $\{\Phi_k, \Delta_k\}$  be derived from the process generated by Algorithm 2, with  $\Delta_k$  - the trust region radius and

$$\Phi_k = \nu f(x_k) + (1 - \nu) \Delta_k^3 \tag{61}$$

where  $\nu \in (0, 1)$  is a deterministic, large enough constant, which we will define later, and  $\Phi_k \geq 0$ . We also define the random time

$$T_\epsilon = \inf\{k \geq 0 : \|\nabla f(X_k)\| \leq \epsilon \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(X_k)) \geq -\epsilon\}, \tag{62}$$

which is a stopping time for the stochastic process defined by Algorithm 2 and hence for  $\{\Phi_k, \Delta_k\}$  as in (61). To bound the expected stopping time  $\mathbb{E}(T_\epsilon)$  for Algorithm 2, we show that Assumption 2.1 is satisfied for  $\{\Phi_k, \Delta_k\}$  in (61) and apply the results of Section 2.

Very similarly to the first order case, we can show that Assumption 2.1(i)–(ii) holds with  $\lambda = \log \gamma$ , and with the following new settings

$$\Delta_\epsilon = \frac{\epsilon}{\zeta}, \quad \text{for } \zeta \geq \max\{\kappa_{eg}, \kappa_{eh}\} + \max\left\{\eta_2 \kappa_{bhm}, \kappa_{bhm}, \frac{8\kappa_{ef}}{\kappa_{scd}(1-\eta_1)}\right\}, \quad (63)$$

with  $\epsilon \in (0, 1]$ , and the (old) assumption that  $\Delta_\epsilon = \gamma^i \delta_0$  for some  $i \leq 0$ . Note that (63),  $\epsilon \in (0, 1]$  and  $\kappa_{bhm} \geq 1$  imply that  $\Delta_\epsilon \leq 1$ .

**Lemma 4.9.** *Let Assumptions 4.1 and 4.2 hold. Let  $\alpha$  and  $\beta$  be such that  $\alpha\beta \geq 1/2$ , then Assumption 2.1(ii) is satisfied for Algorithm 2 with  $W_k = 2(I_k J_k - \frac{1}{2})$ ,  $\lambda = \log \gamma$  and  $p = \alpha\beta$ .*

*Proof.* The proof follows similarly to that of Lemma 3.5 and we show that, conditioned on  $T_\epsilon > k$  (i.e.  $I(T_\epsilon > k) = 1$ ), where  $T_\epsilon$  is now defined in (62), (23) holds with  $\Delta_\epsilon$  defined in (63). The only case that differs (from the first order proof) and needs addressing is when  $\Delta_k \leq \Delta_\epsilon$ . Then, conditioned on  $T_\epsilon > k$ , we have that either  $\|\nabla f(X_k)\| \geq \epsilon$  or  $\lambda_{\min}(\nabla^2 f(X_k)) \leq -\epsilon$  and hence, from the definition of  $\zeta$  in (63), we know that

$$\|\nabla f(x_k)\| \geq \left( \kappa_{eg} + \max\left\{\eta_2 \kappa_{bhm}, \kappa_{bhm}, \frac{8\kappa_{ef}}{\kappa_{scd}(1-\eta_1)}\right\} \right) \delta_k \quad (64)$$

or that

$$-\lambda_{\min}(\nabla^2 f(X_k)) \geq \left( \kappa_{eh} + \max\left\{\eta_2, \frac{8\kappa_{ef}}{\kappa_{scd}(1-\eta_1)}\right\} \right) \delta_k, \quad (65)$$

where we also used that  $\kappa_{bhm} \geq 1$ . Assume that  $I_k = 1$  and  $J_k = 1$ , i.e., both the model and the estimates are good on iteration  $k$ . Since the model  $m_k$  is  $\kappa$ -fully quadratic and  $\delta_k \leq \Delta_\epsilon \leq 1$ , then if (64) holds, we have

$$\|g_k\| \geq \|\nabla f(x_k)\| - \kappa_{eg} \delta_k \geq (\zeta - \kappa_{eg}) \delta_k \geq \max\left\{\eta_2 \kappa_{bhm}, \kappa_{bhm}, \frac{8\kappa_{ef}}{\kappa_{scd}(1-\eta_1)}\right\} \delta_k, \quad (66)$$

and if (65) holds, we have

$$-\lambda_{\min}(H_k) \geq -\lambda_{\min}(\nabla^2 f(X_k)) - \kappa_{eh} \delta_k \geq \max\left\{\eta_2, \frac{8\kappa_{ef}}{\kappa_{scd}(1-\eta_1)}\right\} \delta_k. \quad (67)$$

