

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Convergence Rate Analysis of a Stochastic Trust Region Method via Supermartingales

Jose Blanchet

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027,
jose.blanchet@columbia.edu

Coralia Cartis

Mathematical Institute, University of Oxford, Oxford, UK, cartis@maths.ox.ac.uk

Matt Menickelly

Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, 60439, mmenickelly@anl.gov

Katya Scheinberg

Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, 18015. katyas@lehigh.edu

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, we derive a bound on the expected stopping time of this process. We utilize this framework to analyze the expected global convergence rates of a stochastic variant of a traditional trust region method. 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 only 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 prior analysis, we show that the stochastic process defined by the trust-region method satisfies the assumptions of our proposed general framework. The stopping time in this setting is defined by an iterate satisfying a first-order accuracy condition. We demonstrate the first global complexity bound for a stochastic trust-region method, under the assumption of sufficiently accurate stochastic gradients. Finally, we apply the same framework to derive second-order complexity bounds under additional assumptions.

Key words: stochastic optimization, trust-region methods, stochastic processes, supermartingales, convergence rates

History: This paper was first submitted on March 1, 2018 and existed as a technical report since September 2016.

1. Introduction

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

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

We will assume $f(x)$ is a smooth function, bounded from below, and we will assume $f(x)$ can only be computed with some noise. Let $\tilde{f}(x, \xi)$ be the noisy computable version of f , where the noise ξ is a random variable. A common setting of stochastic optimization can be described by

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

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 (ML) domains. This is because objective functions of optimization problems arising from ML are typically sums of a (possibly) very large number of terms, each term being the loss function evaluated using one data example. These ML objectives can also be viewed as an expected loss which cannot be accurately computed; it can only be evaluated approximately, given a subset of data examples. 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, much 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 Ghadimi and Lan (2013) was the first to provide convergence rates guarantee of a sort for a randomized stochastic gradient method in a nonconvex setting. The analysis of this method, however, utilizes a carefully chosen step size and a randomized stopping scheme, which are quite different from what is used in practice.

From a practical perspective, SGD has a low per-iteration complexity and requires a high number of iterations, making it sequential and ineffective in a distributed setting. On the other extreme, each iteration of a full gradient method has a high per-iteration complexity and requires a low number of iterations, but can be efficiently distributed to reduce the overall wall-clock time. As an alternative to these extremes, several *variance reducing* stochastic methods have been proposed recently within the ML domain, including SAGA Defazio et al. (2014), SVRG Johnson and Zhang (2013) and SARAH Nguyen et al. (2017). These methods exploit the finite-sum structure of typical ML objectives; specifically, SVRG, which apparently has the fastest demonstrable convergence

rate in terms of number of data accesses of these three methods, requires the full gradient of the objective function to be computed on some (but not all) of the iterations. In this sense, SVRG (as well as SARAH) is a hybrid of an SGD method and a full gradient descent method. These hybrids do not easily fit into either computational extreme because the methods alternate between cheap sequential stochastic gradient computations and expensive distributable full gradient computations. For this reason, the superior theoretical computational complexity of SVRG does not necessarily reflect its practical performance. Moreover, the assumption of a dataset fixed prior to optimization - an assumption that underlies finite-sum optimization - conflicts with the ultimate goal of learning, which is to obtain a solution with good generalization performance. The method we describe in this paper is applicable to the *fully stochastic* setting, i.e., we do not assume a finite dataset fixed prior to optimization. Our method implicitly relies on variance reduction achievable simply by choosing adaptive sample sizes that (typically) increase 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 Byrd et al. (2012), Friedlander and Schmidt (2012).

With the rise of interest in nonconvex optimization, the ML community has begun to consider a classical alternative to gradient descent/line search methods - trust region methods Conn et al. (2000), Jen Lin et al. (2007), Dauphin et al. (2014). Their usefulness is largely dictated by their ability to utilize negative curvature in Hessian approximations, potentially escaping the neighborhoods of saddle points Dauphin et al. (2014), which can significantly slow down or even trap a line search method. It has been argued that while saddle points are undesirable, local minima are typically sufficient for the purposes of training certain nonconvex ML models, especially deep neural networks. Several recent works have proposed trust-region methods incorporating stochastic gradient and Hessian estimates Gratton et al. (2017), Xu et al. (2017), but these works assume that the objective function is deterministic. A trust-region method for our setting of stochastic optimization was proposed in Chang et al. (2013), and a more sophisticated adaptive sampling method was recently proposed in Shashaani et al. (2015). In both of these methods, convergence is achieved by repeatedly sampling the function values (and gradients, when applicable) so that the estimates are asymptotically error-free with probability 1. No convergence rates have been derived for these algorithms, likely because of these asymptotic concerns. Trust-region methods with adaptive sampling in a fully stochastic setting, such as may be used in ML contexts, have not yet been explored to our knowledge. We note that, additionally, our analysis applies to the setting where available function value and gradient estimates may be occasionally biased.

We will refer to the primary method analyzed in this paper as STORM (Stochastic Trust-region Optimization with Random Models). STORM was introduced in Chen et al. (2018) and the authors proved almost-sure convergence of STORM to a first-order stationary point. STORM is a stochastic

variance-reducing trust-region method, essentially a minor modification of a classical trust region framework. A similar method was analyzed in Larson and Billups (2015), under more restrictive conditions on $\tilde{f}(x, \xi)$. We believe that our convergence rate analysis framework can be applied to that method as well, but we choose to focus on STORM in this manuscript.

STORM uses adaptive trust-region radii and resembles what is known to be efficient in practice; hence, in this manuscript, we focus only on the theoretical analysis of STORM in both the first-order and second-order convergence regimes. We will demonstrate a convergence rate for STORM with a dependence on ϵ matching that of a deterministic trust-region method. Since STORM is randomized, our convergence rates are exhibited as bounds on the expected number of iterations the algorithm takes to achieve ϵ -accuracy. In contrast and as one example, the convergence rate results for SGD demonstrated in Ghadimi and Lan (2013) exhibit a bound on the expectation of the sum of the norms of all gradients encountered up to the T th iteration, as a function of T . Other weaker forms of convergence rates are established in Xu et al. (2017) and Tripuraneni et al. (2017). In Xu et al. (2017), trust-region and cubic regularization methods utilizing sampled Hessians are considered. The number of samples is selected in such a way that the error in the Hessian approximation is smaller than ϵ with overwhelming probability p . Then, a deterministic convergence rate is established under the assumption that this condition on the Hessian approximation holds in every iteration until ϵ -accuracy is reached. Thus, the established bound on the number of iterations T holds with probability p^T ; with probability $1 - p^T$, no bound is guaranteed. A similar flavor of complexity result is derived in Tripuraneni et al. (2017) for a cubic regularization method, where gradients and Hessians are sampled at a rate dictated by ϵ , and the resulting complexity bound holds only with some probability.

Algorithms in Xu et al. (2017) have some similarities with algorithms analyzed in Cartis and Scheinberg (2018) and Gratton et al. (2017). In Cartis and Scheinberg (2018) and Gratton et al. (2017), the global rates of convergence of trust-region, line-search, and adaptive cubic regularization methods are analyzed under the assumption that available first- and second-order information is inexact, but sufficiently accurate with some probability. However, in all of these works, the analysis relies heavily on the assumption that function values are computed exactly; in particular, the methods monotonically decrease the objective function. This implies that the results in Cartis and Scheinberg (2018) and Gratton et al. (2017) cannot be applied to a stochastic setting. This paper can be seen as an extension of Cartis and Scheinberg (2018) and Gratton et al. (2017) to a fully stochastic setting.

Unlike most of the literature on stochastic methods, we do not make the assumption that function, gradient or Hessian estimates are unbiased. Instead, it is assumed that in 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 formalize later, is very general and does not explicitly specify how such approximations must be obtained. In a setting where unbiased estimators are available, one can utilize sampling techniques described, for example, in Xu et al. (2017), Chen et al. (2018). In Chen et al. (2018), examples are provided where $\tilde{f}(x, \xi)$ is a biased estimator of $f(x)$, arbitrarily erroneous with some small fixed probability; yet, the required approximations can be constructed and the trust-region method still converges to a minimum. Note that because our condition on the approximations holds only with probability p , we provide complexity results in expectation, thus accounting for occasionally poor approximations.

