On Common-Random-Numbers and the Complexity of Adaptive Sampling Trust-Region Methods

Yunsoo Ha, Sara Shashaani*, Raghu Pasupathy†

In the context of simulation optimization (SO), Common Random Numbers (CRN) is the practice of querying the simulation-based oracle with the same random number stream at each point visited by an SO algorithm. This practice is widely believed to facilitate SO algorithm efficiency by preserving structure inherent to the objective function and gradient sample-paths. However, CRN can present coding challenges compared to the widely-used practice of naïve independent sampling. Is the potential CRN efficiency gain worth the potentially significant cost of implementation within stochastic trust-region algorithms? Toward answering this question, we characterize the consistency and complexity of a class of stochastic trust-region algorithms called ASTRO/ASTRO-DF as a function of the use of CRN. We find that the magnitude of CRN's influence depends intimately on the extent of regularity in the underlying sample paths. For instance, CRN's effect is most evident in first-order settings with smooth sample paths, where the algorithm work complexity dramatically improves from $O(\epsilon^{-6})$ to $O(\epsilon^{-2})$. This result is significant considering that the best work complexity of first-order (generic) stochastic trust-region algorithms reported in the literature is $O(\epsilon^{-6})$. CRN's effect is more muted when the sample paths are potentially discontinuous, with the work complexity improving from $O(\epsilon^{-6})$ to $O(\epsilon^{-5})$ in both zeroth-order and first-order settings. In between these extremes, CRN facilitates various improved complexities depending on prevailing conditions of sample-path regularity. We anticipate similar gains in adaptive sampling algorithms other than ASTRO/ASTRO-DF since the derived complexities stem less due to specific algorithmic mechanics, and more due to elements common to all trust-region methods.

1. INTRODUCTION

We consider stochastic optimization problems having the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \mathbb{E}[F(\boldsymbol{x},\xi)] = \int_{\Xi} F(\boldsymbol{x},\xi)\, P(\mathrm{d}\xi), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and bounded from below, and $F : \mathbb{R}^d \times \Xi \to \mathbb{R}$ is a "random function" with the random object ξ having distribution P on the measurable space $(\Xi, \mathcal{F})$. The problem formulation in (1) is the subject of enormous recent attention especially due to the advent of machine learning and data analytics.

Two popular flavors of (1), loosely called *derivative-based* or *first-order*, and *derivative-free* or *zeroth-order*, are of interest in this paper, and reflect the extent of information on F that is available

*Department of Industrial and Systems Engineering, North Carolina State University, sshasha2@ncsu.edu
†Department of Statistics, Purdue University, pasupath@purdue.edu

to a solution algorithm. In the first-order context, it is assumed that $F(\cdot,\xi)$ is differentiable with $\mathbb{E}[\nabla F(\cdot,\xi)] = \nabla f(\cdot)$, and that an algorithm has access to a *first-order stochastic oracle*, that is, the oracle queried at $(\boldsymbol{x},\xi)$ returns $(F(\boldsymbol{x},\xi), \nabla F(\boldsymbol{x},\xi))$. In the derivative-free context, $F(\cdot,\xi)$ is not necessarily differentiable and an algorithm is assumed to have access only to a *zeroth-order stochastic oracle*, that is, when queried at $(\boldsymbol{x},\xi)$ the oracle returns only the function value $F(\boldsymbol{x},\xi)$. In both first-order and zeroth-order contexts, the objective $f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}$ is assumed to be a smooth function. (See Assumption 1 for a precise statement of this assumption.)

The random function $F(\cdot,\xi)$ is called a *sample-path approximation* of $f(\cdot)$; likewise, $\nabla F(\cdot,\xi)$, when it exists, is called a sample-path approximation of $\nabla f(\cdot)$. Sample-average approximations $\bar{F}(\cdot,n)$ and $\bar{\boldsymbol{G}}(\cdot,n)$ of $f(\cdot)$ and $\nabla f(\cdot)$, respectively, can be constructed by averaging sample paths $F(\cdot,\xi_j), j = 1,2,\ldots,n$ and $\nabla F(\cdot,\xi_j), j = 1,2,\ldots,n$:

$$\bar{F}(\boldsymbol{x},n) = \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{x},\xi_i); \quad \bar{\boldsymbol{G}}(\boldsymbol{x},n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{x},\xi_i), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

The estimated variance of $\bar{F}(\cdot,n)$, and the estimated covariance of $\bar{\boldsymbol{G}}(\cdot,n)$, are then computed in the usual way:

$$\hat{\sigma}^2_F(\boldsymbol{x},n) := \frac{1}{n-1}\sum_{j=1}^{n} \left(F(\boldsymbol{x},\xi_i) - \bar{F}(\boldsymbol{x},n)\right)^2; \tag{2}$$

$$\hat{\sigma}^2_{\boldsymbol{G}}(\boldsymbol{x},n) := \frac{1}{n-1}\sum_{j=1}^{n} \left(\boldsymbol{G}(\boldsymbol{x},\xi_i) - \bar{\boldsymbol{G}}(\boldsymbol{x},n)\right)\left(\boldsymbol{G}(\boldsymbol{x},\xi_i) - \bar{\boldsymbol{G}}(\boldsymbol{x},n)\right)^{\intercal}. \tag{3}$$

1.1 Consistency, Iteration Complexity, Work Complexity

An iterative solution algorithm having access to either a stochastic zeroth-order or stochastic first-order oracle, is said to solve Problem (1) if it constructs a stochastic iterate sequence $\{\boldsymbol{X}_k, k \geq 1\}$ that in some rigorous sense converges to a *first-order critical point* of f, that is, a point $\boldsymbol{x}^*$ such that satisfies $\|\nabla f(\boldsymbol{x}^*)\| = 0$. For the purposes of this paper, such an algorithm is said to be *consistent* if the generated sequence $\boldsymbol{X}_k \to \boldsymbol{x}^*$ in probability and *strongly consistent* if $\boldsymbol{X}_k \to \boldsymbol{x}^*$ almost surely. (See Section 3 for a formal definitions of convergence modes).

Suppose the stochastic process of iterates $\{\boldsymbol{X}_k, k \geq 1\}$ is defined on a probability space $(\Omega,\mathcal{F},P)$ equipped with the filtration $\{\mathcal{F}_k, k \geq 1\}$. Suppose also that in generating the iterate $\boldsymbol{X}_k$, an iterative algorithm expends N_k calls to the stochastic oracle, where N_k is an *adaptive sample size*, that is, N_k is $\mathcal{F}_k$-measurable. Then, the total work done after k iterations is also $\mathcal{F}_k$-measurable and given by

$$W_k = \sum_{j=1}^{k} N_j. \tag{4}$$

Then, a consistent algorithm is said to exhibit $O(\epsilon^{-q})$ *work complexity* if

$$W_{K(\epsilon)} \leq \Lambda_W\, \epsilon^{-q}; \quad K(\epsilon) := \inf\{k : \|\nabla f(X_k)\| \leq \epsilon\}, \tag{5}$$

where Λ_W is a well-defined random variable and $\|\cdot\|$ is the L_2 norm. We say the consistent algorithm exhibits $\tilde{O}(\epsilon^{-q})$ complexity if there exists a well-defined random variable Λ_W such that

$$W_{K(\epsilon)} \leq \Lambda_W\, \epsilon^{-q} \log \epsilon^{-1}; \quad K(\epsilon) := \inf\{k : \|\nabla f(X_k)\| \leq \epsilon\}. \tag{6}$$

The conditions in (5) and (6) can be loosely interpreted as the iterative algorithm needing at most $O(\epsilon^{-q})$ calls to the oracle to reach ϵ-accuracy, that is, the generated iterates first enter an ϵ-ball centered on a first-order critical point. An analogous definition holds for iteration complexity whereby a consistent algorithm is said to exhibit $O(\epsilon^{-q})$ *iteration complexity* if there exists a well-defined random variable Λ_T such that

$$K(\epsilon) \leq \Lambda_T \, \epsilon^{-q}; \quad K(\epsilon) := \inf\{k : \|\nabla f(X_k)\| \leq \epsilon\}. \tag{7}$$

Depending on the prevailing computing environment, one or both of work complexity and iteration complexity may be important to an implementer. Although we include results on iteration complexity, our emphasis in this paper is on work complexity. Such emphasis is informed by the fact that the per-iteration effort in most adaptive sampling algorithms varies across iterations, and is not adequately representative of the total computing effort expended to reach ϵ-accuracy.

1.2 Stochastic Trust-region Algorithms in a Nutshell

Trust-region (TR) algorithms are a family of iterative methods for solving smooth nonconvex stochastic optimization problems that have recently gained in popularity due primarily to the robustness stemming their self-tuning nature. All stochastic TR algorithms include the following four steps in each iteration:

(a) (model construction) a model of the objective f is constructed within a "trust region," usually an L_2 ball of radius Δ_k centered around the incumbent iterate $\boldsymbol{X}_k$, by appropriately calling the provided oracle;

(b) (subproblem minimization) the constructed local model of f is approximately minimized within the trust region to yield a candidate point $\widetilde{\boldsymbol{X}}_{k+1}$;

(c) (candidate evaluation) the candidate point $\widetilde{\boldsymbol{X}}_{k+1}$ is accepted or rejected based on a sufficient reduction (of f) test; and

(d) (trust-region management) if accepted, $\widetilde{\boldsymbol{X}}_{k+1}$ replaces $\boldsymbol{X}_k$ as the subsequent incumbent and the trust-region radius Δ_k is either increased or stays the same as a vote of confidence in the model; if $\widetilde{\boldsymbol{X}}_{k+1}$ is rejected, the incumbent remains unchanged during the subsequent iteration, and the trust-region radius shrinks by a factor in an attempt at constructing a better model in the subsequent iteration.

The difference between existing stochastic TR algorithms, e.g., STRONG [1, 2], STORM [3, 4], ASTRO/ASTRO-DF [5, 6], and TRiSH [7] largely stem from the particularities of the steps (a)–(d). As should become clear from our analysis in this paper, the work complexity associated with stochastic TR models depends especially on step (a), particularly on how the model is constructed, how much effort is expended, and what is the resulting accuracy.