As the estimates  $\{f_k^0, f_k^s\}$  are  $\epsilon_F$ -s.o. accurate, with  $\epsilon_F \leq \kappa_{ef}$ , (66) and (67) imply that condition (46) in Lemma 4.4 and (54) in Lemma 4.7 hold, respectively. Thus in both cases, iteration  $k$  is successful, i.e.  $x_{k+1} = x_k + s_k$  and  $\delta_{k+1} = \max\{\delta_{max}, \gamma \delta_k\}$ . If  $I_k J_k = 0$ , then  $\delta_{k+1} \geq \gamma^{-1} \delta_k$  simply by the dynamics of Algorithm 2. Finally, observing that  $P\{I_k J_k\} \geq p = \alpha\beta$  we conclude that (23) implies Assumption 2.1(ii).  $\square$

To show that Assumption 2.1(iii) holds, we need an additional assumption on the accuracy of the function estimates. We also require, for simplicity, an upper bound on the trust-region radius in Algorithm 2<sup>4</sup>.

<sup>4</sup>This restriction can be avoided if one allows a more involved discussion on dominating terms in the proofs of Lemmas 4.2–4.4 and 4.8, and in the proof of the main result.

**Assumption 4.3.** (a) *There exists a constant  $\kappa_F$  such that at any iteration  $k$ ,*

$$\mathbb{E}[|F_k^0 - f(x_k^0)| | \mathcal{F}_{k-1/2}^{M \cdot F}] \leq \kappa_F \delta_k^3$$

and

$$\mathbb{E}[|F_k^s - f(x_k + s_k)| | \mathcal{F}_{k-1/2}^{M \cdot F}] \leq \kappa_F \delta_k^3.$$

(b) *The upper bound  $\delta_{\max}$  in Algorithm 2 is chosen so that  $\delta_{\max} \leq 1$ .*

Note that the bound on the expectation of  $|F_k^0 - f(x_k^0)|$  and  $|F_k^s - f(x_k + s_k)|$ , in principle, implies that the estimates are  $\beta$ -probabilistically  $\epsilon_F$ -s.o. accurate. However, for  $\epsilon_F$  to satisfy the conditions in Assumption 4.2 (b) conditions would have to be imposed on  $\kappa_F$ . Thus, for our purposes here, we choose to have any finite  $\kappa_F > 0$  and to impose the bound only on  $\epsilon_F$ .

Assumption 4.3(a) is needed for the case when we have a bad model and bad estimates, and when the (true) objective may increase after a successful step. Without this assumption, it is possible that the increase in the objective is at most of order  $\delta_k^2$  (due to first order terms), while the decrease (on other successful steps) may be smaller, of order  $\delta_k^3$  (due to second order terms). Such a situation would make it impossible to balance out the increase and decrease in the objective over the course of the algorithm in such a way to ensure that the stochastic process  $\Phi_k$  decreases on average.

We now prove that Assumption 2.1(iii) holds for Algorithm 2.

**Theorem 4.1.** *Let Assumptions 4.1, 4.2 and 4.3 hold. Then, there exist probabilities  $\alpha$  and  $\beta$  and a constant  $\Theta > 0$  such that, conditioned on  $T_\epsilon > k$ , for each iteration  $k$  of Algorithm 2, we have*

$$1(T_\epsilon > k) \mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M \cdot F}] \leq -1(T_\epsilon > k) \Theta \Delta_k^3, \quad (68)$$

where  $T_\epsilon$  is defined in (62), and  $\Phi_k$  in (61).

Moreover, under the particular choice of constants described on page 26, let  $\alpha$  and  $\beta$  satisfy

$$(1 - \alpha)(1 - \beta) \leq \min \left\{ 0.05, 0.0003 \frac{\eta_2 \min\{1, \eta_2\}}{\kappa_F} \right\} \quad (69)$$

and

$$\beta \geq \frac{\kappa_F + 0.008\eta_2 \min\{1, \eta_2\}}{\kappa_F + 0.0085\eta_2 \min\{1, \eta_2\}} \quad (70)$$

Then,  $\zeta = 20\kappa_{bhm} = 20(\kappa_{eh} + L)$  and  $\Theta \geq 6 \cdot 10^{-4} \eta_2 \min\{1, \eta_2\}$ .