The goal of our paper is twofold. First, we introduce a novel framework for bounding the expected complexity of a stochastic optimization method. This framework is based on defining a renewal-reward process associated with the algorithm and an associated stopping time - the time when the algorithm reaches a desired accuracy. Then, under certain assumptions, we derive a bound on the expectation of this stopping time. This framework, in principal, can be used in the analysis of convergence rates of a variety of algorithms. For instance, it applies to all of the algorithms in Cartis and Scheinberg (2018) and Gratton et al. (2017). In recent work Paquette and Scheinberg (2018), this framework has been applied to analyze a stochastic line-search method. In this paper, we use the new general framework to derive a bound on the convergence rate of STORM, by proving that STORM satisfies the framework’s assumptions. In particular, we show that the expected number of iterations required to achieve $\|\nabla f(x)\| \leq \epsilon$ is bounded in $O(\epsilon^{-2}/(2p - 1))$. This bound is an improvement on the result in Ghadimi and Lan (2013) and is similar to a bound in Reddi et al. (2016), Nguyen et al. (2017) in terms of dependence on ϵ ; however, our method never requires the computation of a “true” gradient. Our result is a natural extension of the best-known worst-case complexity of any first-order method for nonconvex optimization Nesterov (2004). In this paper, we also make a significant improvement upon the results in Chen et al. (2018) by relaxing a very restrictive condition on the size of the steps taken by STORM. By again applying our general analytic framework, we also provide a second-order complexity analysis for STORM. In particular, we show that a second-order STORM variant takes an expected number of iterations bounded in $O(\epsilon^{-3}/(2p - 1))$ to ensure $\max\{\|\nabla f(x)\|, -\lambda_{\min}(\nabla^2 f(x))\} \leq \epsilon$; this result requires slightly stronger assumptions on the accuracy of the function estimates, but provides a generalization of the results in Xu et al. (2017), Gratton et al. (2017) to the stochastic case.

Our main complexity results do not yet provide a termination criterion that would guarantee that $\|f(\bar{x})\| \leq \epsilon$, where \bar{x} is the last iterate of STORM. However, our analysis provides a foundation for establishing such a criterion. In particular, while we bound only the expected value of a stopping

time this paper, bounding the tail of the distribution of the stopping time would follow from the analysis here.

The rest of the paper is organized as follows. We begin by defining a framework for a stochastic process and we then derive a bound on its expected stopping time in Section 2. In Section 3, we provide the first-order complexity analysis of STORM by showing that it fits into the framework introduced in Section 2. We then perform a second-order complexity analysis in Section 4.

Notation Throughout the paper we use $\|\cdot\|$ to denote the Euclidian norm. Several constants are denoted by κ with a subscript meant to indicate the object that the given constant bounds. In particular, we use the 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 use $\mathbb{1}(A)$ to denote the indicator of a random event A occurring.

2. A Renewal-Reward Martingale Process

In this section, we consider a random process and an associated stopping time T . 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 manuscript to analyze the convergence rate of STORM and to argue that the framework presented in this section can be applied to the convergence analysis of a variety of stochastic algorithms. We start by providing the formal definition of a stopping time of a discrete time stochastic process.

DEFINITION 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 σ -field generated by X_1, \dots, X_m , for each $m \geq 0$.

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 $k \geq 0$. Let $\{W_k\}_{k=0}^\infty$ be a sequence defined on the same probability space as $\{(\Phi_k, \Delta_k)\}$ such that $W_0 = 1$ and

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

where \mathcal{F}_k is the σ -algebra generated by $\{(\Phi_0, \Delta_0, W_0), \dots, (\Phi_k, \Delta_k, W_k)\}$.¹ Note that due to (2), the W_k 's are mutually independent and are moreover independent of the sequence $\{(\Phi_j, \Delta_j)\}_{j=0}^{k-1}$.

¹ One can always enlarge the σ -algebras by adding sources of randomness which are independent from \mathcal{F}_k and consider such enlarged σ -algebras. In order to not add further notation, we prefer to work with \mathcal{F}_k as defined.

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

ASSUMPTION 1.

(i) There exist 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 $k \geq 0$:

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

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

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

or, equivalently,

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

In other words, Assumption 1 states that the nonnegative stochastic process Φ_k gets reduced by at least $\Theta h(\Delta_k)$ at each step, provided $T_\epsilon > k$. Also, Δ_k tends to increase whenever it is smaller than some threshold Δ_ϵ . 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, $\Delta_k \geq \Delta_\epsilon$ occurs frequently, and hence it occurs sufficiently frequently that $\mathbb{E}(\Phi_{k+1} - \Phi_k)$ can be bounded by a negative fixed value (dependent on ϵ). This will allow us to apply Wald's identity (stated momentarily) and hence derive a bound on $\mathbb{E}(T_\epsilon)$. To formalize this, we introduce a renewal process in which renewals occur at times k when $\Delta_k \geq \Delta_\epsilon$. We consider the sum of rewards V_j 's obtained between two renewals.

In order to define this renewal process, we first define an auxiliary process $\{Z_k\}_{k=0}^\infty$ by letting $Z_0 = j_\epsilon$ and setting

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

Note that the process $\{Z_k\}_{k=0}^\infty$ is a birth-death process on the set $\{k : k \leq j_\epsilon\}$. We then define the renewal process $\{A_n\}_{n=0}^\infty$ by letting $A_0 = 0$ and setting $A_n = \inf\{m > A_{n-1} : Z_m = j_\epsilon\}$. By (3) we have that

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

where we have used a simple inductive argument to obtain the second inequality. In other words, on $T_\epsilon > k$, the process A_n only counts the iterations for which $\Delta_k \geq \Delta_\epsilon$. The interarrival times of this renewal process are defined for all $n \geq 1$ by

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

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

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

That is, $N(k)$ counts 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 1. *Let τ_n be defined as above. Then, for all n*

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

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

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

On the other hand, by 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 τ_1 conditioned on $Z_1 = j_\epsilon - 1$ is the same as the distribution of $\bar{\tau}$. Thus, we simplify the above expression to conclude that

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

We now bound the expected number of renewals that occur before the time T_ϵ .

LEMMA 2.

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

For ease of notation, let $k \wedge T_\epsilon = \min\{k, T_\epsilon\}$. Consider the stochastic process defined by $R_0 = \Phi_0$ 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 Θ is from Assumption 1(iii). Observe that R_k is a non-negative supermartingale with respect to $\{\mathcal{F}_k\}$; to see this, we first write

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

Then,

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

where the last equality follows because T_ϵ is a stopping time and so the expectation of T_ϵ is \mathcal{F}_k -measurable.

Since $\{T_\epsilon \geq k+1\} = \{T_\epsilon > k\} = \{T_\epsilon \leq k\}^c \in \mathcal{F}_k$ we obtain

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

where we have used (5). Combining (6) and (7) we have that

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

as claimed. We then obtain, 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 even on the event $\{T_\epsilon = \infty\}$. Therefore, by the Monotone Convergence Theorem,

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

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

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

where 1 was added to $N(T_\epsilon - 1)$ in the last equality because $A_0 = 0$. Inserting this in (8),

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

which concludes the proof.

We now state and prove a well-known theorem concerning expected stopping time, known as Wald's Identity (e.g., see Theorem 2.2.4 in Alsmeyer (2010)). We provide a proof here because Wald's Identity is typically shown in the literature under the assumption that the stopping time is almost surely finite. Dropping this assumption is particularly important in our framework, as this assumption is equivalent to assuming that the optimization algorithm which generates the stochastic process in fact converges. It is convenient and useful not to have to prove the convergence result before establishing the convergence rates bounds, since convergence immediately follows from the existence of bounds on expected stopping time.