1.3 CRN and the Random Field Interpretation

The object $F(\cdot, \xi)$ appearing in (1) is a *random field*, that is, a collection $\{F(x, \xi), x \in \mathbb{R}^d\}$ of random variables labeled by $x \in \mathbb{R}^d$. This viewpoint forms the foundation of CRN use in optimization as a mechanism to preserve inherent structure in $F(\cdot, \xi)$, otherwise lost through independent sampling.

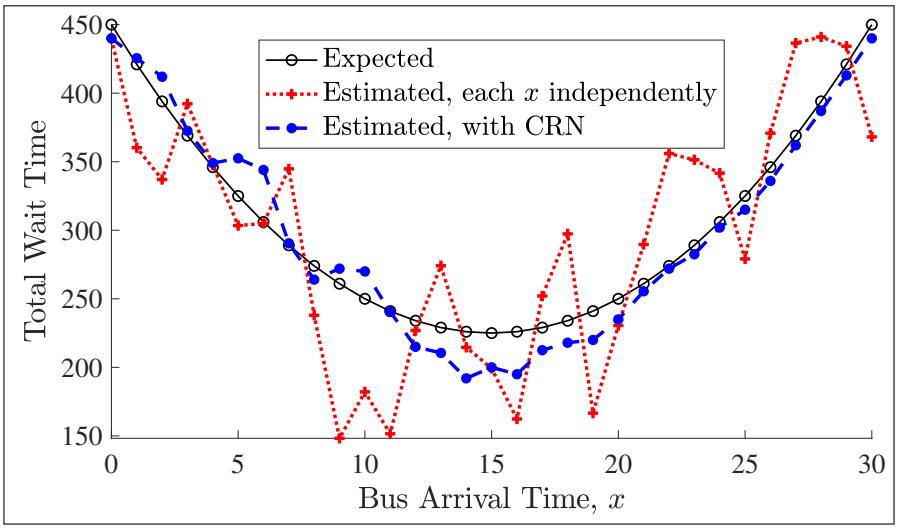

FIGURE 1: Illustration of the effect of CRN on an example problem adapted from [8]. The black curve is the true function f, the blue curve is f's estimate constructed with CRN, and the red curve is f's estimate constructed with independent sampling.

To illustrate, suppose a stochastic TR algorithm attempts to construct a local model (step (a) in Section 1.2) by calling the oracle at $2d+1$ points denoted $x_1, x_2, \ldots, x_{2d+1}$. Recall that executing the oracle once entails specifying two inputs: a point $x \in \mathbb{R}^d$ and a "random number" ξ. Using CRN when calling the oracle means "holding the random number fixed" (at say ξ) as the oracle is called at the points $x_1, x_2, \ldots, x_{2d+1}$, that is, the inputs to the oracle are $(x_1, \xi), (x_2, \xi), \ldots, (x_{2d+1}, \xi)$. The oracle will then return $F(x_1, \xi), F(x_2, \xi), \ldots, F(x_{2d+1}, \xi)$. This is in contrast to independent sampling with $(x_1, \xi_1), (x_2, \xi_2), \ldots, (x_{2d+1}, \xi_{2d+1})$ inputs to the oracle, where $\xi_1, \xi_2, \ldots, \xi_{2d+1}$ are independent and identically distributed random variables. The idea of CRN extends when calling the oracle multiple, say N, times at each $x_j, j = 1, 2, \ldots, x_{2d+1}$ in a similar manner. The inputs to the oracle become $(x_1, \xi_j), (x_2, \xi_j), \ldots, (x_{2d+1}, \xi_j),\ j = 1, 2, \ldots, N$ and the resulting function estimates become

$$\bar{F}(\boldsymbol{x}_i, N) := \frac{1}{N} \sum_{j=1}^{N} F(x_i, \xi_j),\ i = 1, 2, \ldots, 2d+1.$$

For a concrete example, Figure 1 depicts the expected waiting time of passengers arriving (to board a bus) according to a Poisson process as a function of placement x of six buses in a fixed time interval $[0, 30]$. The black curve in Figure 1 represents the expected wait time $f(\boldsymbol{x})$, the blue curve represents the estimated wait time $\bar{F}(\boldsymbol{x}, N)$ generated with CRN and the red curve represents the estimated wait time $\bar{F}(\boldsymbol{x}, N)$ generated with independent sampling. When looking to identify bus arrival times that minimize total expected wait time, an estimator obtained by minimizing the blue curve is likely to be much closer to the true minimum than that obtained by minimizing the red curve.

Since models in step (a) of Section 1.2 are constructed to stipulated accuracy, CRN often directly translates to lower overall sampling effort, leading to the widely held opinion that CRN aids numerical implementation. What is perhaps most interesting is that our results in this paper suggest that the gains are not just during implementation, but are reflected clearly in the work complexity of the resulting algorithms.

2. SUMMARY OF INSIGHT

In what follows, we summarize the principal insights of this paper.

(a) Table 1 presents the work complexities implied by Theorem 7 of the paper organized by smoothness of the function sample-path (None/Continuous/Smooth), CRN use (or lack thereof), and oracle information (first-order or zeroth order).

	First-order (ASTRO)			Zeroth-order (ASTRO-DF)		
	Function sample-path property					
	None	Continuous	Smooth	None	Continuous	Smooth
CRN	$\widetilde{\mathcal{O}}(\epsilon^{-5})$	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	$\widetilde{\mathcal{O}}(\epsilon^{-2})$	$\widetilde{\mathcal{O}}(\epsilon^{-5})$	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	$\widetilde{\mathcal{O}}(\epsilon^{-4})$
No CRN	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$

TABLE 1: Work complexity rates for ASTRO and ASTRO-DF.

Various aspects of Table 1 are salient. First, when CRN is not used, consistent with the single existing result in the literature, the work complexity is $\widetilde{\mathcal{O}}(\epsilon^{-6})$ across the board. Second, the complexity improves dramatically to $\widetilde{\mathcal{O}}(\epsilon^{-2})$ when using CRN in the first-order context with smooth sample-paths. (It is now well-known that $\mathcal{O}(\epsilon^{-2})$ is the best achievable complexity for stochastic optimization with a first-order stochastic oracle.) Third, there appears to be a steady progression of benefit due to using CRN, with smooth sample-paths providing the maximum benefit, followed by contexts having continuous sample-paths, and finally those having potentially discontinuous sample-paths. Fourth, there appears to no difference in complexities between between first-order and zeroth-order contexts except in the smooth context.

(b) Theorem 6 demonstrates that the iteration complexity of both ASTRO and ASTRO-DF is $O(\epsilon^{-2})$, matching the best achievable for any algorithm that expends a constant amount of sampling effort in each iteration.

(c) The complexity and consistency proofs in the paper follow from arguably weak conditions. For instance, no assumptions are made on success probabilities of the underlying model appearing in Step (a) of Section 1.2, or on its independence from function estimates. (This is in contrast to proofs in [4] which rely on such "black-box" assumptions.)

(d) The optimal complexity $O(\epsilon^{-2})$ is achieved by classic algorithms such as stochastic gradient descent (SGD) with independent sampling, whereas the corresponding complexity in Table 1 is $\tilde{O}(\epsilon^{-6})$. This large discrepancy in complexities seems to arise from the standard manner in which the sufficient reduction test (Step (c) in Section 1.2) is conducted in many TR algorithms using a quality stipulation on both the function and gradient estimates.

3. MATHEMATICAL PRELIMINARIES

In this section, we provide the notation, key definitions, standing assumptions and some supporting results that will be invoked in the paper.

3.1 Notation

We use bold font for vectors; $\boldsymbol{x} = (x_1, x_2, \cdots, x_d) \in \mathbb{R}^d$ denotes a d-dimensional vector of real numbers. Let $\boldsymbol{e}^i \in \mathbb{R}^d$ for $i = 1, \ldots, d$ denote the unit basis vector in the ith coordinate. We use calligraphic fonts for sets and sans serif fonts for matrices. Our default norm operator $\|\cdot\|$ is an L_2 norm in the Euclidean space. We use $a \wedge b := \min\{a, b\}$ and denote $\mathcal{B}(\boldsymbol{x}^0; \Delta) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}^0\|_2 \leq \Delta\}$ as the closed ball of radius $\Delta > 0$ with center $\boldsymbol{x}^0$. For a sequence of sets $\{\mathcal{A}_n\}$, the set $\{\mathcal{A}_n \text{ i.o.}\}$ denotes $\limsup_n \mathcal{A}_n$. For sequences $\{a_k\}$ and $\{b_k\}$ of nonnegative reals, $a_k \sim b_k$ denotes $\lim_{k\to\infty} a_k/b_k = 1$. We say $f(\boldsymbol{x}) = \mathcal{O}(g(\boldsymbol{x}))$ if there exist positive numbers ε and m such that $|f(\boldsymbol{x})| \leq m g(\boldsymbol{x})$ for all $\boldsymbol{x}$ with $0 < |\boldsymbol{x}| < \varepsilon$. $\mathcal{C}(\mathbb{R}^d)$ denotes the set of all continuous functions on $\mathbb{R}^d$. We use capital letters for random scalars and vectors. XY denotes independent random variables X and Y. For a sequence of random vectors $\{\boldsymbol{X}_k\}$, $k \in \mathbb{N}$, $\boldsymbol{X}_k \xrightarrow{wp1} \boldsymbol{X}$ denotes almost sure convergence. "iid" abbreviates independent and identically distributed and "wp1" abbreviates with probability 1.

3.2 Key Definitions

Progress of TR algorithms relies on a local model constructed using function value estimates, often as a quadratic approximation:

$$M_k(\boldsymbol{X}_k + \boldsymbol{s}) = \bar{F}_k^0(N_k) + \boldsymbol{s}^\intercal \boldsymbol{G}_k + \frac{1}{2}\boldsymbol{s}^\intercal \mathsf{H}_k \boldsymbol{s}, \text{ for all } \boldsymbol{s} \in \mathcal{B}(0; \Delta_k) \tag{8}$$

where $\boldsymbol{G}_k$ and H_k are the model gradient and Hessian at the incumbent solution $\boldsymbol{X}_k$ and Δ_k is the size of the neighborhood around $\boldsymbol{X}_k$ where the model is deemed credible.