*Proof.* Since (68) easily holds if  $T_\epsilon \leq k$ , we assume in what follows that  $T_\epsilon > k$  and so  $\tau(x_k) > \epsilon$ , where  $\tau(x)$  is defined in (38). We will consider two possible cases:  $\tau(x_k) \geq \zeta \delta_k$  and  $\tau(x_k) < \zeta \delta_k$ , where  $\zeta$  is defined in (63). We show that (68) holds in both cases and so for all  $k < T_\epsilon$ . Let  $\nu \in (0, 1)$  be such that

$$\frac{\nu}{1 - \nu} \geq \frac{\gamma^3}{\min\{\zeta C_1, \zeta C_4, C_2\}}, \quad (71)$$

with  $C_1$  defined as in Lemma 4.3,  $C_4$  in Lemma 4.6, and  $C_2$  in Lemma 4.8. Note that on all successful iterations,  $x_{k+1} = x_k + s_k$  and  $\delta_{k+1} = \min\{\gamma \delta_k, \delta_{max}\}$  with  $\gamma > 1$ , hence

$$\phi_{k+1} - \phi_k \leq \nu(f(x_{k+1}) - f(x_k)) + (1 - \nu)(\gamma^3 - 1)\delta_k^3. \quad (72)$$

On all unsuccessful iterations,  $x_{k+1} = x_k$  and  $\delta_{k+1} = \frac{1}{\gamma}\delta_k$ , i.e.

$$\phi_{k+1} - \phi_k = (1 - \nu)\left(\frac{1}{\gamma^3} - 1\right)\delta_k^3 \equiv b_1 < 0. \quad (73)$$

**Case 1:**  $\tau(x_k) = \max\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\} \geq \zeta\delta_k$ , **where  $\zeta$  is defined in (63).**

- a.  $I_k = 1$  and  $J_k = 1$ , i.e., both the model and the estimates are good on iteration  $k$ . From the definition of  $\zeta$  and Case 1, we know that either (64) or (65) hold. Since  $I_k = 1$  and  $\delta_{\max} \leq 1$  (Assumption 4.3(b)), (64) and (65) imply that either condition (44) in Lemma 4.3 or condition (49) in Lemma 4.6 hold. Therefore, the trial step  $s_k$  leads to a decrease in  $f$  as in (45) or as in (50). Again from  $I_k = 1$  and  $\delta_{\max} \leq 1$ , (64) or (65) imply that (66) or (67) hold, respectively. As  $J_k = 1$ , and  $\epsilon_F \leq \kappa_{ef}$ , (66) and (67) imply that condition (46) in Lemma 4.4 or (54) in Lemma 4.7 hold, respectively. Thus in both cases, iteration  $k$  is successful, i.e.  $x_{k+1} = x_k + s_k$  and  $\delta_{k+1} = \max\{\delta_{max}, \gamma\delta_k\}$ .

Combining (45) and (72), we get

$$\phi_{k+1} - \phi_k \leq -\nu C_1 \|\nabla f(x_k)\| \delta_k^2 + (1 - \nu)(\gamma^3 - 1)\delta_k^3, \quad (74)$$

with  $C_1$  defined in Lemma 4.3. Since  $\|\nabla f(x_k)\| \geq \zeta\delta_k$  we have

$$\phi_{k+1} - \phi_k \leq [-\nu C_1 \zeta + (1 - \nu)(\gamma^3 - 1)]\delta_k^3 \leq b_1, \quad (75)$$

with  $b_1$  defined in (73), for  $\nu \in (0, 1)$  satisfying (71).

Combining (50) and (72), we get

$$\phi_{k+1} - \phi_k \leq \nu C_4 \lambda_{\min}(\nabla^2 f(X^k)) \delta_k^2 + (1 - \nu)(\gamma^3 - 1)\delta_k^3, \quad (76)$$

with  $C_4$  defined in Lemma 4.6. Again, since  $-\lambda_{\min}(\nabla^2 f(X^k)) \geq \zeta\delta_k$  we have

$$\phi_{k+1} - \phi_k \leq [-\nu C_4 \zeta + (1 - \nu)(\gamma^3 - 1)]\delta_k^3 \leq b_1, \quad (77)$$

with  $b_1$  defined in (73), for  $\nu \in (0, 1)$  satisfying (71).