THEOREM 1. *Wald's Identity* 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 $\mathbb{E}(Y_i) = \mu_i \in [0, \infty]$ 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

$$\mathbb{E}(S_N) = \mathbb{E}(s_N).$$

Let $m > 0$ be an arbitrary integer and define $Y_i(m) = \min(Y_i, m)$, $N_m = \min(N, m)$, $\mu_i(m) = \mathbb{E}(Y_i(m))$, $S_n(m) = Y_1(m) + \dots + Y_n(m)$, 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

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

Because of monotonicity,

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

as $m \rightarrow \infty$. If $N = \infty$, we interpret $S_N = \sup_{n \geq 0} S_n$. Similarly,

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

as $m \rightarrow \infty$. By the Monotone Convergence Theorem we then conclude that

$$\mathbb{E}(S_N) = \mathbb{E}(s_N).$$

REMARK 1. If $\mu_i = \mu$, then $\mathbb{E}(S_N) = \mu \mathbb{E}(N)$. If $\mu = 0$, then $Y_i = 0$ almost surely and $S_N = 0$. Therefore, if $\mu = 0$, we interpret $\mu \mathbb{E}(N) = 0$, even if $\mathbb{E}(N) = \infty$. This interpretation is also consistent with the case in which $N = 0$ almost surely - in this case $0 = \mu \mathbb{E}(N) = \mathbb{E}(S_N)$, even if $\mu = \infty$.

We now apply Wald's Identity to $S_n = A_n = \sum_{i=0}^n \tau_i$ to obtain the main result of this section.

THEOREM 2. *Let Assumption 1 hold. Then*

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

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 the random variable $N(T_\epsilon - 1) + 1$ is a stopping time with respect to $\{\mathcal{G}_n\}_{n \geq 0}$. To see this, note that because $N(k) \leq k$, we have the equality of events

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

The inclusion follows because $N(k) + 1$ is a stopping time with respect to $\{\mathcal{F}_{A_n}\}_{n \geq 0}$ ($A_n \geq n$ implies $\mathcal{F}_n \subseteq \mathcal{F}_{A_n}$), and T_ϵ is a stopping time with respect to $\{\mathcal{G}_n\}_{n \geq 0}$ by construction.

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 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 obtain from Lemmas 1 and 2 that

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

The statement of the theorem follows from the last inequality.

3. The first-order STORM algorithm

We now state and analyze a stochastic trust-region (TR) algorithm (Algorithm 1), which is very similar to its deterministic counterpart Conn et al. (2000). Algorithm 1 uses inexact (noisy) information about f and its derivatives, just as the commonly stated deterministic method uses exact information. Algorithm 1 and the assumptions on its steps that we will impose below are intended to yield convergence to a first-order stationary point. In this section, we analyze the global rate of convergence of Algorithm 1 to such a point, while in Section 4, we will extend Algorithm 1 to yield convergence to second-order critical points.

Algorithm 1 STOCHASTIC DFO WITH RANDOM MODELS, CHEN ET AL. (2018)

- 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, that is,

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

for some constant $\kappa_{fcd} \in (0, 1]$. This condition is standard in the analysis of TR methods, easy to enforce in practice, and is discussed in detail in the literature Conn et al. (2000), Nocedal and Wright (2006). Iterations k in which $x_{k+1} = x_k + s_k$ are called *successful*.

Algorithm 1 generates a random process. Randomness stems from the random models and random estimates constructed on each iteration, which in turn are based on random information obtained from the stochastic function $f(x, \xi)$. M_k will denote a random model in the k th iteration and we 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 Δ_k , and the steps S_k are also random quantities; correspondingly, $x_k = X_k(\omega)$, $\delta_k = \Delta_k(\omega)$, and $s_k = S_k(\omega)$ will denote their respective realizations. Moreover, we let random variables $\{F_k^0, F_k^s\}$ denote respective estimates of $f(X_k)$ and $f(X_k + S_k)$, and we denote their realizations by $f_k^0 = F_k^0(\omega)$ and $f_k^s = F_k^s(\omega)$. Hence, 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 $\{F_k\} \triangleq \{(F_k^0, F_k^s)\}$, the resulting stochastic process has a desirable convergence rate. These conditions require that the models M_k and estimates (F_k^0, F_k^s) are sufficiently accurate with sufficiently high probability, conditioned on the past.

In the analysis of a deterministic TR method, the function value $f(x)$ never increases after an iteration; the main challenge of the analysis of Algorithm 1 lies in the fact that this monotonic

property is certainly not guaranteed in the presence of stochasticity. The key to our analysis lies in the assumption that accuracy improves in coordination with the perceived progress of the algorithm. Our analysis is based on properties of supermartingales - in particular, supermartingales for which the increments depend on the change in function value between iterations, 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. Thus, we make the following assumptions on f :

ASSUMPTION 2. *Over all iterates x_k generated by Algorithm 1, the gradient $\nabla f(\cdot)$ is L -Lipschitz continuous and*

$$f(x_k) \geq 0$$

The assumptions of Lipschitz continuity of ∇f and boundedness of f from below are standard. For simplicity and without loss of generality, we assume that the lower bound on $f(\cdot)$ is nonnegative.

3.1. Assumptions on the first-order STORM algorithm

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

DEFINITION 2. 1) A function m_k is a κ -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 ϵ_F -accurate estimates of $f(x_k)$ and $f(x_k + s_k)$, respectively, for a given δ_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. A sequence of random models $\{M_k\}$ is said to be α -probabilistically κ -fully linear with respect to the corresponding sequence $\{B(X_k, \Delta_k)\}$ if the events

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

satisfy the condition

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

DEFINITION 4. A sequence of random estimates $\{F_k\}$ is said to be β -probabilistically ϵ_F -accurate with respect to the corresponding sequence $\{X_k, \Delta_k, S_k\}$ if the events

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

satisfy the condition

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

where ϵ_F is a fixed constant.

We can now state our key assumption on the nature of the stochastic (and deterministic) information used by our algorithm.

ASSUMPTION 3. *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 α -probabilistically κ -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 β -probabilistically ϵ_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 $\beta \in (0, 1)$.*

We comment on what is meant by “sufficiently large” in Assumption 3. Under Assumption 3, $P\{I_k J_k = 1 | \mathcal{F}_{k-1}^{M \cdot F}\} \geq \alpha\beta$ and $P\{I_k + J_k = 0 | \mathcal{F}_{k-1}^{M \cdot F}\} \leq (1 - \alpha)(1 - \beta)$. In iteration k , if $I_k J_k = 1$, then the algorithm behaves like an (inexact) deterministic algorithm in that iteration. In the other extreme, if $I_k + J_k = 0$, then not only may Algorithm 1 produce a *bad step* (a step in which the objective function value increases), but Algorithm 1 may accept this bad step by mistaking it for an *improving step* (a step that decreases the function value). In the remaining two cases in which exactly one of $I_k = 0$ or $J_k = 0$ holds, then either the model is good but the estimates are faulty, or the estimates are good and the model is faulty. In either case, an improving step is still possible, but a bad step is impossible. In the worst case, no step is taken and the trust region radius is reduced. The main idea of our framework is to choose the probabilities of the occurrence of $I_k J_k = 1$ and $I_k + J_k = 0$ according to the possible corresponding increase or decrease in $f(x)$ so that, in expectation, $f(x)$ is sufficiently decreased; this is achieved by selecting α and β “sufficiently large”. An important observation is that α and β do not have to increase as the algorithm progresses; with the same constant - but sufficiently small - probabilities, our models and estimates can be arbitrarily erroneous.

REMARK 2. In Chen et al. (2018), the analysis of Algorithm 1 required an additional assumption that $\eta_2 \geq \kappa_{ef}$; for simplicity, it was 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 manuscript, we improve on the analysis of Chen et al. (2018) and drop this assumption, allowing η_2 to be set to a small value. Note that small values of η_2 imply small values of ϵ_F due to Assumption 3(c), representing a potential trade-off in the selection of η_2 . In one extreme, this relationship indicates that if $\epsilon_F = 0$ (that is, there is no error in the function value estimates), then η_2 can be selected arbitrarily small.