In the first-order context where we have access to a stochastic first-order oracle, the model gradient $\boldsymbol{G}_k \equiv \bar{\boldsymbol{G}}_k(N_k) = \frac{1}{N_k}\sum_{i=1}^{N_k} \boldsymbol{G}(\boldsymbol{X}_k, \xi_i)$, is simply the unbiased gradient estimate and H_k replaced by it approximation using BFGS [9, 10], i.e.,

$$\mathsf{B}_k = \mathsf{B}_{k-1} - (\boldsymbol{S}_{k-1}^\intercal \mathsf{B}_{k-1} \boldsymbol{S}_{k-1})^{-1} \mathsf{B}_{k-1} \boldsymbol{S}_{k-1} \boldsymbol{S}_{k-1}^\intercal \mathsf{B}_{k-1} + (\boldsymbol{Y}_{k-1}^\intercal \boldsymbol{S}_{k-1})^{-1} \boldsymbol{Y}_{k-1} \boldsymbol{Y}_{k-1}^\intercal,$$

where $\boldsymbol{Y}_{k-1} = \bar{\boldsymbol{G}}_k(N_k) - \bar{\boldsymbol{G}}_{k-1}(N_{k-1})$ and $\boldsymbol{S}_{k-1} = \boldsymbol{X}_k - \boldsymbol{X}_{k-1}$.

For a zeroth-order stochastic oracle, this local model can be constructed by fitting a quadratic surface on the function estimates at neighboring points, detailed in Definition 1.

Definition 1. *(*stochastic interpolation models*) Given $\boldsymbol{X}_k = \boldsymbol{X}_k^0 \in \mathbb{R}^d$ and $\Delta_k > 0$, let $\Phi(\boldsymbol{x}) = (\phi_0(\boldsymbol{x}), \phi_1(\boldsymbol{x}), \ldots, \phi_q(\boldsymbol{x}))$ be a polynomial basis on $\mathbb{R}^d$. With $p = q$ and a design set $\mathcal{X}_k := \{\boldsymbol{X}_k^0, \boldsymbol{X}_k^1, \ldots, \boldsymbol{X}_k^p\} \subset \mathcal{B}(\boldsymbol{X}_k; \Delta_k)$, we seek $\boldsymbol{\alpha}_k = \begin{bmatrix}\alpha_{k,0} & \alpha_{k,1} & \ldots & \alpha_{k,p}\end{bmatrix}$ such that*

$$\mathcal{M}(\Phi, \mathcal{X}_k)\boldsymbol{\alpha}_k = \begin{bmatrix}\bar{F}_k^0(N_k) & \bar{F}_k^1(N_k^1) & \cdots & \bar{F}_k^p(N_k^p)\end{bmatrix}^\intercal,$$

where for $i = 1, 2, \ldots, p$, N_k^i is the k-th iteration's adaptive sample size at the i-th design point, $\bar{F}_k^i(N_k^i) := \bar{F}(\boldsymbol{X}_k^i, N_k^i)$, and $\mathcal{M}(\Phi, \mathcal{X}_k) = [\boldsymbol{\phi}_k^0, \boldsymbol{\phi}_k^1, \ldots, \boldsymbol{\phi}_k^p]^\intercal$ with $\boldsymbol{\phi}_k^i = [\phi_1(\boldsymbol{X}_k^i), \phi_2(\boldsymbol{X}_k^i), \ldots, \phi_q(\boldsymbol{X}_k^i)]$. If the matrix $\mathcal{M}(\Phi, \mathcal{X}_k)$ is nonsingular, the set $\mathcal{X}_k$ is poised in $\mathcal{B}(\boldsymbol{X}_k; \Delta_k)$. The function $M_k : \mathcal{B}(\boldsymbol{X}_k; \Delta_k) \to \mathbb{R}$, defined as $M_k(\boldsymbol{x}) = \sum_{i=0}^p \alpha_{k,i}\phi_i(\boldsymbol{x})$ is a stochastic polynomial interpolation model of f on $\mathcal{B}(\boldsymbol{X}_k; \Delta_k)$. For representation of M_k in (8), $\boldsymbol{G}_k = \begin{bmatrix}\alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,d}\end{bmatrix}^\intercal$ be the subvector of $\boldsymbol{\alpha}_k$ and H_k be a symmetric matrix of size $d \times d$ with elements uniquely defined by $\alpha_{k,d+1}, \alpha_{k,d+2}, \cdots, \alpha_{k,p}$.

Taylor bounds for first-order local model errors need to be replicated for the zeroth-order models for sufficient model quality. This is classically done through the concept of fully-linear models [11], whose stochastic variant we list in Definition 2.

Definition 2. *(*stochastic fully linear models*) Given $\boldsymbol{X}_k \in \mathbb{R}^d$ and $\Delta_k > 0$, let model $M_k(\cdot)$ be obtained following Definition 1 and define $m_k(\cdot)$ as its limiting function where $N_k^i = \infty$ for $i = 0, 1, \cdots, p$. We say M_k is a stochastic fully linear model of f in $\mathcal{B}(\boldsymbol{X}_k; \Delta_k)$ if there exist positive constants κ_{eg} and κ_{ef} independent of $\boldsymbol{X}_k$ and Δ_k such that*

$$\|\nabla f(\boldsymbol{x}) - \nabla m_k(\boldsymbol{x})\| \leq \kappa_{eg}\Delta_k, \text{ and } \|f(\boldsymbol{x}) - m_k(\boldsymbol{x})\| \leq \kappa_{ef}\Delta_k^2 \ \forall \boldsymbol{x} \in \mathcal{B}(\boldsymbol{X}_k; \Delta_k). \tag{9}$$

Certain geometry of the design set will fulfill the fully-linear property of the local model [11]. Furthermore, to keep the model gradient in tandem with the trust-region radius Δ_k that ultimately reduces to 0, an additional check of $\|\boldsymbol{G}_k\|$ and Δ_k is often performed for the zeroth-order oracles (see criticality steps in **(author?)** [12]). The minimization is a constrained optimization and often, solving it to a point of Cauchy reduction is sufficient for TR methods to converge.

Definition 3. *(*Cauchy reduction*) Given $\boldsymbol{X}_k \in \mathbb{R}^d$ and $\Delta_k > 0$ and a model $M_k(\cdot)$ obtained following Definition 1, $\boldsymbol{S}_k^c$ is called the Cauchy step if*

$$M_k(\boldsymbol{X}_k) - M_k(\boldsymbol{X}_k + \boldsymbol{S}_k^c) \geq \frac{1}{2}\|\boldsymbol{G}_k\| \left(\frac{\|\boldsymbol{G}_k\|}{\|\mathsf{H}_k\|} \wedge \Delta_k \right). \tag{10}$$

We assume that $\|\boldsymbol{G}_k\|/\|\mathsf{H}_k\| = +\infty$ when $\|\mathsf{H}_k\| = 0$. We call the RHS of (10), the Cauchy reduction. The Cauchy step is obtained by minimizing the model $M_k(\cdot)$ along the steepest descent direction within $\mathcal{B}(\boldsymbol{X}_k; \Delta_k)$ and hence easy and quick to obtain.

With these preliminaries, we assess adaptive sampling rules defined below for function values and gradients in zeroth and first-order cases.

Definition 4. *(*Filtration and Stopping Time*). A filtration $\{\mathcal{F}_k\}_{k\geq 1}$ is defined as an increasing family of σ-algebras of $\mathcal{F} \cup \mathcal{F}^g$, i.e., $\mathcal{F}_k \subset \mathcal{F}_{k+1} \subset \cdots$ for all k. We interpret $\mathcal{F}_k$ as "all the information available at time k." A filtered space $(\Omega, \{\mathcal{F}_k\}_{k\geq 1}, \mathbb{P})$ is a probability space equipped with a filtration. A map $N : \Omega \to \{0, 1, 2, \ldots, \infty\}$ is called a stopping time with respect to $\mathcal{F}_k$ if the event $\{N = n\} := \{\omega : N(\omega) = n\} \in \mathcal{F}_k$ for all $n \leq \infty$.*

3.3 Standing Assumptions

The following assumptions specify the nature of the underlying function and random fields that will be used througout the paper.

Assumption 1. *(κ_{Lg}: Lipschitz constant of gradients) The function f is twice continuously differentiable in an open domain $\mathcal{X} \supseteq \mathcal{B}(\boldsymbol{x}_0; \Delta_{\max})$, ∇f is Lipschitz continuous in $\mathcal{X}$ with constant $\kappa_{Lg} \in (0, \infty)$, i.e., $\|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\| \leq \kappa_{Lg}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$, $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$.*

At the end of each iteration k, the stochastic process $(\{\boldsymbol{X}_k^i\}_{i=0,1,\cdots,p}, \widetilde{\boldsymbol{X}}_{k+1}, \widetilde{N}_{k+1}, \Delta_{k+1})$ becomes $\mathcal{F}_k$ measurable. On the other hand, the sampling error

$$\bar{E}_k^i(N_k^i) = \frac{1}{N_k^i}\sum_{j=1}^{N_k^i} E_{k,j}^i = \frac{1}{N_k}\sum_{j=1}^{N_k^i} F(\boldsymbol{X}_k^i, \xi_j) - f(\boldsymbol{X}_k^i),$$

is a martingale with $E^i_{k,j} \in \mathcal{F}_{k,j}$ and $\mathbb{E}[E^i_{k,j}|\mathcal{F}_{k,j-1}] = 0$, where $\mathcal{F}_k := \mathcal{F}_{k,0} \subset \mathcal{F}_{k,1} \subset \cdots \subset \mathcal{F}_{k+1}$. Similarly, in the presence of a first-order stochastic oracle,

$$\bar{\boldsymbol{E}}^g_k(N_k) = \frac{1}{N_k}\sum_{j=1}^{N_k} \boldsymbol{E}^g_{k,j} = \frac{1}{N_k}\sum_{j=1}^{N_k} \boldsymbol{G}(\boldsymbol{X}_k, \xi_j) - \nabla f(\boldsymbol{X}_k),$$

is a martingale with $\boldsymbol{E}^g_{k,j} \in \mathcal{F}_{k,j}$ and $\mathbb{E}[\boldsymbol{E}^g_{k,j}|\mathcal{F}_{k,j-1}] = 0$. We make the next two assumptions on the higher moments of the stochastic noise resembling the Bernstein condition.