- b.  $I_k = 1$  and  $J_k = 0$ , i.e., we have a good model and bad estimates on iteration  $k$ . Then the analysis of case (a) applies, and either by Lemma 4.3 or 4.6,  $s_k$  yields a sufficient decrease in  $f$ . However, the step can be erroneously rejected, because of inaccurate probabilistic estimates, in which case we have an unsuccessful iteration and (73) holds. Since (71) holds, (73) applies whether the iteration is successful or not.

- c.  $I_k = 0$  and  $J_k = 1$ , i.e., we have a bad model and good estimates on iteration  $k$ . In this case, again, iteration  $k$  can be either successful or unsuccessful; in the latter case, (73) holds. In the former, since the estimates are  $\epsilon_F$ -accurate and (41) holds, then by Lemma 4.8 and Assumption 4.3(b), (59) holds with some  $C_2 > 0$ , and so,

$$\phi_{k+1} - \phi_k \leq [-\nu C_2 + (1 - \nu)(\gamma^3 - 1)]\delta_k^3 \leq b_1, \quad (78)$$

due to the choice of  $\nu$  satisfying (71).

d.  $I_k = 0$  and  $J_k = 0$ , i.e., both the model and the estimates are bad on iteration  $k$ . Inaccurate estimates can cause the algorithm to accept a bad step, which may lead to an increase both in  $f$  and in  $\delta_k$ . Hence in this case  $\phi_{k+1} - \phi_k$  may be positive. We can derive a bound on the increase of  $f(x_k)$  on successful steps in terms of the error of the estimates

$$\begin{aligned}\phi_{k+1} - \phi_k &\leq \nu(f(x_k + s_k) - f(x_k)) + (1 - \nu)(\gamma^3 - 1)\delta_k^3 \\ &\leq \nu((f(x_k + s_k) - f_k^s) + (f_k^s - f_k^0) + (f(x_k) - f_k^0)) + (1 - \nu)(\gamma^3 - 1)\delta_k^3 \\ &\leq \nu(|f(x_k + s_k) - f_k^s| + |f(x_k) - f_k^0|) + (1 - \nu)(\gamma^3 - 1)\delta_k^3.\end{aligned}\quad (79)$$

On unsuccessful steps (73) still applies, which means that the right-hand side of (79) dominates and (79) holds whether the iteration is successful or not. Note that here, in (d), we have not used the definition of Case 1.

Now we are ready to take the expectation of  $\Phi_{k+1} - \Phi_k$  in Case 1. Case (d) occurs with probability at most  $(1 - \alpha)(1 - \beta)$  and in that case  $\phi_{k+1} - \phi_k$  is bounded from above as in (79). Cases (a), (b) and (c) occur otherwise and in those cases  $\phi_{k+1} - \phi_k$  is bounded from above by  $b_1 < 0$ , with  $b_1$  defined in (73). Hence, we obtain

$$\begin{aligned}&\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\tau(X_k) \geq \zeta \Delta_k\}] \\ = &\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, I_k + J_k = 0] + \mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\tau(X_k) \geq \zeta \Delta_k\}, I_k + J_k > 0] \\ \leq &(1 - \alpha)(1 - \beta) (\nu \mathbb{E}[|f(x_k + s_k) - f_k^s| + |f(x_k) - f_k^0| | \mathcal{F}_{k-1}^{M.F}] + (1 - \nu)(\gamma^3 - 1) \mathbb{E}[\Delta_k^3 | \mathcal{F}_{k-1}^{M.F}]) \\ + &(1 - (1 - \alpha)(1 - \beta))(1 - \nu) \left(\frac{1}{\gamma^3} - 1\right) \mathbb{E}[\Delta_k^3 | \mathcal{F}_{k-1}^{M.F}]\end{aligned}$$

Recalling Assumption 4.3, noting that  $\mathbb{E}[\Delta_k^3 | \mathcal{F}_{k-1}^{M.F}] = \Delta_k^3$  and rearranging the terms we obtain

$$\begin{aligned}&\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\tau(X_k) \geq \zeta \Delta_k\}] \\ \leq &\left( (1 - \alpha)(1 - \beta) 2\nu\kappa_F + (1 - \nu) \left(\frac{1}{\gamma^3} - 1\right) + (1 - \alpha)(1 - \beta)(1 - \nu) \left(\gamma^3 - \frac{1}{\gamma^3}\right) \right) \Delta_k^3\end{aligned}$$