3.2. Useful existing results

Algorithm 1 was analyzed in Chen et al. (2018) and it was demonstrated that there exists a selection of α and β such that under Assumption 3 (and the additional assumption that $\eta_2 \geq \kappa_{ef}$, see Remark 2), the sequence of random iterates $\{X_k\}$ generated by Algorithm 1 almost surely satisfies $\lim_{k \rightarrow \infty} \|\nabla f(X_k)\| = 0$. This is an almost sure first-order convergence result.

Our primary goal in this manuscript is to bound the expected number of steps taken by Algorithm 1 before $\|\nabla f(X_k)\| \leq \epsilon$ occurs. Our secondary goal, as mentioned in Remark 2 is to relax the assumption $\eta_2 \geq \kappa_{ef}$. We will modify the analysis that led to the above stationarity result in Chen et al. (2018). First, we state (without proof) several auxiliary lemmas from Chen et al. (2018).

LEMMA 3. [Good model \Rightarrow function value reduction $\propto \|g_k\|$] 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 4. [Good model \Rightarrow function value reduction $\propto \|\nabla f(x_k)\|$] Under Assumption 3(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 5. [Good model + good estimates \Rightarrow successful step] Under Assumption 3(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 ϵ_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 6. [Good estimates + successful step \Rightarrow function value reduction $\propto \delta_k^2$] Under Assumption 3(a), suppose that the estimates $\{f_k^0, f_k^s\}$ are ϵ_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 like

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

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 β . First, note that the constants κ_{ef}, κ_{eg} , and κ_{bhm} can be chosen arbitrarily large, but should be ideally chosen as small as possible while guaranteeing Assumption 3. Let us assume that κ_{ef}, κ_{eg} and κ_{bhm} can be all chosen in $\Theta(L)^2$, where L is the Lipschitz constant of $\nabla f(x)$ from Assumption 2. We acknowledge that L is generally not explicitly known; see Conn et al. (2009) for a discussion of the construction of fully-linear models in the case of unavailable derivative estimates. Once these constants are chosen, ϵ_F is chosen to satisfy the conditions in Assumption 3(c). Note that if η_2 is chosen in $\Theta(L)$, then Algorithm 1 only takes steps when (roughly) $\delta_k \leq \frac{\|g_k\|}{L}$; this is similar to restricting step sizes to $\frac{1}{L}$ in a gradient descent method. With these choices for constants, we see from Assumption 3(c) that the estimates need to be slightly more accurate than the models, but the order of required accuracy is similar (in $\Theta(L)$, but with a tighter constant). However, a choice of $\eta_2 \in \Theta(L)$ may not be desirable - in trust-region methods, step sizes are meant to be chosen adaptively; hence, it is desirable to allow larger steps, which is done by setting η_2 as small as possible. But via Assumption 3(c), this requires a proportionally small selection of ϵ_F ; that is, the function value estimates will have to be *substantially* more accurate than the models. Yet another trade-off in choosing a value of η_2 will become apparent in our main complexity results; we will see that the expected improvement per iteration may depend on η_2 . However, selecting “reasonable” values of η_2 will remove 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 frequently-used values for these constants in practice. We also assume that $\kappa_{bhm} \leq 12\kappa_{ef}$ and $\eta_2 \leq \kappa_{eg}$. To simplify expressions further we will suppose $\kappa_{ef} = \kappa_{eg}$. Clearly, if κ_{ef} or κ_{eg} happen to be smaller (that is, the models give better approximations of the true function), then bounds

² Note that it is possible to have κ_{ef} and κ_{eg} of different magnitudes. In particular, when κ_{eg} is small, we obtain correspondingly accurate gradients, but κ_{ef} remains in $\Theta(L)$. Our analysis and results apply then as well.

somewhat better than the ones derived here can be derived. We are interested in deriving bounds in the pessimistic case in which κ_{ef} or κ_{eg} may be large. We note that our analysis can be performed for any other values of the above constants; we again stress that these choices of constants have been made entirely for convenience and simplicity.

The conditions on α and β 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 Δ_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 constant, sufficiently close to 1, to be defined later. Clearly $\Phi_k \geq 0$. We simplify the notation $\mathcal{F}_k^{M \cdot F}$ to \mathcal{F}_k .

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_ϵ is a stopping time for the stochastic process defined by Algorithm 1, and is hence a stopping time 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 1 is satisfied for $\{\Phi_k, \Delta_k\}$, allowing us to apply the results of Section 2.

So, we show that Assumptions 1(i)-(ii) hold with the following choice of Δ_ϵ :

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

Note that, by our choice of algorithmic parameters, (22) is satisfied by $\zeta = 20\kappa_{eg}$.

For simplicity of 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 ζ within a factor of γ of its lower bound in (22). It follows that for any k , $\Delta_k = \gamma^{i_k} \Delta_\epsilon$ for some integer i_k . Choose λ in Assumption 1(i)-(ii) so that $e^\lambda = \gamma$. Assumption 1(i) then holds automatically due to the definition of $\{\Phi_k, \Delta_k\}$ and the choice of δ_{max} imposed by Algorithm 1. For Assumption 1(ii) to hold, we need to show that the dynamics (3) hold for Δ_k , which we do in the following lemma.

LEMMA 7. *Let Assumptions 2 and 3 hold. Let α and β be such that $\alpha\beta > 1/2$. Then Assumption 1(ii) is satisfied for $W_k = 2(I_k J_k - \frac{1}{2})$, $\lambda = \log(\gamma)$, and $p = \alpha\beta$.*

Clearly, inequality (3) holds when $\mathbb{1}(T_\epsilon > k) = 0$. We will show that, conditioned on $T_\epsilon > k$ (i.e. $\mathbb{1}(T_\epsilon > k) = 1$) 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 in which $\delta_k > \Delta_\epsilon$, we have $\delta_k \geq \gamma \Delta_\epsilon$ and hence $\delta_{k+1} \geq \Delta_\epsilon$. Now, suppose that $\delta_k \leq \Delta_\epsilon$. Then, because $T_\epsilon > k$, we have $\|\nabla f(x_k)\| > \epsilon$ and hence from the definition of ζ , we have 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 $J_k = 1$, that is, both the model and the estimates are good in the k th iteration. Because the model m_k is κ -fully linear,

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

Moreover, because the estimates $\{f_k^0, f_k^s\}$ are ϵ_F -accurate with $\epsilon_F \leq \kappa_{ef}$, we conclude that condition (17) in Lemma 5 holds. Thus, the k th iteration is successful, that is, $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \max\{\delta_{max}, \gamma\delta_k\}$.

If $I_k J_k = 0$, then $\delta_{k+1} \geq \gamma^{-1}\delta_k$ 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 that Assumption 1(ii) holds.

We now show that Assumption 1(iii) holds; this is the key theorem in this section and is similar to Theorem 4.11 in Chen et al. (2018), except we drop the restrictive conditions imposed on η_2 mentioned in Remark 2 and simplify the proof. We will omit the parts of the proof that are identical to those of Theorem 4.11 in Chen et al. (2018).

THEOREM 3. *There exist probabilities α and β such that, if Assumptions 2 and 3 hold with these α and β , then there exists a constant $\Theta > 0$ such that*

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

conditioned on $T_\epsilon > k$.

Moreover, under our particular choice of constants, let α and β satisfy

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

and

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

Then³, $\Theta = \frac{1}{1800} \min \{\eta_2\beta, \kappa_{eg}^{-1}\}$.

³ Note that because $\beta > \frac{1}{2}$ always, if $\eta_2 \geq 2\kappa_{eg}^{-1}$, then $\Theta = \frac{1}{1800}\kappa_{eg}^{-1}$ independently of η_2 . This implies that small values for η_2 are permissible if the value of κ_{eg} is large.

Since (24) holds trivially if $T_\epsilon \leq k$, we assume in this proof that $\|\nabla f(X_k)\| > \epsilon$. We split the analysis into 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 (24) 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 4.