Assumption 2. *Let the random iterate be $\boldsymbol{X}_k \in \mathcal{F}_{k-1}$ (in ASTRO-DF, let the design set be $\{\boldsymbol{X}^i_k\}_{i=0,1,\cdots,p} \in \mathcal{F}_{k-1}$). Then the stochastic errors $E^i_{k,j}$ are independent of $\mathcal{F}_{k-1}$, $\mathbb{E}[E^i_{k,j}|\mathcal{F}_{k,j-1}] = 0$, and there exists $\sigma^2_f > 0$ and $b_f > 0$ such that for a fixed n,*

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[|E^i_{k,j}|^m|\mathcal{F}_{k,j-1}] \leq \frac{m!}{2} b_f^{m-2}\sigma^2_f, \ \forall m = 2, 3, \cdots, \forall k.$$

Assumption 3. *Let $\boldsymbol{X}_k \in \mathcal{F}_{k-1}$ and $[\boldsymbol{E}^g_k]_r$ be the r-th element of the stochastic gradient error. Then $[\boldsymbol{E}^g_{k,j}]_r\mathcal{F}_{k-1}$ for any $r \in \{1, \ldots, d\}$ and $\mathbb{E}\left[[\boldsymbol{E}^g_{k,j}]_r \mid \mathcal{F}_{k,j-1}\right] = 0$. There also exist $\sigma^2_g > 0$ and $b_g > 0$ such that for a fixed n and any $r \in \{1, \ldots, d\}$,*

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[|[\boldsymbol{E}^g_{k,j}]_r|^m|\mathcal{F}_{k,j-1}\right] \leq \frac{m!}{2} b_g^{m-2}\sigma^2_g, \ \forall m = 2, 3, \cdots, \forall k.$$

Random variables fulfilling Assumptions 2 and 3 exhibit a subexponential tail behavior. Next, to analyze the sums and maxima of the stochastic error estimates, we need characterize the tail probability of a sequence of dependent random variables.

Assumption 4. *For a given solution at iteration k and constant $c > 0$, there exists a large $c_0 > 0$ such that for all $\lambda_k \leq n \leq N_k$, $\frac{1}{n-1}\sum_{j=1}^{n-1}|E_{k,j}|$ is stochastically decreasing in $|E_{k,n}|$, meaning that for a fixed c_2, $\mathbb{P}\{\frac{1}{n-1}\sum_{j=1}^{n-1}|E_{k,j}| > c_2 \mid |E_{k,n}| > c_1\}$ is decreasing in c_1; and likewise is $\frac{1}{n-1}\sum_{j=1}^{n-1}\|\boldsymbol{E}^g_{k,j}\|$ stochastically decreasing in $\|\boldsymbol{E}^g_{k,n}\|$.*

Note, this assumption allows for dependence of the consecutive stochastic error estimates with certain tail independence structure, in order to be able to apply a similar tail probability result for heavy-tailed (sub-exponential) random variables. The dependence structure is non-restrictive leading to the subexponentiality of the summands eliminating the impact of dependence of the tail behavior of the sums. Now let us introduce two assumptions which are needed to prove consistency.

3.4 Supporting Results

Lemma 1. *(Bernstein's inequality for martingales). Let Assumption 2 hold. Then, for all $c > 0$ and a fixed $n \in \mathbb{N}$, $\mathbb{P}\left\{\bar{E}_k(n) \geq c\right\} \leq \exp\left\{\frac{-nc^2}{2(cb+\sigma^2)}\right\}$.*

Note, Lemma 1 allows for the stochastic errors to be dependent, thereby facilitating using this result for our estimation error resulting from the stopping time N_k.

Lemma 2. *(Sums and Maxima of Dependent Subexponentials [13, 14]) Let Assumptions 2 and 4 hold for random variables $E_{k,j}$. Then $\forall n \in [\lambda_k, N_k]$,*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n}|E_{k,j}| > c\right\} \sim \mathbb{P}\left\{\sup_{\lambda_k \le n \le N_k}\frac{1}{n}\sum_{j=1}^{n}|E_{k,j}| > c\right\} \sim \mathbb{E}\left[\sum_{\lambda_k}^{N_k}\mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n}|E_{k,j}| > c\right\} \mid \mathcal{F}_k\right].$$

We review the reduction in the variance of difference between two adjacent points due to the use of common random numbers and assumptions on the sample-paths, a result that is often used for the gradients approximated with finite-differencing.

Theorem 1. *(Variance in Differences [15]) Let the function value at two adjacent points $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{s}$ be simulated to obtain $F(\boldsymbol{x},\xi_i)$ and $F(\boldsymbol{x}+\boldsymbol{s},\xi_j)$. Then the variance of the difference is*

$$Var(F(\boldsymbol{x}+\boldsymbol{s},\xi_j) - F(\boldsymbol{x},\xi_i)) = \begin{cases} \mathcal{O}(1) & \text{if } \xi_i\xi_j, \\ \mathcal{O}(\|\boldsymbol{s}\|) & \text{if } \xi_i = \xi_j, \\ \mathcal{O}(\|\boldsymbol{s}\|^2) & \text{if } \xi_i = \xi_j \text{ and } F(\cdot,\xi) \in \mathcal{C}(\mathbb{R}^d) \text{ for each } \xi \in \Xi. \end{cases}$$

4. ASTRO and ASTRO-DF

Before describing the details of ASTRO and ASTRO-DF, we focus on the adaptive sampling rule for each. These rules contain constants $\beta_f, \beta_g \in [0,2]$ that will be determined based on the use of CRN and sample-path behavior.

$$N_k = \min\left\{n \ge \lambda_k : \left(\frac{\sqrt{\mathrm{Tr}(\hat{\sigma}^2_{\boldsymbol{G}}(\boldsymbol{X}_k,n))}}{\sqrt{n}} \le \kappa_{ag}\frac{\Delta_k^{\beta_g}}{\sqrt{\lambda_k}}\right) \cap \left(\frac{\hat{\sigma}_F(\boldsymbol{X}_k,n)}{\sqrt{n}} \le \kappa_{af}\frac{\Delta_k^{\beta_f}}{\sqrt{\lambda_k}}\right)\right\}, \tag{11}$$

$$N_k^i = \min\left\{n \ge \lambda_k : \frac{\hat{\sigma}_F(\boldsymbol{X}_k^i,n)}{\sqrt{n}} \le \kappa_{af}\frac{\Delta_k^{\beta_f}}{\sqrt{\lambda_k}}\right\}. \tag{12}$$

The sample sizes specified above are stopping times lower bounded by a deterministically increasing sequence $\{\lambda_k\}$ that grows logarithmically in k. In particular, $\mathcal{O}(\lambda_k) = (\log k)^{1+\epsilon_\lambda}$ for some $\epsilon_\lambda \in (0,1)$. In (11), $\mathrm{Tr}(\cdot)$ denotes the trace of the covariance matrix (2), and the adaptive rule intends to keep a ratio on gradient estimation error reminiscent of a student T bounded by a constant κ_{ag} and β_g power of the TR radius. Both (11) and (12) ensure that the standard error of the function estimate at each design point is kept in lock-step with the β_f-th power of the TR radius. We note that $\beta_f = 2$ in the original version of this algorithm [5]. In both cases, the slow logarithmic deflation of the right-hand-side thresholds by $1/\sqrt{\lambda_k}$ ensures that the sample sizes do not stay small due to chance. The difficulty of analyzing the algorithms ASTRO and ASTRO-DF thereof is due to the complexity of analyzing moments estimation errors $\bar{\boldsymbol{E}}_k^g(N_k)$ and $\bar{E}_k^i(N_k^i)$.

For ease of exposition, we will drop $\boldsymbol{X}_k$ from function and gradient estimates and replace $\bar{F}(\boldsymbol{X}_k,n)$, $\bar{F}(\widetilde{\boldsymbol{X}}_{k+1},n)$, and $\bar{\boldsymbol{G}}(\boldsymbol{X}_k,n)$ with $\bar{F}_k^0(n)$, $\bar{F}_k^s(n)$, and $\bar{\boldsymbol{G}}_k(n)$. Hence, the ASTRO and ASTRO-DF algorithm listings are given in Algorithm 1 and Algorithm 2, respectively, following the iterative steps (a)-(d) in Section 1.2. The notable differences between the two algorithms are in model construction and adaptive sample size.

Algorithm 1 ASTRO – Adaptive Sampling Trust-region Optimization

Require: Initial guess $\boldsymbol{x}_0 \in \mathbb{R}^d$, initial and maximum radius $\Delta_{\max} > \Delta_0 > 0$, model "fitness" threshold $\eta \in (0,1)$, expansion and shrinkage constants $\gamma_1 > 1 > \gamma_2 > 0$, lower bound λ_k, adaptive sampling constants $\kappa_{ag}, \kappa_{af} > 0$, and criticality threshold $\mu > 0$.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: *Model Construction:* Obtain $\bar{F}_k^0(N_k)$ and $\boldsymbol{G}_k = \bar{\boldsymbol{G}}_k(N_k)$ with sample size N_k following from (11), and build a local model M_k as defined in (8).

3: *Subproblem Minimization:* Approximate the k-th step by minimizing the model inside the TR with $\boldsymbol{S}_k = \operatorname{argmin}_{\|\boldsymbol{s}\| \le \Delta_k} M_k(\boldsymbol{X}_k + \boldsymbol{s})$, and set $\widetilde{\boldsymbol{X}}_{k+1} = \boldsymbol{X}_k + \boldsymbol{S}_k$.

4: *Candidate Evaluation:* Estimate the function at the candidate point using adaptive sampling to obtain $\bar{F}_k^s(\widetilde{N}_{k+1})$. Compute the success ratio

$$\hat{\rho}_k = \frac{\bar{F}_k^s(\widetilde{N}_{k+1}) - \bar{F}_k^0(N_k)}{M_k(\widetilde{\boldsymbol{X}}_{k+1}) - M_k(\boldsymbol{X}_k)}. \tag{13}$$

5: *Update:* Set

$$(\boldsymbol{X}_{k+1}, \Delta_{k+1}) = \begin{cases} (\widetilde{\boldsymbol{X}}_{k+1}, \gamma_1 \Delta_k \wedge \Delta_{\max}) & \text{if } \hat{\rho}_k > \eta \text{ and } \mu \|\boldsymbol{G}_k\| \ge \Delta_k, \\ (\boldsymbol{X}_k, \Delta_k \gamma_2) & \text{otherwise.} \end{cases}$$

Set $k = k + 1$.