Choosing  $0 < \alpha \leq 1$  and  $0 < \beta \leq 1$  such that

$$(1 - \alpha)(1 - \beta) \leq \min \left\{ \frac{1}{2(\gamma^3 + 1)}, \frac{1 - \nu}{8\kappa_F\nu} \left(1 - \frac{1}{\gamma^3}\right) \right\}, \quad (80)$$

we conclude that

$$\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M.F}, \{\tau(X_k) \geq \zeta \Delta_k\}] \leq -\frac{1}{4}(1 - \nu) \left(1 - \frac{1}{\gamma^3}\right) \Delta_k^3. \quad (81)$$

**Case 2:**  $\tau(x_k) = \max\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\} < \zeta \delta_k$ , where  $\zeta$  is defined in (63).

i.  $J_k = 1$ , namely, we have good estimates but the model may be bad. This case follows similarly to Case 1(c), and (78) holds on successful steps. Thus decrease  $b_1$  can again be guaranteed for  $\Phi_k$  whether the iteration is successful or not.

ii.  $J_k = 0$ , namely, we have bad estimates and the model may be bad too. In this case, the situation of Case 1(d) may occur and both  $f$  and  $\delta_k$  may increase. Then we can upper bound the potential increase in  $\Phi_k$  by (79) on both successful and unsuccessful steps<sup>5</sup>.

Now we are ready to take the expectation of  $\Phi_{k+1} - \Phi_k$  in Case 2. Case 2(i) occurs with probability at least  $\beta$ , and then  $\phi_{k+1} - \phi_k$  is bounded above by  $b_1 < 0$ , with  $b_1$  defined in (73). Case 2(ii) happens with probability at most  $(1 - \beta)$ , and possible increase in  $\Phi_k$  as in (79). We obtain

$$\begin{aligned} & \mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M \cdot F}, \{\tau(X_k) < \zeta \Delta_k\}] \leq \\ & (1 - \beta) (\nu \mathbb{E}[|f(x_k + s_k) - f_k^s| + |f(x_k) - f_k^0| | \mathcal{F}_{k-1}^{M \cdot F}]) + (1 - \nu)(\gamma^3 - 1) \mathbb{E}[\Delta_k^3 | \mathcal{F}_{k-1}^{M \cdot F}] \\ & + \beta(1 - \nu) \left(\frac{1}{\gamma^3} - 1\right) \mathbb{E}[\Delta_k^3 | \mathcal{F}_{k-1}^{M \cdot F}] \end{aligned}$$

From Assumption 4.3, and  $\mathbb{E}[\Delta_k^3 | \mathcal{F}_{k-1}^{M \cdot F}] = \Delta_k^3$  we obtain

$$\begin{aligned} & \mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M \cdot F}, \{\tau(X_k) < \zeta \Delta_k\}] \leq \\ & \left\{ (1 - \beta)[2\nu\kappa_F + (1 - \nu)(\gamma^3 - 1)] + \beta(1 - \nu) \left(\frac{1}{\gamma^3} - 1\right) \right\} \Delta_k^3 \end{aligned} \quad (82)$$

Choosing  $\beta \in (0, 1]$  such that

$$\frac{\beta}{1 - \beta} \geq \frac{2\gamma^3[2\nu\kappa_F + (1 - \nu)(\gamma^3 - 1)]}{(1 - \nu)(\gamma^3 - 1)}, \quad (83)$$

we conclude that

$$\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M \cdot F}, \{\tau(X_k) < \zeta \Delta_k\}] \leq -\frac{1}{2}\beta(1 - \nu) \left(1 - \frac{1}{\gamma^3}\right) \Delta_k^3. \quad (84)$$

Thus, in conclusion, for  $\nu$  satisfying (71) and  $\alpha$  and  $\beta$  satisfying (80) and (83), the expected decrease in  $\Phi_k$  in (68) holds, with

$$\Theta = \frac{1}{4} \min\{2\beta, 1\} (1 - \nu) \left(1 - \frac{1}{\gamma^3}\right).$$