Let x_k , δ_k , s_k , g_k , and ϕ_k denote realizations of random quantities X_k , Δ_k , S_k , G_k , and Φ_k , respectively. Consider an arbitrary 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 $\delta_{k+1} = \frac{1}{\gamma} \delta_k$, that is,

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

Case 1: $\|\nabla f(x_k)\| \geq \zeta \delta_k$ with ζ satisfying (22). Let α and β 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 4 and $C_3 = 1 + \frac{3L}{2\zeta}$. We consider four subcases:

a. $I_k = 1$ and $J_k = 1$, i.e., both the model and the estimates are good in the k th iteration. The proof is almost identical to that of Theorem 4.11 in Chen et al. (2018). However, because we do not assume that $\eta_2 \geq \kappa_{bhm}$, there is a slight modification due to our definition of ζ .

Because Lemma 4 and Lemma 5 hold, we have that

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

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 in the k th iteration. The proof is identical to that of Theorem 4.11 in Chen et al. (2018), 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 Chen et al. (2018).

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 Chen et al. (2018) 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 Chen et al. (2018) we combine the outcomes of the four subcases 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 the last inequality is because $\|\nabla f(X_k)\| \geq \zeta \Delta_k$.

We now derive bounds on the expectation of $\Phi_{k+1} - \Phi_k$ in the remaining case. This proof differs from the analysis of Case 2 performed in Theorem 4.11 of Chen et al. (2018) due to our weaker assumptions on η_2 .

Case 2: $\|\nabla f(x_k)\| < \zeta \delta_k$ with ζ satisfying (22). Note that if $\|g_k\| < \eta_2 \delta_k$, then the k th iteration is unsuccessful and (27) holds. Hence, we assume that $\|g_k\| \geq \eta_2 \delta_k$. We consider only two subcases: in the first subcase, we show that if the function value estimates are good, then (27) holds. In the second subcase, because $\|\nabla f(x_k)\| < \zeta \delta_k$, the increase in ϕ_k can be bounded from above by a multiple of δ_k^2 . Thus, by selecting an appropriate value for the probability β , we establish the same bound on expected decrease in Φ_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, δ_k is reduced. Applying the same argument as in the Case 1(c), we have that (27) 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 ϕ_k . Using the Taylor expansion of f about x_k , the Lipschitz continuity of $\nabla f(x)$ and taking into account the bound $\|\nabla f(x_k)\| < \zeta \delta_k$ we have

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

Thus, the change in ϕ is bounded like

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

We are now ready to bound the expectation of $\phi_{k+1} - \phi_k$, as we did in Case 1. In this Case 2, however, we only need to combine (31), which holds with probability at most $(1 - \beta)$, with (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 $\beta \in (0, 1]$ so that

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

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

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

To complete the proof of the theorem, it remains to substitute the appropriate constants into 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 α and β 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 $\nu > \frac{1}{2}$ in (31) implies that

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

Case 2: Recalling the expression for C_3 , recalling the values for the constants ζ and $\gamma = 2$, and choosing ν 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 ν is chosen so that $1-\nu = \frac{\eta_2}{320+\eta_2}$, from (34) and $\eta_2 < 320$ (since $\nu > \frac{1}{2}$),

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

Thus 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.

Our almost-sure stationarity result follows immediately from Theorem 3 with the same proof as given in Chen et al. (2018); however, we do not assume that $\eta_2 \geq \kappa_{ef}$:

THEOREM 4. *Let Assumptions 2 and 3 hold. Let α and β satisfy the conditions of Theorem 3. Then, the sequence of random iterates $\{X_k\}$ generated by Algorithm 1 almost surely satisfies*

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

Using Theorem 3, we can moreover demonstrate the validity of Assumption 1(iii), which is more directly related to the primary goal in this manuscript. We state the result below for completeness and convenience of reference.

LEMMA 8. *Let the assumptions of Theorem 3 hold. Then Assumption 1(iii) is satisfied, with $\Theta = \frac{1}{1800} \min \{\eta_2 \beta, \kappa_{eg}^{-1}\}$ for the process $\{\Phi_k, \Delta_k\}$, where Φ_k is defined as in (20) with ν satisfying (25) and $h(\delta) = \delta^2$.*

3.4. Complexity result for first-order STORM algorithm

We immediately arrive at the following theorem.

THEOREM 5. *Consider Algorithm 1 and the corresponding stochastic process. Let T_ϵ be defined as in (21). Then, under the assumptions of Theorem 3,*

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

where $\Theta = \frac{1}{1800} \min \{\eta_2 \beta, \kappa_{eg}^{-1}\}$, Φ_0 defined as in (20) with $k = 0$, with ν satisfying (25).

3.5. Example of models and estimates satisfying Assumption 3

Although Assumption 3 was sufficiently general to allow us to develop our general complexity analysis, Assumption 3 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 to produce sufficiently accurate gradient and Hessian approximations. For example, Lemma 4 in Tripuraneni et al. (2017) uses matrix concentration results from Tropp (2015) to show that given a 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 $\nabla \tilde{f}(x, \xi)$ (denoted by g) satisfies

$$\|g - \nabla f(x)\| \leq \epsilon$$

with probability p , where $\tilde{\mathcal{O}}$ hides a term dependent on $-\log(1-p)$. A similar result was established for the Hessian sample average approximation. Another similar result for the function estimates, given variance σ_f , is a simpler version of the same inequalities and can be derived using Chebyshev's inequality.

Using these results, we can obtain α -probabilistically fully-linear models as follows. We compute f_k by averaging $\mathcal{O}(\frac{\sigma_f^2}{\Delta_k^4} \log(\frac{1}{1-\sqrt{\alpha}}))$ samples $\tilde{f}(x_k, \xi)$, and we independently 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 α . The fully-linear condition $|m(y) - f(y)| \leq \kappa_{ef} \Delta_k^2$ follows automatically with an appropriately chosen κ_{ef} . Note that all of these sample sizes are determined by quantities that are either known by the algorithm or can be accurately estimated.

Similarly, we can obtain β -probabilistically ϵ_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 $\tilde{f}(x_k + s_k, \xi)$, respectively.

In the case of simulation optimization, when $\nabla f(x, \xi)$ is not available, κ -fully-linear models m_k can be constructed via polynomial interpolation Conn et al. (2009); α -probabilistically κ -fully-linear models can be similarly obtained by combining interpolation and sufficiently accurate function value estimates (see, e.g. Shashaani et al. (2015)).

Another setting that is explored in Chen et al. (2018) is when $f(x)$ (and possibly, $\nabla \tilde{f}(x)$) are computed accurately via some procedure, but this procedure 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, then STORM still converges almost surely.

4. The second-order STORM algorithm

We now introduce a variant of Algorithm 1 that is intended to achieve second-order criticality in the fully stochastic setting; we use the same notation as in Algorithm 1.

In this second-order setting, the putative subproblem solution generally needs to provide more than just Cauchy decrease (9). In particular, we require that in the k th iteration, for all realizations m_k (as defined in Step 2 of Algorithm 1) of M_k , we are able to compute a step s_k satisfying

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

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 model⁴ Conn et al. (2000).

Our analysis (but not the algorithm itself) will use, in lieu of ∇f , the following measure of proximity to a second order stationary point for the objective f :

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

The corresponding optimality measure for the model m_k , following Bandeira et al. (2014), is defined analogously as

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

The additional term in (39) not present in (38) is necessary because there is no upper bound on the model Hessians on all iterations, as in the first-order case (κ_{bhm}). We will only ever apply (39) to the iterate x_k , in which case (39) 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, a modification of the first-order STORM algorithm.

Algorithm 2 SECOND ORDER STOCHASTIC DFO WITH RANDOM MODELS

Like Algorithm 1, but 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 s_k 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 set $\delta_{k+1} = \gamma^{-1} \delta_k$; $k \leftarrow k + 1$ and go to step 2.

The analysis for Algorithm 2 variant will again use the framework proposed in Section 2, thus serving as another illustration of the applicability of this generic framework. Before proceeding, we need to describe the additional assumptions required for our second-order analysis. In particular, we will need to assume one more order of smoothness than was assumed in Assumption 2:

ASSUMPTION 4. *The function f satisfies Assumption 2 and f is twice continuously differentiable. The Hessian $\nabla^2 f$ is L_H -Lipschitz continuous.*

⁴ 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 .

4.1. Assumptions on the second order STORM algorithm

We introduce a measure of second-order accuracy of the models m_k (see Conn et al. (2009), Billups et al. (2011), Larson and Billups (2015) for more details).