6: **end for**

Algorithm 2 ASTRO – Derivative-free (ASTRO-DF)

Require: Same as ASTRO.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: *Model Construction:* Select the design set $\mathcal{X}_k$, for all $i = 0, 1, \cdots, p$ estimate $\bar{F}_k^i(N_k^i)$, and build a local model M_k following Definition 1.

3: *Subproblem Minimization:* Same as ASTRO, find $\widetilde{\boldsymbol{X}}_{k+1} = \boldsymbol{X}_k + \boldsymbol{S}_k$.

4: *Candidate Evaluation:* Estimate $\bar{F}_k^s(\widetilde{N}_{k+1})$ and compute the success ratio (13).

5: *Update:* Same as ASTRO. Set $k = k + 1$.

6: **end for**

4.1 Stochastic noise

The first main result is to control the stochastic error by keeping $\bar{F}_k^0(N_k)$ bounded wp1. We will now show that this result holds with a logarithmic λ_k.

Theorem 2. *Let $\{\boldsymbol{X}_k\}$ be the sequence of solutions generated by Algorithm 1 or 2. Let Assumptions 2 and 4 hold and set λ_k such that $\mathcal{O}(\lambda_k) = (\log\ k)^{1+\epsilon_\lambda}$ for some $\epsilon_\lambda \in (0,1)$. Then $\mathbb{P}\{\liminf_{k\to\infty} \bar{F}_k^0(N_k) = -\infty\} = 0$. Moreover, $\mathbb{P}\{|\bar{E}_k(N_k)| \ge c_f \Delta_k^{\beta_f}\ \ i.o.\} = 0$.*

Proof. Given that f is bounded from below, the postulate will hold if and only if we can show $\mathbb{P}\{|\bar{E}_k(N_k)| > c_f \text{ i.o.}\} = 0$ for any $c_f > 0$. Observe from Theorem 2.7 and 2.8 in [5] that

$\hat{\sigma}_F(\boldsymbol{X}_k, N_k)/\sigma_F(\boldsymbol{X}_k) \xrightarrow{wp1} 1$ as $k \to \infty$. Then for a fixed $c > 0$ and large enough k, we have

$$\begin{aligned}
\mathbb{P}\{|\bar{E}_k(N_k)| > c_f \mid \mathcal{F}_{k-1}\} &\leq \mathbb{P}\left\{\sup_{n \geq \lambda_k} |\bar{E}_k(n)| > c_f \mid \mathcal{F}_{k-1}\right\} \\
&\leq \sum_{n \geq \lambda_k} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} |E_{k,j}| > c_f \mid \mathcal{F}_{k-1}\right\} \\
&\leq \sum_{n \geq \lambda_k} 2e^{\left(-n\frac{c_f^2}{2(c_f b_f + \sigma_f^2)}\right)}. \qquad (14)
\end{aligned}$$

Then for iterates generated from Algorithm 1 or 2 knowing that $N_k \geq \frac{\sigma_{mf}^2 \lambda_k}{2\kappa_{af}^2 \Delta_k^{2\beta_f}}$ for large enough k, where σ_{mf} is such that $\sigma_{mf}^2 \leq \inf_{\boldsymbol{x} \in \mathbb{R}^d} \sigma_F^2(\boldsymbol{x}) \leq \sigma_f^2$, we can write

$$\begin{aligned}
\sum_{n \geq \lambda_k} 2\exp\left(-n\frac{c_f^2}{2(c_f b_f + \sigma_f^2)}\right) &\leq \sum_{n \geq \sigma_{mf}^2 \lambda_k (\sqrt{2}\kappa_{af} \Delta_k^{\beta_f})^{-2}} 2\exp\left(-n\frac{c_f^2}{2(c_f b_f + \sigma_f^2)}\right) \\
&\leq 2\frac{cb + \sigma^2}{c^2}\mathbb{P}\{\mathrm{Exp}(\frac{c^2}{2(cb+\sigma^2)}) \geq \lambda_k\} \\
&\leq 2\frac{c_f b_f + \sigma_f^2}{c_f^2}\exp\left(-\lambda_k \frac{c_f^2}{c_f b_f + \sigma_f^2}\frac{\sigma_{mf}^2}{2\kappa_{af}^2 \Delta_k^{2\beta_f}}\right) \\
&\leq 2\frac{c_f b_f + \sigma_f^2}{c_f^2}\frac{2\kappa_{af}^2 \Delta_{\max}^{2\beta_f}}{\sigma_{mf}^2} k^{-(1+\epsilon_\lambda)\frac{c_f^2}{c_f b_f + \sigma_f^2}\frac{\sigma_{mf}^2}{2\kappa_{af}^2 \Delta_{\max}^{2\beta_f}}}.
\end{aligned}$$

Using Borel-Cantelli's Lemma for martingales and observing that the right hand side of (14) is summable in k (as long as the power of k is less than -1), we conclude that $\liminf_{k\to\infty} |\bar{E}_k(N_k)| < \infty$ almost surely. We also observe that $\mathbb{P}\{|\bar{E}_k(N_k)| > c\Delta_k^{\beta_f} \text{ i.o.}\} = 0$ since the $\Delta_k^{2\beta_f}$ in the enumerator of the exponent in the third inequality will ultimately get cancelled out with that in the denominator. □

Remark 1. *Because of the way N_k is defined, we observe from the proof of Theorem 2 that $\mathbb{P}\{|\bar{E}_k(N_k)| \geq c_\lambda \Delta_k^2 \ |\mathcal{F}_k\} \leq \lambda_k^{-1}$ for some $c_\lambda > 0$; i.e., the estimate is accurate with probability $1 - \lambda_k^{-1}$ that tends to one as $k \to \infty$. This is in contrast to the assumption of probabilisitcally accurate estimates with a fixed probability [4] in the STROM algorithm. Although a fixed probability for accuracy of the estimated values seems less stringent than an increasing probability of accurate estimates, the latter will more carefully keep the total work involved at bay. We will show in this paper that we can obtain a canonical work complexity under certain assumptions with this property. A recent study [16] suggests that STORM enjoys a $\widetilde{\mathcal{O}}(\epsilon^{-6})$ complexity bound for the first-order stochastic oracles barring any dependence between the stochastic models, the functions estimates, the function estimates, and the history. In contrast, we will show that ASTRO can enjoy a stronger complexity analysis and ASTRO-DF will be at least as good as STORM but by exploiting the dependence and random number generation, it can indeed have an improved work complexity of $\widetilde{\mathcal{O}}(\epsilon^{-5})$ or better.*

Corollary 3. *Let Assumptions 3 and 4 hold. Let $\{\boldsymbol{X}_k\}$ be the sequence of solutions generated by Algorithm 1 and set λ_k such that $\mathcal{O}(\lambda_k) = (\log\ k)^{1+\epsilon_\lambda}$ for some $\epsilon_\lambda \in (0,1)$. Then for any $c_g > 0$*

$$\mathbb{P}\{\|\bar{\boldsymbol{E}}_k^g(N_k)\| > c_g\Delta_k^{\beta_g}\ \ i.o.\} = 0 \text{ for } \beta_g = 0, 1. \tag{15}$$

Proof. We begin by noticing that from Assumption 3

$$\mathbb{P}\{\|\bar{\boldsymbol{E}}_k^g(N_k)\| > c_g\Delta_k^{\beta_g} \mid \mathcal{F}_{k-1}\} \leq \sum_{i=1}^{d} \mathbb{P}\left\{|[\bar{\boldsymbol{E}}_k^g(N_k)]_i| > \frac{c_g\Delta_k^{\beta_g}}{d} \mid \mathcal{F}_{k-1}\right\},$$

where $[\bar{\boldsymbol{E}}_k^g(N_k)]_i$ is the i-th element in the stochastic gradient error. Following the same steps in the proof for Theorem 2, (15) holds with $N_k \geq \sigma_{mg}^2\lambda_k(2\kappa_{af}^2\Delta_k^{2\beta_g})^{-1}$ for large enough k, where σ_{mg} is such that $\sigma_{mg}^2 \leq \inf_{\boldsymbol{x}\in\mathbb{R}^d} \mathrm{Tr}(\sigma_{\boldsymbol{G}}^2(\boldsymbol{x})) \leq d\sigma_g^2$. □

Lemma 3. *Let $\{\boldsymbol{X}_k\}$ be the sequence of solutions generated by Algorithm 1 or Algorithm 2. Let Assumptions 2-4 hold and set λ_k such that $\mathcal{O}(\lambda_k) = (\log\ k)^{1+\epsilon_\lambda}$ for some $\epsilon_\lambda \in (0,1)$. Then given any $c_{fd} > 0$ we obtain that*

$$\mathbb{P}\{|\bar{E}_k(N_k) - \bar{E}(\boldsymbol{x}, N_k)| \geq c_{fd}\Delta_k^2\ \ i.o.\} = 0,$$

for any $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{X}_k; \Delta_k)$ with CRN and either of the following

(i)^{df} $\beta_f = 3/2$;

(ii)^{df} $\beta_f = 1$ and $F(\cdot, \xi) \in \mathcal{C}(\mathbb{R}^d)$ for each $\xi \in \Xi$.

Moreover, with ASTRO, we have that given any $c_{gd} > 0$

$$\mathbb{P}\{\|\bar{\boldsymbol{E}}_k^g(N_k) - \bar{\boldsymbol{E}}^g(\boldsymbol{x}, N_k)\| \geq c_{gd}\Delta_k\ \ i.o.\} = 0, \tag{16}$$

for any $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{X}_k; \Delta_k)$ with CRN and either of the following

(i) $\beta_g = 1/2$;

(ii) $\beta_g = 0$ and $\boldsymbol{G}(\cdot, \xi) \in \mathcal{C}(\mathbb{R}^d)$ for each $\xi \in \Xi$.