Now, let us particularize the constants as on page 26. Firstly, using  $\eta_2 \leq 18$  and  $\kappa_{bhm} = \kappa_{eh} + L$ , we deduce that  $\zeta := 20\kappa_{bhm} = 20(\kappa_{eh} + L)$  satisfies (63). Furthermore, letting  $C_1 = C_4 = \frac{1}{10}$  satisfies the conditions in Lemmas 4.3 and 4.6, and from Lemma 4.8 and particular choice of  $\epsilon_F$  we conclude,  $C_2 = \frac{1}{80}\eta_2 \min\{1, \eta_2\}$ . Thus, from (71) and  $\epsilon_F \leq \kappa_{eh} \leq \kappa_{bhm}$ ,  $\nu$  must satisfy

$$\frac{\nu}{1 - \nu} \geq \frac{8}{\min\{2\kappa_{bhm}, C_2\}} = \frac{320}{\eta_2 \min\{1, \eta_2\}}.$$

We let  $\nu = \frac{320}{320 + \eta_2 \min\{1, \eta_2\}} \in (0, 1)$ . Then (80) is equivalent to

$$(1 - \alpha)(1 - \beta) \leq \min \left\{ \frac{1}{18}, \frac{7\eta_2 \min\{1, \eta_2\}}{2^{11} \cdot 10\kappa_F} \right\}$$

---

<sup>5</sup>Note that under additional assumptions on  $\kappa_{ef}$  and  $\eta_2$ , one can further refine the analysis here and take into account the decrease in  $\Phi_k$  that could then be achieved when  $I_k = 1$

which is implied by (69). The bound (83) becomes

$$\frac{\beta}{1-\beta} \geq 16 \left[ \frac{640\kappa_F}{7\eta_2 \min\{1, \eta_2\}} + 1 \right],$$

which is implied by

$$\beta \geq \frac{2 \cdot 10^3 \kappa_F + 16\eta_2 \min\{1, \eta_2\}}{2 \cdot 10^3 \kappa_F + 17\eta_2 \min\{1, \eta_2\}}$$

which is implied by (70). The value of  $\Theta$  follows also using  $\eta_2 \min\{1, \eta_2\} \leq 18$ .  $\square$

Note the difference to first order results (Theorem 3.1): the effect of the stronger assumption on the estimates (Assumption 4.3 (a)) can be clearly seen in our results, in the presence of  $\kappa_F$  in the simplified bounds. Note that, due to the choice of constants and requirements on the accuracy of the estimates, the  $\eta_2$  terms are assumed to be smaller than terms involving Lipschitz constants  $\bar{L}$  and  $\kappa_{eh}$ , and hence, they remain present in the bounds. Given the definition of  $\eta_2$  in the algorithm, we can regard it as a means to control/ensure model quality.

The main complexity result for second order STORM follows.

**Theorem 4.2** (Complexity of second order STORM algorithm). *Consider Algorithm 2 and the corresponding stochastic process. Let  $T_\epsilon$  be defined as in (62) with  $\epsilon \in (0, 1]$ . Then, under the assumptions of Theorem 4.1, for sufficiently large  $\alpha \in (0, 1]$  and  $\beta \in (0, 1]$  with  $\alpha\beta > 1/2$ , we have*

$$\mathbb{E}[T_\epsilon] \leq \frac{\alpha\beta}{2\alpha\beta - 1} \left( \frac{\Phi_0 \zeta^3}{\Theta \epsilon^3} + 1 \right), \quad (85)$$

where  $\Phi_0$  is defined in (61) with  $k = 0$ ,  $\nu$  in (71) and  $\zeta$  in (63). Moreover, under the particular choice of constants described on page 26, and in Theorem 4.1, (85) becomes

$$\mathbb{E}[T_\epsilon] \leq 8 \cdot 10^3 \frac{\alpha\beta}{2\alpha\beta - 1} \left( \frac{\Phi_0(\kappa_{eh} + L)^3}{\Theta \epsilon^3} + 1 \right),$$

where  $\Theta \geq 6 \cdot 10^{-4} \eta_2 \min\{1, \eta_2\}$ .

*Proof.* The validity of the Assumption 2.1(iii) follows from Theorem 4.1, with  $h(\delta) = \delta^3$  and  $\Delta_\epsilon$  defined in (63). Then Lemma 4.9 and the discussion preceding it imply that Theorem 2.2 applies, which provides (85).  $\square$

The lim inf-type probability one convergence result trivially follows.

**Corollary 4.1** (Convergence of second order STORM algorithm). *Under conditions of Theorem 4.2 Algorithm 2 generates a subsequence convergent to a second order stationary point, almost surely.*

As in Section 3.5, similar set-ups that sub-sample function, gradient and Hessian estimates can be provided that satisfy Assumptions 4.2 and 4.3 for second order STORM [3].