DEFINITION 5. 1) A function m_k is a κ -fully quadratic model of f on $B(x_k, \delta_k)$ provided, for $\kappa = (\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ and $\forall y \in B(x_k, \delta_k)$,

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

2) The estimates f_k^0 and f_k^s are ϵ_F -s.o.-accurate (s.o. for “second-order”) estimates of $f(x_k)$ and $f(x_k + s_k)$, respectively, for a given δ_k provided

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

DEFINITION 6. A sequence of random models $\{M_k\}$ is said to be α -**probabilistically** κ -**fully quadratic** (see Bandeira et al. (2013)) 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 7. A sequence of random estimates $\{F_k^0, F_k^s\}$ is said to be β -**probabilistically** ϵ_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 ϵ_F is a fixed constant.

We no longer utilize Assumption 3(a), i.e., we no longer explicitly assume that model Hessians H_k are bounded in norm. Instead, we demonstrate in the following lemma that $\|H_k\|$ is uniformly bounded from above as a direct consequence of m_k being a fully quadratic model of f :

LEMMA 9. Bandeira et al. (2014) Let Assumption 4 hold. Given constants κ_{eh} , κ_{eg} , κ_{ef} , and δ_{\max} , there exists a constant $\kappa_{bhm} \geq 1$ such that, uniformly over every k and every realization m_k of M_k such that m_k is $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ -fully quadratic model of f on $B(x_k, \delta_k)$ and $\delta_k \leq \delta_{\max}$, we have

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

The proof follows trivially from the definition of fully quadratic models and the fact that $\|\nabla^2 f\| \leq L$; this follows. For our convergence analysis, we again need to impose conditions on the stochastic (and deterministic) information used by STORM:

ASSUMPTION 5. *Within Algorithm 2:*

- (a) *The sequence of random models $\{M_k\}$ generated by Algorithm 2 is α -probabilistically κ -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 β -probabilistically ϵ_F -s.o. accurate for $\epsilon_F < \min\{\kappa_{ef}, \frac{1}{4}\eta_1\eta_2\kappa_{scd} \min\{\eta_2, 1\}\}$ and sufficiently large $\beta \in (0, 1)$.*

Note that, as in our first-order analysis, we allow for unrestricted values of η_2 in Algorithm 2, involving a potential trade-off with an increased accuracy requirement on the function estimates.

4.2. Useful preliminary results for second order STORM analysis

The analysis of Algorithm 2 is similar to the analysis of Algorithm 1. However, there are more cases to consider and the convergence rate to the second-order stationary point is different, as one would expect from the second-order convergence analysis of a deterministic TR method. There is one more significant difference, an additional assumption on function estimates, to be detailed in the next section. First, we state and prove analogues of Lemmas 3–6 for function decrease in terms of first- and second-order optimality. The first three lemmas are almost identical to Lemmas 3–5, with the notable exceptions that 1) the models are assumed to be fully quadratic instead of fully linear, 2) the model decrease condition (37) is used, and 3) the condition $\|H_k\| \leq \kappa_{bhm}$ is only valid in iterations k for which the model m_k is fully-quadratic (as seen in Lemma 9). For completeness, we have included the proofs of Lemmas 10–12 in the Appendix.

LEMMA 10. [Good quadratic model \Rightarrow function reduction $\propto \|g_k\|$] *Let Assumption 4 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 11. [Good quadratic model \Rightarrow function reduction $\propto \|\nabla f(x_k)\|$] *Let Assumption 4 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 12. [Good quadratic model + good s.o. estimates \Rightarrow successful step] *Let Assumption 4 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 ϵ_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 negative curvature in the model and second-order accurate estimates.

LEMMA 13. [Good quadratic model \Rightarrow function reduction $\propto \lambda_{\min}(H_k)$] *Let Assumption 4 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

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

Whenever $\lambda_{\min}(H_k) < 0$, the 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 κ -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).

LEMMA 14. [Good quadratic model \Rightarrow function reduction $\propto \lambda_{\min}(\nabla^2 f(x_k))$] *Let Assumption 4 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 $C_4 \leq \frac{\kappa_{scd}}{4} \cdot \frac{8\kappa_{ef}}{8\kappa_{ef} + \kappa_{scd}\kappa_{eh}}$.

Using Corollary 8.5.6 from Golub and Loan (1989), the definition of a κ -fully-quadratic model implies that

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

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

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

Thus, the conditions of Lemma 13 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).

LEMMA 15. [Good quadratic model + good s.o. estimates \Rightarrow successful step] *Let Assumption 4 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 ϵ_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.

From the model decrease condition (37),

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

From the definition of the model m_k being $(\kappa_{ef}, \kappa_{eg})$ -fully quadratic,

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

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

Since the estimates are ϵ_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. Thus, $\rho_k \geq \eta_1$. Moreover, the first term in (54) and (40) imply $\tau_k^m \geq (-\lambda_{\min}(H_k)) \geq \eta_2\delta_k$. The conclusion of the lemma follows.

LEMMA 16. [Good s.o. estimates + successful step \Rightarrow function decrease $\propto \delta_k^3$] Suppose the estimates $\{f_k^0, f_k^s\}$ are ϵ_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 (the k th iteration is successful), then the improvement in f is bounded below like

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

where

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

If the k th iteration is successful, then $\rho \geq \eta_1$ and either $\min\{\|g_k\|, \frac{\|g_k\|}{\|H_k\|}\} \geq \eta_2\delta_k$ or $-\lambda_{\min}(H_k) \geq \eta_2\delta_k$. Let us first suppose that $\min\{\|g_k\|, \frac{\|g_k\|}{\|H_k\|}\} \geq \eta_2\delta_k$. 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 used the supposition $\delta_k \leq 1$.

Let us now suppose that $-\lambda_{\min}(H_k) \geq \eta_2\delta_k$. Then,

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

Thus, in either case, using the fact that the estimates are ϵ_F -s.o. accurate, we have

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

where C_2 is defined in (60).

Choosing constants To simplify our calculations, just as we did in our first-order analysis, 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 5, 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 κ_{bhm} , as such a bound cannot be chosen freely; as seen in Lemma 9, we have $\kappa_{bhm} = \kappa_{eh} + L \leq 2 \max\{\kappa_{eh}, \bar{L}\}$.

4.3. Defining and analyzing the process $\{\Phi_k, \Delta_k\}$ for second-order convergence

Seeing as how the order of the function decrease that can be guaranteed on good iterations of Algorithm 2 changed from δ_k^2 in the first-order analysis (c.f. Lemma 3) to δ_k^3 in the second-order analysis (c.f. Lemma 16), we must modify the process Φ_k accordingly. We let $\{\Phi_k, \Delta_k\}$ be derived from the process generated by Algorithm 2, where once again Δ_k denotes the trust-region radius, but this time we let

$$\Phi_k = \nu f(X_k) + (1 - \nu) \Delta_k^3, \quad (61)$$

where $\nu \in (0, 1)$ is a deterministic constant sufficiently close to 1, which we will define later. Once again, it is clear that $\Phi_k \geq 0$. We define a random time

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

which is a stopping time for the stochastic process defined by Algorithm 2, and is hence also a stopping time for $\{\Phi_k, \Delta_k\}$. To bound the expected stopping time $\mathbb{E}(T_\epsilon)$ for Algorithm 2, we show that Assumption 1 is satisfied for $\{\Phi_k, \Delta_k\}$ and apply the results of Section 2.

Similarly to our first-order analysis, we can show that Assumption 1(i)–(ii) holds with $\lambda = \log \gamma$, and with the 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 $\epsilon \in (0, 1]$. We reuse our assumption from the first-order analysis that $\Delta_\epsilon = \gamma^i \delta_0$ for some $i \leq 0$. Note that (63), $\epsilon \in (0, 1]$ and $\kappa_{bhm} \geq 1$ together imply that $\Delta_\epsilon \leq 1$.