Proof. We first know that Assumption 2 holds for $E_{k,j}(\boldsymbol{X}_k) - E_{k,j}(\boldsymbol{x})$ with the parameter $2\sigma_f^2$ for any $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{X}_k; \Delta_k)$ and $j \in \mathbb{N}$. The proof is completed trivially using the triangle inequality. Given that N_k solely relies on the estimates at $\boldsymbol{X}_k$ and not on $\boldsymbol{x}$, it can be naturally concluded that Assumption 4 also applies to $E_{k,j}(\boldsymbol{X}_k) - E_{k,j}(\boldsymbol{x})$. Now we obtain that from Theorem 1 for any $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{X}_k; \Delta_k)$ with CRN,

$$\mathrm{Var}\left(\bar{E}_k(N_k) - \bar{E}(\boldsymbol{x}, N_k)\right) = \begin{cases} \mathcal{O}(\Delta_k^2) & \text{if } F(\cdot, \xi) \in \mathcal{C}(\mathbb{R}^d) \text{ for each } \xi \in \Xi; \\ \mathcal{O}(\Delta_k) & \text{otherwise.} \end{cases} \tag{17}$$

Let us first consider the case under condition $(i)^{df}$. We know from (17) that $2\sigma_f^2 \leq c_{\sigma 1}\Delta_k$ for some $c_{\sigma 1} > 0$. Then by substituting $c_{\sigma 1}\Delta_k$ in place of σ_f^2 and continuing the inequality in (14), we can show that the theorem is satisfied. For the case under condition $(ii)^{df}$, we can also obtain the same result with $2\sigma_f^2 \leq c_{\sigma 2}\Delta_k^2$ for some $c_{\sigma 2} > 0$. The same steps with Assumption 3 yield (16) under any one of conditions (i) and (ii). □

5. CONSISTENCY

In this section, we prove the strong consistency of ASTRO and ASTRO-DF for four cases where the power of the trust-region size in (11) and (12) can vary as a result of lower variance in the model gradient estimate obtained by common random numbers and continuity assumption of the function sample-paths. Before delving into the strong consistency, let us introduce an assumption concerning the random gradient observations paths. This assumption will help us specify one of the four cases for ASTRO.

Assumption 5. *(κ_{ubg}: path-wise gradient Lipschitz constants) The sample-path function $F(\cdot,\xi)$ is differentiable $\forall \boldsymbol{x} \in \mathbb{R}^d$, wp1. The sample-path gradient function $\boldsymbol{G}(\cdot,\xi)$ is $\kappa_{lcG}(\xi)$-Lipschitz in $\mathbb{R}$, i.e., $\|\boldsymbol{G}(\boldsymbol{x}_1,\xi) - \boldsymbol{G}(\boldsymbol{x}_2,\xi)\| \leq \kappa_{lcG}(\xi)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|,\ \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$.*

5.1 Almost sure convergence for ASTRO and ASTRO-DF

In both ASTRO and ASTRO-DF, the subproblem's recommended solution is required to obtain sufficient model reduction. This is formally stated by the following assumption.

Assumption 6. *(κ_{fcd}: fraction of the Cauchy decrease) There exists a constant $\kappa_{fcd} \in (0,1]$ for all k such that*

$$M_k(\boldsymbol{X}_k) - M_k(\widetilde{\boldsymbol{X}}_{k+1}) \geq \kappa_{fcd}[M_k(\boldsymbol{X}_k) - M_k(\boldsymbol{X}_k + \boldsymbol{S}_k^c)],$$

where $\boldsymbol{S}_k^c$ is the Cauchy step.

Local model built on quadratic approximation in ASTRO and ASTRO-DF rely on the classic assumptions on model Hessian below.

Assumption 7. *(ASTRO Hessian) $\|\mathsf{B}_k\| \leq \kappa_{\mathsf{B}}\ \forall k$ and some $\kappa_{\mathsf{B}} \in (0,\infty)$ wp1.*

Assumption 8. *(ASTRO-DF Hessian) $\|\mathsf{H}_k\| \leq \kappa_{\mathsf{H}}\ \forall k$ and some $\kappa_{\mathsf{H}} \in (0,\infty)$ wp1.*

The following result characterizes the stochastic model error with estimation error. This result with Theorem 2 ensures that, given a sufficiently large number of iterations k, the local model becomes the stochastic fully linear model almost surely, i.e., achieving (9) wp1.

Lemma 4. *(Lemma 2.9 in [5]) Let $\mathcal{X} = \{\boldsymbol{X}^0, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^p\}$ be a Λ-poised set on $\mathcal{B}(\boldsymbol{X}^0;\Delta)$ and let Assumption 1 hold. Let $m(\cdot)$ be a polynomial interpolation model of f on $\mathcal{B}(\boldsymbol{X}^0;\Delta)$. Let $M(\cdot)$ be the corresponding stochastic polynomial interpolation model of f on $\mathcal{B}(\boldsymbol{X}^0;\Delta)$ constructed using observations $\bar{F}(\boldsymbol{X}^i, n(\boldsymbol{X}^i)) = f(\boldsymbol{X}^i) + \bar{E}^i$ for $i = 0,1,\ldots,p$.*

(i) $|M(\boldsymbol{x}) - m(\boldsymbol{x})| \leq (p+1)\Lambda \max_{i=0,1,\ldots,p} |\bar{F}(\boldsymbol{X}^i, n(\boldsymbol{X}^i)) - f(\boldsymbol{X}^i)|\ \forall \boldsymbol{x} \in \mathcal{B}(\boldsymbol{X}^0;\Delta)$;

(ii) There exist constants $\kappa_{eg1} > 0$ and $\kappa_{eg2} > 0$ such that for $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{X}^0;\Delta)$,

$$\|\nabla M(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\| \leq \kappa_{eg1}\Delta + \kappa_{eg2}\frac{\sqrt{\sum_{i=1}^p (\bar{E}^i - \bar{E}^0)^2}}{\Delta}. \tag{18}$$

We now provide the convergence theory for ASTRO and ASTRO-DF under any one of the four conditions (Table 2). Throughout the analysis of ASTRO and ASTRO-DF, we will refer to four conditions that play a crucial role in establishing its consistency, iteration complexity, and work complexity.

CRN or IS	With IS	With CRN		
Sample-path property	None	None	Continuous	Smooth
ASTRO	$(a)\beta_f = 2$ $\beta_g = 1$	$(b)\beta_f = 1.5$ $\beta_g = 1$	$(c)\beta_f = 1$ $\beta_g = 1$	$(d)\beta_f = 0$ $\beta_g = 0$
ASTRO-DF	$(a)^{df}\beta_f = 2$	$(b)^{df}\beta_f = 1.5$	$(c)^{df}\beta_f = 1$	$(d)^{df}\beta_f = 1$

TABLE 2: Sampling rule conditions for ASTRO and ASTRO-DF.

Theorem 4 (Strong Consistency)**.** *Considering the sampling conditions in Table 2, the sequence* $\{\boldsymbol{X}_k\}$ *of iterates satisfies the following wp1:*

$$\liminf_{k\to\infty} \|\nabla f(\boldsymbol{X}_k)\| = 0; \tag{19}$$

$$\lim_{k\to\infty} \|\nabla f(\boldsymbol{X}_k)\| = 0, \tag{20}$$

- *for ASTRO: if Assumptions 2-6, and 7 hold and under conditions* (a)*-*(d)*.*
- *for ASTRO-DF: if Assumptions 1, 2, 4, 6, and 8 hold and under conditions* $(a)^{df}$*-*$(d)^{df}$*.*

Before delving into the proof of Theorem 4 we make some observations.

(*sc*-a) An important result needed for the strong consistency is showing that if the trust-region radius becomes too small with respect to the gradient (of the function or model for ASTRO and ASTRO-DF respectively), then a success event and progress in optimization must occur leading to expansion in the trust-region in the proceeding iteration. This will lead to stating that before $\|\nabla f(\boldsymbol{X}_k)\|$ hits an ϵ-distance from 0, the TR radius remains bounded below as a function of ϵ. For both ASTRO and ASTRO-DF, we demonstrate this in Section 5.2, specifically in Lemma 5 and the proof for ASTRO-DF, respectively.

(*sc*-b) In the conditions, β_g and β_f represent the order of sampling size in (11) and (12), exerting a substantial influence on the work complexity. When considering ASTRO, β_f takes precedence over β_g across all conditions (a)-(d) (Table 2). This observation naturally leads to the conclusion that the work complexity of ASTRO is primarily determined by β_f rather than β_g. This aspect will be further discussed in Section 6.3.

5.2 Proof of Theorem 4

We first prove the wp1 convergence of ASTRO, i.e., Theorem 4. Our first result for consistency of ASTRO proves that iterations where the TR radius becomes small compared to the estimated gradient will become very successful iterations wp1.

Lemma 5. *Let Assumptions 2-6, and 7 hold. Define the set*

$$\mathcal{V}_1 := \left\{\omega : \exists\ \{k_j\}\ s.t.\ \left(\Delta_{k_j}(\omega) \le \frac{\|\bar{\boldsymbol{G}}_{k_j}(N_{k_j})\|\kappa_{fcd}(1-\eta)}{\kappa_{Lg}+\kappa_{\mathsf{B}}}\right)\bigcap\left(\hat{\rho}_{k_j}(\omega) < \eta\right)\right\}.$$

Then $\mathbb{P}\{\mathcal{V}_1\} = 0$ *under any one of the conditions* (a)*-*(d)*.*