## 5 Conclusion

We have proposed a general framework based on a stochastic process that can be used to bound expected complexity of optimization algorithms. This framework can be applied beyond the algorithms discussed in this paper and has already been used in a new work on stochastic line search [24]. We then applied this framework to establish that a stochastic trust region method, with dynamic stochastic estimates of the gradient, has essentially the same complexity as any other first order method in non convex setting. Similarly, given dynamic stochastic estimates of the gradient and Hessian the second order stochastic trust region method converges to second order stationary point and its expected complexity matches the deterministic case. While our algorithm requires the stochastic estimates to be progressively more accurate, it never requires to compute the full gradient, hence it applies in purely stochastic settings.

## References

- [1] Gerold Alsmeyer. Renewal, recurrence and regeneration, 2010.
- [2] A. Bandeira, K. Scheinberg, and L.N. Vicente. Convergence of trust-region methods based on probabilistic models. Technical report, Lehigh University, 2013.
- [3] Afonso S Bandeira, Katya Scheinberg, and Luis Nunes Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.
- [4] S. Billups, J. Larsson, and P Graf. Derivative-free optimization of expensive functions with computational error using weighted regression. 2011.
- [5] R. Byrd, G. M. Chin, J. Nocedal, and Y. Wu. Sample size selection in optimization methods for machine learning. 134:127–155, 2012.
- [6] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 2018.
- [7] K.H. Chang, M.K. Li, and H. Wan. Stochastic trust-region response-surface method (strong) - a new response-surface framework for simulation optimization. *INFORMS Journal on Computing*, 25(2):230–243, 2013.
- [8] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, Jun 2018.
- [9] A. R. Conn, N. I. M. Gould, and P. T. Toint. *Trust Region Methods*. MPS/SIAM Series on Optimization. SIAM, Philadelphia, 2000.
- [10] A.R. Conn, K. Scheinberg, and L.N. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.
- [11] Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and

- K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2933–2941. Curran Associates, Inc., 2014.
- [12] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.
- [13] M Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):1380–1405, 2012.
- [14] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- [15] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 2nd edition, 1989.
- [16] S Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Complexity and global rates of trust-region methods based on probabilistic models. *IMA Journal of Numerical Analysis*, 2017.
- [17] Chih jen Lin, Ruby C. Weng, S. Sathiya Keerthi, and Alexander Smola. Trust region newton method for large-scale logistic regression. In *In Proceedings of the 24th International Conference on Machine Learning (ICML, 2007)*.
- [18] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.
- [19] J. Larson and S.C. Billups. Stochastic derivative-free optimization using a trust region framework. *Computational Optimization and Applications.*, 2015.
- [20] Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.
- [21] Lam Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.
- [22] Lam Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *ArXiv e-prints*, 2017. <https://arxiv.org/abs/1705.07261>.
- [23] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, New York, NY, USA, 2nd edition, 2006.
- [24] Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *arXiv:1807.07994*, 2018.
- [25] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Stochastic variance reduction for nonconvex optimization. In *ICML*, pages 314–323, 2016.

- [26] S. Shashaani, F. S. Hashemi, and R. Pasupathy. Astro-df: A class of adaptive sampling trust-region algorithms for derivative-free simulation optimization. 2015.
- [27] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. *arXiv:1711.02838*, 2017.
- [28] Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, May 2015.
- [29] Peng Xu, Farbod Roosta-Khorasani, and Michael W. Mahoney. Newton-type methods for non-convex optimization under inexact hessian information. *arXiv*, 2017. arXiv:1708.07164.

## 6 Appendix

This Appendix contains proofs of several lemmas that are novel, but whose proofs are similar to existing results. We include them here for completeness.

**Proof of Lemma 4.2.** Using the optimal decrease condition (37), the upper bound on model Hessian from Lemma 4.1, and the fact that  $\|g_k\| \geq \kappa_{bhm}\delta_k$ , we have

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{scd}}{2}\|g_k\| \min \left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\} = \frac{\kappa_{scd}}{2}\|g_k\|\delta_k.$$

Since the model is  $\kappa$ -fully quadratic, the improvement in  $f$  achieved by  $s_k$  is

$$\begin{aligned} & f(x_k + s_k) - f(x_k) \\ &= f(x_k + s_k) - m(x_k + s_k) + m(x_k + s_k) - m(x_k) + m(x_k) - f(x_k) \\ &\leq 2\kappa_{ef}\delta_k^3 - \frac{\kappa_{scd}}{2}\|g_k\|\delta_k \leq -\frac{\kappa_{scd}}{4}\|g_k\|\delta_k, \end{aligned}$$

where the last inequality is implied by  $\delta_k^2 \leq \delta_k \leq \frac{\kappa_{scd}}{8\kappa_{ef}}\|g_k\|$ . □