LEMMA 17. *Let Assumptions 4 and 5 hold. Let α and β be such that $\alpha\beta \geq 1/2$. Then, Assumption 1(ii) is satisfied for the stochastic process generated by Algorithm 2 with $W_k = 2(I_k J_k - \frac{1}{2})$, $\lambda = \log \gamma$, and $p = \alpha\beta$.*

The proof is similar to that of Lemma 7. We show that, conditioned on $T_\epsilon > k$ (that is, $\mathbb{1}(T_\epsilon > k) = 1$) where T_ϵ is defined in (62), (23) holds with Δ_ϵ as defined in (63). The case that differs from the proof of Lemma 7 and needs to be addressed here is when $\delta_k \leq \Delta_\epsilon$. In this case, 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 ζ in (63), we have that at least one of

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

or

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

holds, where we used the fact that $\kappa_{bhm} \geq 1$.

Suppose that $I_k = 1$ and $J_k = 1$, that is, both the model and the estimates are good in the k th iteration. Since the model m_k is κ -fully quadratic and $\delta_k \leq \Delta_\epsilon \leq 1$, then if (64) holds, we have

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

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 ϵ_F -s.o. accurate with $\epsilon_F \leq \kappa_{ef}$, (66) implies that condition (46) in Lemma 12 holds and (67) implies that condition (54) in Lemma 15 holds. Thus, in either case, iteration k is successful, that is, $x_{k+1} = x_k + s_k$ 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 1(ii).

To show that Assumption 1(iii) holds, we need an additional assumption on the accuracy of the function estimates. In addition, we will make a simplifying assumption of an upper bound on the trust-region radius⁵ in Algorithm 2.

ASSUMPTION 6. *We assume that:*

(a) *There exists a constant κ_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

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

⁵ This restriction can be avoided if one allows a more involved discussion on dominating terms in the proofs of Lemmas 10–12 and 16, and in the proof of the main result.

(b) The upper bound δ_{\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 β -probabilistically ϵ_F -s.o. accurate. However, for ϵ_F to satisfy the conditions in Assumption 5(b), additional conditions would have to be imposed on κ_F . Thus, for our purposes here, we choose to allow any finite $\kappa_F > 0$, and impose a bound only on ϵ_F .

Assumption 6(a) is needed for the case when we have a bad model and bad estimates in the k th iteration, the case in which the (true) objective may increase after a successful step. Without Assumption 6(a), it is possible that the increase in the objective is, in the worst case, on the order of δ_k^2 (due to first-order terms). Meanwhile, in the worst case, the objective decrease attained on other, successful steps is only guaranteed to be on the order of order δ_k^3 (due to second-order terms). Such a situation would make it impossible to balance the increase and decrease in the objective over the course of the algorithm in such a way to ensure that the stochastic process Φ_k decreases on average.

We now prove that Assumption 1(iii) holds for Algorithm 2.

THEOREM 6. *Let Assumptions 4, 5 and 6 hold. Then, there exist probabilities α and β and a constant $\Theta > 0$ such that for each iteration k of Algorithm 2, we have*

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

conditioned on $T_\epsilon > k$, where T_ϵ is defined in (62) and Φ_k is defined in (61).

Moreover, with the particular choice of constants described on page 30, let α and β 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 $\Theta \geq 6 \cdot 10^{-4} \eta_2 \min\{1, \eta_2\}$.

Since (68) clearly holds if $T_\epsilon \leq k$, we suppose in what follows that $T_\epsilon > k$. Then, $\tau(x_k) > \epsilon$, where $\tau(x)$ is defined in (38). We consider two possible cases: $\tau(x_k) \geq \zeta\delta_k$ and $\tau(x_k) < \zeta\delta_k$, where ζ is defined in (63). We show that (68) holds in either case, from which we can conclude that (68) holds 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 11, C_4 defined as in Lemma 14, and C_2 defined as in Lemma 16. 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 $\delta_{k+1} = \frac{1}{\gamma}\delta_k$, that is,

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

Case 1: $\tau(x_k) = \max\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\} \geq \zeta\delta_k$

a. $I_k = 1$ and $J_k = 1$, that is, both the model and the estimates are good in the k th iteration. From the definition of ζ and Case 1, we know that either (64) or (65) hold. Since $I_k = 1$ and $\delta_{\max} \leq 1$ (via Assumption 6(b)), (64) implies that condition (44) in Lemma 11 holds and (65) implies that condition (49) in Lemma 14 holds. Therefore, the trial step s_k leads to a decrease in f as in (45) or a decrease in f as in (50), respectively. Again from $I_k = 1$ and $\delta_{\max} \leq 1$, (64) or (65) imply that (66) or (67) hold, respectively. Since $J_k = 1$ and $\epsilon_F \leq \kappa_{ef}$, (66) and (67) imply that condition (46) in Lemma 12 and (54) in Lemma 15 hold, respectively. Thus in either case, iteration k is successful, that is, $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \max\{\delta_{\max}, \gamma\delta_k\}$.

Combining (45) and (72), we have that

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

with C_1 defined in Lemma 11. Since $\|\nabla f(x_k)\| \geq \zeta\delta_k$ we have that

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

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

Combining (50) and (72), we have that

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

with C_4 defined in Lemma 14. Again, since $-\lambda_{\min}(\nabla^2 f(X^k)) \geq \zeta\delta_k$ we have that

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

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

b. $I_k = 1$ and $J_k = 0$, that is, we have a good model and bad estimates in the k th iteration. In this case, the analysis of case (a) again applies; either Lemma 11 or 14 demonstrate that s_k yields a sufficient decrease in f . However, the step can be erroneously rejected, because of inaccurate function 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 $J_k = 1$, that is, we have a bad model and good estimates in the k th iteration. In this case, as in case (b), the k th iteration can be either successful or unsuccessful; in the latter

case, (73) holds. In the former, since the estimates are ϵ_F -accurate and (41) holds, then by Lemma 16 and Assumption 6(b), (59) holds with some $C_2 > 0$. Thus,

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

because $\nu \in (0, 1)$ satisfies (71).

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

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

Even in unsuccessful iterations, (73) still applies; this means that the right-hand side of (79) dominates and (79) holds whether the k th iteration is successful or not. Note that nowhere in the analysis of case (d) have we used the definition of Case 1.

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

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

Recalling Assumption 6, noting that $\mathbb{E}[\Delta_k^3 | \mathcal{F}_{k-1}^{M \cdot F}] = \Delta_k^3$, and rearranging terms, we obtain

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

Choosing $\alpha \in (0, 1]$ and $\beta \in (0, 1]$ 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

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

Case 2: $\tau(x_k) = \max\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\} < \zeta \delta_k$

i. $J_k = 1$, that is, we have good estimates but the model may be bad. The analysis of this case is similar to the analysis of Case 1(c), and so (78) holds on successful steps. Thus, decrease bounded by b_1 can again be guaranteed for ϕ_k whether the iteration is successful or not.

ii. $J_k = 0$, that is, we have bad estimates and the model may also be bad. In this case, the analysis of Case 1(d) again applies, because both f and δ_k may increase. We can once again upper bound the potential increase in ϕ_k using (79) on both successful and unsuccessful steps⁶.

We are now ready to compute the expectation of $\Phi_{k+1} - \Phi_k$ in Case 2. Case 2(i) occurs with probability at least β ; in Case 2(i), $\phi_{k+1} - \phi_k$ is bounded above by $b_1 < 0$, with b_1 defined in (73). Case 2(ii) happens with probability at most $(1 - \beta)$; in Case 2(ii), the possible increase in ϕ_k is bounded like (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 6, and because $\mathbb{E}[\Delta_k^3 | \mathcal{F}_{k-1}^{M \cdot F}] = \Delta_k^3$ 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

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

In conclusion, for ν satisfying (71) and α and β satisfying (80) and (83) respectively, the expected decrease in Φ_k in (68) holds, with

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

⁶ Note that under additional assumptions on κ_{ef} and η_2 , one can further refine the analysis here to account for the decrease in ϕ_k that could be achieved when $I_k = 1$.