Proof. First, let us delve into the cases marked as conditions (a)-(c). For the purpose of deriving a contradiction, suppose that the set $\mathcal{V}_1$ has positive measure, and let $\omega \in \mathcal{V}_1$. We suppress ω in the following statements for ease of notation. Recall the stochastic model defined on Step 2 of Algorithm 1,

$$M_k(\widetilde{\boldsymbol{X}}_{k+1}) = \bar{F}_k^0(N_k) + \bar{\boldsymbol{G}}_k(N_k)^\intercal \boldsymbol{S}_k + \frac{1}{2}\boldsymbol{S}_k^\intercal \mathsf{B}_k \boldsymbol{S}_k. \tag{21}$$

where the step size, $\boldsymbol{S}_k$, satisfies $\|\boldsymbol{S}_k\| \leq \Delta_k$, for all k. Now, we see that

$$\begin{aligned} \bar{F}_k^s(\widetilde{N}_{k+1}) = \bar{F}_k^0(N_k) + \nabla f(\boldsymbol{X}_k)^\intercal \boldsymbol{S}_k + \int_0^1 \left[\nabla f(\boldsymbol{X}_k + \tau \boldsymbol{S}_k) - \nabla f(\boldsymbol{X}_k)\right]^\intercal \boldsymbol{S}_k \, d\tau \\ + \bar{E}_k^s - \bar{E}_k^0 + \bar{\boldsymbol{G}}_k(N_k)^\intercal \boldsymbol{S}_k - \bar{\boldsymbol{G}}_k(N_k)^\intercal \boldsymbol{S}_k, \end{aligned} \tag{22}$$

where $\bar{E}_k^s = \bar{F}_k^s(\widetilde{N}_{k+1}) - f(\widetilde{\boldsymbol{X}}_{k+1})$ and $\bar{E}_k^0 = \bar{F}_k^0(N_k) - f(\boldsymbol{X}_k)$. For sufficiently large k and any $c_g > 0$ and $c_{fd} > 0$, (21) and (22) along with the triangle inequality imply that wp1:

$$\begin{aligned} \left|\bar{F}_k^s(\widetilde{N}_{k+1}) - M_k(\widetilde{\boldsymbol{X}}_{k+1})\right| &\leq \left| \int_0^1 [\nabla f(\boldsymbol{X}_k + \tau \boldsymbol{S}_k) - \nabla f(\boldsymbol{X}_k)]^\intercal \boldsymbol{S}_k \, d\tau - \frac{1}{2}\boldsymbol{S}_k^\intercal \mathsf{B}_k \boldsymbol{S}_k \right| \\ &\quad + |\bar{E}_k^s - \bar{E}_k^0| + |[\nabla f(\boldsymbol{X}_k) - \bar{\boldsymbol{G}}_k(N_k)]^\intercal \boldsymbol{S}_k| \\ &\leq \int_0^1 \left\|\nabla f(\boldsymbol{X}_k + \tau \boldsymbol{S}_k) - \nabla f(\boldsymbol{X}_k)\right\| \|\boldsymbol{S}_k\| d\tau + \frac{1}{2}\|\boldsymbol{S}_k\|^2 \kappa_{\mathsf{B}} \\ &\quad + |\bar{E}_k^s - \bar{E}_k^0| + \|\nabla f(\boldsymbol{X}_k) - \bar{\boldsymbol{G}}_k(N_k)\| \|\boldsymbol{S}_k\| \\ &\leq \int_0^1 \kappa_{Lg} \|\boldsymbol{S}_k\|^2 \tau d\tau + \frac{1}{2}\kappa_{\mathsf{B}} \Delta_k^2 + c_{fd}\Delta_k^2 + c_g \Delta_k \|\boldsymbol{S}_k\| \\ &\leq \frac{1}{2}(\kappa_{Lg} + \kappa_{\mathsf{B}} + 2c_{fd} + 2c_g)\Delta_k^2, \end{aligned} \tag{23}$$

where the first inequality follows from the Cauchy-Schwartz, the second inequality follows from Assumption 7, the third inequality follows from Theorem 2, 3, and 3, and Assumption 1. Without loss of generality, we can set $2c_g + 2c_{fd} = \kappa_{Lg} + \kappa_{\mathsf{B}}$. Define the set

$$\mathcal{D}_1 := \left\{ |\bar{F}_k^s(N_k) - M_k(\widetilde{\boldsymbol{X}}_{k+1})| \leq \left(\kappa_{Lg} + \kappa_{\mathsf{B}}\right)\Delta_k^2, \text{ for large enough } k \right\},$$

where $\mathbb{P}\{\mathcal{D}_1\} = 1$. Select $\omega \in \mathcal{D}_1 \cap \mathcal{V}_1$ for the following arguments. Recall $\kappa_{fcd} \in (0, 1]$ and $\eta \in (0, 1)$, and so, $\kappa_{fcd}(1-\eta) < 1$. For a subsequence $\{k_j\}$ that meets the criteria in the definition of $\mathcal{V}_1$, we can write

$$\Delta_{k_j} \leq \frac{\|\bar{\boldsymbol{G}}_{k_j}(N_{k_j})\| \kappa_{fcd}(1-\eta)}{\kappa_{Lg} + \kappa_{\mathsf{B}}} < \frac{\|\bar{\boldsymbol{G}}_{k_j}(N_{k_j})\|}{\kappa_{Lg} + \kappa_{\mathsf{B}}} \leq \frac{\|\bar{\boldsymbol{G}}_{k_j}(N_{k_j})\|}{\|\mathsf{B}_{k_j}\|}. \tag{24}$$

We then observe that

$$\begin{aligned} |\hat{\rho}_{k_j} - 1| &= \left| \frac{\bar{F}_{k_j}^s(N_{k_j}) - M_{k_j}(\widetilde{\boldsymbol{X}}_{k_j+1})}{M_{k_j}(\boldsymbol{X}_{k_j}) - M_{k_j}(\widetilde{\boldsymbol{X}}_{k_j+1})} \right| \\ &\leq \frac{(\kappa_{Lg} + \kappa_{\mathsf{B}})\Delta_{k_j}^2}{\kappa_{fcd}\|\bar{\boldsymbol{G}}_{k_j}(N_{k_j})\| \left(\frac{\|\bar{\boldsymbol{G}}_{k_j}(N_{k_j})\|}{\|\mathsf{B}_{k_j}\|} \wedge \Delta_{k_j} \right)} \leq \frac{(\kappa_{Lg} + \kappa_{\mathsf{B}})\Delta_{k_j}}{\kappa_{fcd}\|\bar{\boldsymbol{G}}_{k_j}(N_{k_j})\|} \leq \frac{(\kappa_{Lg} + \kappa_{\mathsf{B}})(1-\eta)}{\kappa_{Lg} + \kappa_{\mathsf{B}}}, \end{aligned} \tag{25}$$

where the equality follows from the fact we assumed $M_k(\boldsymbol{X}_k) = \bar{F}_k^0(N_k)$ for all k, the first inequality follows from the Cauchy decrease condition (10) and the definition of $\mathcal{D}_1$, and the second inequality follows from (24). Thus, we can conclude from (25) that $\hat{\rho}_{k_j} \geq \eta$, whereas $\omega \in \mathcal{V}_1$ implies that $\hat{\rho}_{k_j} < \eta$. Therefore, $\mathbb{P}\{\mathcal{V}_1\} = 0$, and hence, the assertion of the theorem holds.

Now let us explore the situation, where the sample-path gradient function is Lipschitz continuous, which corresponds to the condition (d). By Assumption 5 and the fact that $\bar{F}_k^0(N_k)$ and $\bar{F}_k^s(N_k)$ are constructed as the sample mean of N_k i.i.d copies of $F(\boldsymbol{X}_k, \xi)$ and $F(\widetilde{\boldsymbol{X}}_{k+1}, \xi)$ with the same $\xi_1, \ldots, \xi_{N_k}$ (CRN), we see that

$$\bar{F}_k^s(N_k) = \bar{F}_k^0(N_k) + \bar{\boldsymbol{G}}_k(N_k)^\intercal \boldsymbol{S}_k + \int_0^1 \left[\bar{\boldsymbol{G}}(\boldsymbol{X}_k + \tau \boldsymbol{S}_k, N_k) - \bar{\boldsymbol{G}}_k(N_k)\right]^\intercal \boldsymbol{S}_k \, d\tau. \tag{26}$$

For sufficiently large k and given any $c_{gd} > 0$, (21) and (26) along with the triangle inequality imply that wp1:

$$\begin{aligned}
\left|\bar{F}_k^s(N_k) - M_k(\widetilde{\boldsymbol{X}}_{k+1})\right| &= \left| \int_0^1 [\bar{\boldsymbol{G}}(\boldsymbol{X}_k + \tau \boldsymbol{S}_k, N_k) - \bar{\boldsymbol{G}}_k(N_k)]^\intercal \boldsymbol{S}_k d\tau - \frac{1}{2} \boldsymbol{S}_k^\intercal \mathsf{B}_k \boldsymbol{S}_k \right| \\
&\leq \int_0^1 \left\|\bar{\boldsymbol{G}}(\boldsymbol{X}_k + \tau \boldsymbol{S}_k, N_k) - \bar{\boldsymbol{G}}_k(N_k)\right\| \left\|\boldsymbol{S}_k\right\| d\tau + \frac{1}{2} \|\boldsymbol{S}_k\|^2 \kappa_{\mathsf{B}} \\
&\leq \int_0^1 \left\|\nabla f(\boldsymbol{X}_k + \tau \boldsymbol{S}_k) - \nabla f(\boldsymbol{X}_k)\right\| \|\boldsymbol{S}_k\| d\tau \\
&\qquad + \int_0^1 \left\|\bar{\boldsymbol{E}}^g(\boldsymbol{X}_k + \tau \boldsymbol{S}_k, N_k) - \bar{\boldsymbol{E}}^g(\boldsymbol{X}_k, N_k)\right\| \left\|\boldsymbol{S}_k\right\| d\tau + \frac{1}{2} \|\boldsymbol{S}_k\|^2 \kappa_{\mathsf{B}} \\
&\leq \int_0^1 (\kappa_{Lg} + c_{gd}) \|\boldsymbol{S}_k\|^2 \tau d\tau + \frac{1}{2} \kappa_{\mathsf{B}} \Delta_k^2 \leq \frac{1}{2} (\kappa_{Lg} + c_{gd} + \kappa_{\mathsf{B}}) \Delta_k^2,
\end{aligned}$$

where the second inequality follows from Assumption 5 and Theorem 3. Hence, we have a similar results with (23) under condition (d) by defining $c_{gd} := \kappa_{Lg} + \kappa_{\mathsf{B}}$. Afterward, the remainder of the proof follows a similar line of reasoning as the previous cases mentioned under condition (a)-(c), albeit with one crucial modification: the utilization of κ_{ubg} in place of κ_{Lg}. □

The next corollary asserts that for iterates with true gradient larger than ϵ, we can characterize a bound for Δ_k in terms of ϵ. This corollary guarantees that if the gradient estimate is bounded away from zero, the TR radius cannot be too small, wp1Thus, as long as a sequence, $\{\boldsymbol{X}_k\}$, generated by ASTRO is not close to a first-order critical point, the size of the search space will not crash to zero. This result will set the ground for ASTRO's global convergence theorem.