**Proof of Lemma 4.3.** The definition of a  $\kappa$ -fully-quadratic model yields that

$$\|g_k\| \geq \|\nabla f(x)\| - \kappa_{eg}\delta_k^2.$$

Since condition (44) implies that  $\|\nabla f(x_k)\| \geq \max \left\{ \kappa_{bhm} + \kappa_{eg}, \frac{8\kappa_{ef}}{\kappa_{scd}} + \kappa_{eg} \right\} \delta_k$ , using  $\delta_k \leq 1$ , we have

$$\|g_k\| \geq \max \left\{ \kappa_{bhm}, \frac{8\kappa_{ef}}{\kappa_{scd}} \right\} \delta_k.$$

Hence, the conditions of Lemma 4.2 hold and we have

$$f(x_k + s_k) - f(x_k) \leq -\frac{\kappa_{scd}}{4}\|g_k\|\delta_k. \tag{86}$$

Since  $\|g_k\| \geq \|\nabla f(x)\| - \kappa_{eg}\delta_k$  in which  $\delta_k$  satisfies (44), we also have

$$\|g_k\| \geq \max \left\{ \frac{\kappa_{bhm}}{\kappa_{bhm} + \kappa_{eg}}, \frac{8\kappa_{ef}}{8\kappa_{ef} + \kappa_{scd}\kappa_{eg}} \right\} \|\nabla f(x_k)\|. \tag{87}$$

Combining (86) and (87) yields (45).  $\square$

**Proof of Lemma 4.4.** Since  $\delta_k \leq \frac{\|g_k\|}{\kappa_{bhm}}$ , the model decrease condition (37) and the uniform bound on  $H_k$  under Lemma 4.1 immediately yield that

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{scd}}{2} \|g_k\| \min \left\{ \frac{\|g_k\|}{\kappa_{bhm}}, \delta_k \right\} = \frac{\kappa_{scd}}{2} \|g_k\| \delta_k. \quad (88)$$

The model  $m_k$  being  $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic implies that

$$|f(x_k) - m_k(x_k)| \leq \kappa_{ef} \delta_k^3, \quad \text{and} \quad (89)$$

$$|f(x_k + s_k) - m_k(x_k + s_k)| \leq \kappa_{ef} \delta_k^3. \quad (90)$$

Since the estimates are  $\epsilon_F$ -s.o. accurate with  $\epsilon_F \leq \kappa_{ef}$ , we obtain

$$|f_k^0 - f(x_k)| \leq \kappa_{ef} \delta_k^3, \quad \text{and} \quad |f_k^s - f(x_k + s_k)| \leq \kappa_{ef} \delta_k^3. \quad (91)$$

We have

$$\begin{aligned} \rho_k &= \frac{f_k^0 - f_k^s}{m_k(x_k) - m_k(x_k + s_k)} \\ &= \frac{f_k^0 - f(x_k)}{m_k(x_k) - m_k(x_k + s_k)} + \frac{f(x_k) - m_k(x_k)}{m_k(x_k) - m_k(x_k + s_k)} + \frac{m_k(x_k) - m_k(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} \\ &\quad + \frac{m_k(x_k + s_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} + \frac{f(x_k + s_k) - f_k^s}{m_k(x_k) - m_k(x_k + s_k)}, \end{aligned}$$

which, combined with (88)-(91), implies

$$|\rho_k - 1| \leq \frac{8\kappa_{ef}\delta_k^3}{\kappa_{scd}\|g_k\|\delta_k} \leq 1 - \eta_1,$$

where we have used the assumptions  $\delta_k^2 \leq \delta_k \leq \frac{\kappa_{scd}(1-\eta_1)}{8\kappa_{ef}} \|g_k\|$  to deduce the last inequality.

Hence,  $\rho_k \geq \eta_1$ . Moreover, since  $\|g_k\| \geq \eta_2 \kappa_{bhm} \delta_k$ , then  $\tau_k^m \geq \min \left\{ \|g_k\|, \frac{\|g_k\|}{\kappa_{bhm}} \right\} \geq \eta_2 \delta_k$  and the  $k$ -th iteration is successful.  $\square$