Now, let us particularize these results with the constants given on page 30. Using $\eta_2 \leq 18$ and $\kappa_{bhm} = \kappa_{eh} + L$, we deduce that $\zeta := 20\kappa_{bhm} = 20(\kappa_{eh} + L)$ satisfies (63). A selection of $C_1 = C_4 = \frac{1}{10}$ satisfies the conditions in Lemmas 11 and 14. By Lemma 16 and our particular choice of ϵ_F , we select $C_2 = \frac{1}{80}\eta_2 \min\{1, \eta_2\}$. Thus, from (71) and because $\epsilon_F \leq \kappa_{eh} \leq \kappa_{bhm}$, ν must satisfy

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

We select $\nu = \frac{320}{320 + \eta_2 \min\{1, \eta_2\}} \in (0, 1)$. With these selections, (80) is equivalent to

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

which is implied by our choice in (69). Likewise, with these selections, the bound (83) is equivalent to

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

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 our choice in (70). Our choice of Θ also follows by noting that $\eta_2 \min\{1, \eta_2\} \leq 18$.

We contrast the second-order results of Theorem 6 with the first-order results of Theorem 3. The effect of the stronger assumption on the estimates in Assumption 6 (a) is clearly seen in the appearance of κ_F in the denominator of (69). Also, due to our particular choice of constants and requirements on the accuracy of the estimates, η_2 was assumed to be smaller than the Lipschitz constants \bar{L} and κ_{eh} ; hence, η_2 appears in the numerator of (69), while Lipschitz constants do not. In this light, η_2 can be interpreted as an additional means to control/ensure model quality, which is perhaps unsurprising given the definition of η_2 in Algorithm 2.

We now state and prove the main complexity result for Algorithm 2.

THEOREM 7 (Complexity of second-order STORM algorithm). *Consider Algorithm 2 and its corresponding stochastic process. Let T_ϵ be defined as in (62) with $\epsilon \in (0, 1]$. Then, under the assumptions of Theorem 6, 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 Φ_0 is defined in (61) with $k=0$, ν is defined in (71) and ζ is defined in (63).

Moreover, with the particular choices of constants described on page 30 and in Theorem 6, (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 $\Theta \geq 6 \cdot 10^{-4} \eta_2 \min\{1, \eta_2\}$.

The validity of Assumption 1(iii) follows from Theorem 6, with $h(\delta) = \delta^3$ and Δ_ϵ defined in (63). Lemma 17 and the discussion preceding it imply that Theorem 2 applies, from which we conclude (85).

A liminf-type almost sure convergence result trivially follows.

COROLLARY 1 (Convergence of second-order STORM algorithm). *Under the conditions of Theorem 7, the iterates $\{X_k\}$ generated by Algorithm 2 almost surely contain a subsequence convergent to a second-order stationary point of f .*

As in our discussion in Section 3.5, similar techniques for computing function, gradient, and Hessian estimates can be derived that satisfy Assumptions 5 and 6 for Algorithm 2 Bandeira et al. (2014).

5. Conclusion

In this manuscript, we proposed a general framework based on a stochastic process that can be used to bound the expected complexity of optimization algorithms. This framework can be applied beyond the algorithms discussed in this paper and has already been used in recent work on a stochastic line search method Paquette and Scheinberg (2018). We then applied this framework to demonstrate 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 a nonconvex setting. We then showed that a second-order stochastic trust-region method converges to second-order stationary point and moreover demonstrated that the expected complexity of this second-order method essentially matches the known complexity of second-order methods for second-order methods in nonconvex optimization settings. While the algorithms we analyzed require stochastic estimates to be progressively more accurate, the algorithms never require the computation of a full gradient; hence, the algorithms apply to purely stochastic settings.

References

- Alsmeyer G (2010) Renewal, recurrence and regeneration.
- Bandeira A, Scheinberg K, Vicente L (2013) Convergence of trust-region methods based on probabilistic models. Technical report, Lehigh University.
- Bandeira AS, Scheinberg K, Vicente LN (2014) Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization* 24(3):1238–1264.
- Billups S, Larsson J, Graf P (2011) Derivative-free optimization of expensive functions with computational error using weighted regression.
- Byrd R, Chin GM, Nocedal J, Wu Y (2012) Sample size selection in optimization methods for machine learning 134:127–155.

- Cartis C, Scheinberg K (2018) Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming* .
- Chang K, Li M, Wan H (2013) Stochastic trust-region response-surface method (strong) - a new response-surface framework for simulation optimization. *INFORMS Journal on Computing* 25(2):230–243.
- Chen R, Menickelly M, Scheinberg K (2018) Stochastic optimization using a trust-region method and random models. *Mathematical Programming* 169(2):447–487.
- Conn A, Scheinberg K, Vicente L (2009) *Introduction to Derivative-Free Optimization* (Philadelphia, PA, USA: Society for Industrial and Applied Mathematics), ISBN 0898716683, 9780898716689.
- Conn AR, Gould NIM, Toint PT (2000) *Trust Region Methods*. MPS/SIAM Series on Optimization (Philadelphia: SIAM).
- Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K, eds., *Advances in Neural Information Processing Systems 27*, 2933–2941 (Curran Associates, Inc.), URL <http://papers.nips.cc/paper/5486-identifying-and-attacking-the-saddle-point-problem-in-high-dimensional-non-convex-optimization.pdf>.
- Defazio A, Bach F, Lacoste-Julien S (2014) SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *NIPS*, 1646–1654.
- Friedlander M, Schmidt M (2012) Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing* 34(3):1380–1405.
- Ghadimi S, Lan G (2013) Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23(4):2341–2368.
- Golub GH, Loan CFV (1989) *Matrix Computations* (Johns Hopkins University Press, Baltimore), 2nd edition.
- Gratton S, Royer CW, Vicente LN, Zhang Z (2017) Complexity and global rates of trust-region methods based on probabilistic models. *IMA Journal of Numerical Analysis* .
- Lin C, Weng RC, Keerthi SS, Smola A (2007) Trust region newton method for large-scale logistic regression. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*.
- Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds., *Advances in Neural Information Processing Systems 26*, 315–323 (Curran Associates, Inc.), URL <http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-variance-reduction.pdf>.
- Larson J, Billups S (2015) Stochastic derivative-free optimization using a trust region framework. *Computational Optimization and Applications* .

- Nesterov Y (2004) *Introductory Lectures on Convex Optimization* (Boston, MA: Kluwer Academic Publishers).
- Nguyen L, Liu J, Scheinberg K, Takáč M (2017a) SARAH: A novel method for machine learning problems using stochastic recursive gradient. *ICML*.
- Nguyen L, Liu J, Scheinberg K, Takáč M (2017b) Stochastic recursive gradient algorithm for nonconvex optimization. *ArXiv e-prints* <https://arxiv.org/abs/1705.07261>.
- Nocedal J, Wright SJ (2006) *Numerical Optimization*. Springer Series in Operations Research (New York, NY, USA: Springer), 2nd edition.
- Paquette C, Scheinberg K (2018) A stochastic line search method with expected complexity analysis. *arXiv:1807.07994* .
- Reddi SJ, Hefny A, Sra S, Póczos B, Smola AJ (2016) Stochastic variance reduction for nonconvex optimization. *ICML*, 314–323.
- Shashaani S, Hashemi FS, Pasupathy R (2015) Astro-df: A class of adaptive sampling trust-region algorithms for derivative-free simulation optimization .
- Tripuraneni N, Stern M, Jin C, Regier J, Jordan MI (2017) Stochastic cubic regularization for fast nonconvex optimization. *arXiv:1711.02838* .
- Tropp JA (2015) An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.* 8(1-2):1–230, ISSN 1935-8237, URL <http://dx.doi.org/10.1561/22000000048>.
- Xu P, Roosta-Khorasani F, Mahoney MW (2017) Newton-type methods for non-convex optimization under inexact hessian information. *arXiv* arXiv:1708.07164.

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 10. Using the optimal decrease condition (37), the upper bound on model Hessian from Lemma 9, 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 κ -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 11. The definition of a κ -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 10 hold and we have

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

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

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

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

Proof of Lemma 12. Since $\delta_k \leq \frac{\|g_k\|}{\kappa_{bhm}}$, the model decrease condition (37) and the uniform bound on H_k under Lemma 9 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

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

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

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

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

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