Corollary 5. *Let Assumptions 2-6, and 7 hold. For some $\kappa_{lbg} > 0$, define the sets*

$$\begin{aligned}
\mathcal{V}_2 &:= \{\omega : \ \exists \ \epsilon_0(\omega) > 0 \ s.t. \ \|\nabla f(\boldsymbol{X}_k(\omega))\| \geq \epsilon_0 \ \forall \ k\}, \\
\mathcal{V}_3 &:= \left\{\omega : \exists \ \{k_j\} \ s.t. \left(\|\bar{\boldsymbol{G}}_{k_j}(N_{k_j}(\omega))\| \geq \kappa_{lbg}\right) \bigcap \left(\Delta_{k_j}(\omega) < \kappa_{lbd} := \frac{\gamma_2 \kappa_{fcd} \kappa_{lbg}(1-\eta)}{1 + \kappa_{Lg} + \kappa_{\mathsf{B}}}\right)\right\}.
\end{aligned} \tag{27}$$

Then $\mathbb{P}\{\mathcal{V}_2 \cap \mathcal{V}_3\} = 0$ under any one of the conditions (a)-(d).

Proof. We begin by noticing that $\lim_{k\to\infty} \|\bar{\boldsymbol{G}}_k(N_k) - \nabla f(\boldsymbol{X}_k)\| = 0$, almost surely from Theorem 3. Now, assume by contradiction that the set $\mathcal{V}_2 \cap \mathcal{V}_3$ has a positive measure. Let $\omega \in \mathcal{V}_2 \cap \mathcal{V}_3$ and $\epsilon_0(\omega)$ and $\{k_j(\omega)\}$ be as defined. From the fact that $\lim_{k\to\infty} \|\bar{\boldsymbol{G}}_k(N_k) - \nabla f(\boldsymbol{X}_k)\| = 0$ wp1, there exists $K(\omega) > 0$ such that $\left\|\bar{\boldsymbol{G}}_k(N_k(\omega))\right\| \geq \epsilon_0(\omega)/2$ for all $k \geq K(\omega)$. Now let $\kappa_{lbg} = \epsilon_0(\omega)/2$ and choose $t > K(\omega)$ such that $t \notin \{k_j(\omega)\}$ but $t+1 \in \{k_j(\omega)\}$. Therefore the following hold:

$$\begin{aligned} \left\|\bar{\boldsymbol{G}}_t(N_t(\omega))\right\| \geq \kappa_{lbg}; \quad \Delta_t(\omega) \geq \kappa_{lbd} \\ \left\|\bar{\boldsymbol{G}}_{t+1}(N_{t+1}(\omega))\right\| \geq \kappa_{lbg}; \quad \Delta_{t+1}(\omega) < \kappa_{lbd}. \end{aligned} \tag{28}$$

From (28) we observe that $\Delta_t(\omega) > \Delta_{t+1}(\omega)$. On the other hand, from Step 5 of Algorithm 1 we know it is true that

$$\Delta_t(\omega) \leq \frac{\Delta_{t+1}(\omega)}{\gamma_2} \leq \frac{\kappa_{fcd}(1-\eta)}{1+\kappa_{Lg}+\kappa_{\mathsf{B}}} \left\|\bar{\boldsymbol{G}}_t(N_t(\omega))\right\| \Rightarrow \frac{\Delta_t(\omega)}{\left\|\bar{\boldsymbol{G}}_t(N_t(\omega))\right\|} \leq \frac{\kappa_{fcd}(1-\eta)}{1+\kappa_{Lg}+\kappa_{\mathsf{B}}}.$$

Given that Lemma 5 holds true for any of the conditions, we must have $\omega \in \mathcal{V}_1^c$. This implies that $\hat{\rho}_t(\omega)$ is greater than η, indicating the success of iteration t. Consequently, we have $\Delta_t(\omega)$ being less than $\Delta_{t+1}(\omega)$, leading to a contradiction. Therefore, we can conclude that $\mathbb{P}\{\mathcal{V}_2 \cap \mathcal{V}_3\} = 0$ holds true under any one of the conditions (a)-(d). □

We have now reached a point where we can confidently establish the wp1 convergence of ASTRO. The following proof solidifies our claim.

Proof. [of Theorem 4 for ASTRO.] We first proceed to prove (19) by contradiction under any one of the conditions (a)-(c). We need to assume at least one sample path for which the true gradient is bounded from below for all k. We observe that this is the same as assuming $\mathcal{V}_2$, as defined in the statement of Corollary 5 has a positive probability. From Lemma 5 and Corollary 5, we observe that $\mathcal{V}_1^c \cap \mathcal{V}_2 \cap \mathcal{V}_3^c$ has a nonzero measure. We let $\omega_0 \in \mathcal{V}_1^c \cap \mathcal{V}_2 \cap \mathcal{V}_3^c$ and $\epsilon_0(\omega_0)$ as defined in (27), and we suppress ω_0 in the following statements for ease of notation. Since $\lim_{k\to\infty} \|\bar{\boldsymbol{G}}_k(N_k) - \nabla f(\boldsymbol{X}_k)\| = 0$ wp1, we can find $K_1 > 0$ such that $\|\bar{\boldsymbol{G}}_k(N_k)\| \geq \epsilon_1 > \epsilon_0$ for all $k \geq K_1$. Moreover, $\omega_0 \in \mathcal{V}_3^c$ implies that there exists $K_2 > 0$ such that for all $k \geq K := \max\{K_1, K_2\}$,

$$\Delta_k \geq \frac{\gamma_2 \kappa_{fcd}(1-\eta)}{1+\kappa_{Lg}+\kappa_{\mathsf{B}}} \epsilon_1. \tag{29}$$

We now split the analysis in two cases, depending on whether the choice of ω_0 has finite or infinite number of successful iterations.

Case I – ω_0 has finitely many success events: Let $K_3 > 0$ be such that $\Delta_k < \Delta_{k+1}$ for all $k \geq K_3$. This implies that $\Delta_k \to 0$ for $k \geq \max\{K, K_3\}$ which contradicts (29).

Case II – ω_0 has an infinite number of successful iterations: Let $\mathcal{S} = \{k : \hat{\rho}_k \geq \eta\}$ and in this case $|\mathcal{S}| = \infty$. We observe from Assumption 6 and the constant lower bound on TR in (29) that for all $k \in \mathcal{S}$ and $k \geq K$,

$$\bar{F}_k^0(N_k) - \bar{F}_k^s(\widetilde{N}_{k+1}) \geq \eta \kappa_{fcd} \epsilon_1^2 \min\left\{\frac{1}{1+\kappa_{\mathsf{B}}}, \frac{\gamma_2 \kappa_{fcd}(1-\eta)}{1+\kappa_{Lg}+\kappa_{\mathsf{B}}}\right\}. \tag{30}$$

Let the right-hand-side of (30) be called u and $\mathcal{S}_{K,k}$ be for the set of indices of successful iterations between K and k. Moreover, we know from Theorem 2 that there exists K_4 such that $|\bar{E}_k^0| + |\bar{E}_k^s| <$

$u/2$ for all $k \geq K_4$. Without loss of generality, we can assume that $K_4 > K$. We can then obtain

$$\begin{aligned} u|\mathcal{S}_{K,k}| &\leq \sum_{i \in \mathcal{S}_{K,k}} (\bar{F}_i^0(N_i) - \bar{F}_i^s(\widetilde{N}_{i+1})) = \sum_{i \in \mathcal{S}_{K,k}} (f(\boldsymbol{X}_i) - f(\boldsymbol{X}_{i+1}) + \bar{E}_i^0 - \bar{E}_i^s) \\ &\leq f(\boldsymbol{x}_0) - f^* + \sum_{i \in \mathcal{S}_{K,k}} (|\bar{E}_i^0| + |\bar{E}_i^s|) \\ &\leq f(\boldsymbol{x}_0) - f^* + \sum_{i \in \mathcal{S}_{K,K_4}} (|\bar{E}_i^0| + |\bar{E}_i^s|) + u|\mathcal{S}_{K_4,k}|/2, \\ &\Rightarrow (u - 2c)|\mathcal{S}_{K,k}| \leq f(\boldsymbol{x}_0) - f^*. \end{aligned}$$

where the first inequality comes from (30). We then obtain

$$\frac{u}{2}|\mathcal{S}_{K,k}| \leq f(\boldsymbol{x}_0) - f^* + \sum_{i \in \mathcal{S}_{K,K_4}} (|\bar{E}_i^0| + |\bar{E}_i^s|). \tag{31}$$

As $k \to \infty$ we observe that $|\mathcal{S}_{K,k}| \to \infty$ implying that the right-hand side of (31) diverges as well, which contradicts the statement $f(\boldsymbol{x}_0) - f^* + \sum_{i \in \mathcal{S}_{K,K_4}} (|\bar{E}_i^0| + |\bar{E}_i^s|)$ is finite. Hence, (19) must hold. Similarly, for condition (d), we can obtain same result by applying the same steps, but this time we use $\widetilde{N}_{k+1} = N_k$ for all $k \in \mathbb{N}$.

We now proceed to prove (20) under condition (a). We first need to assume that there is at least a subsequence that has gradients bounded away from zero for contradiction. Particularly, suppose that there exists a set, $\hat{\mathcal{D}}$, of positive measure, $\omega_1 \in \hat{\mathcal{D}}$, $\epsilon_0(\omega_1) > 0$, and a subsequence of successful iterates, $\{t_j(\omega_1)\}$, such that $\|\nabla f(\boldsymbol{X}_{t_j(\omega_1)}(\omega_1))\| > 2\epsilon_0(\omega_1)$, for all $j \in \mathbb{N}$. We denote $t_j = t_j(\omega_1)$ and suppress ω_1 in the following statements for ease of notation. Due to the lim-inf type of convergence just proved in (19), for each t_j, there exists a first successful iteration, $\ell_j := \ell(t_j) > t_j$, such that, for large enough k,

$$\|\nabla f(\boldsymbol{X}_k)\| > 2\epsilon_0, \quad t_j \leq k < \ell_j, \tag{32}$$

and

$$\|\nabla f(\boldsymbol{X}_{\ell_j})\| < \epsilon_0. \tag{33}$$

Define $\mathcal{A}_j := \left\{k \in \mathcal{S} : t_j \leq k < \ell_j\right\} \subset \mathcal{S}$. Let j be sufficiently large and let $k \in \mathcal{A}_j$. We then obtain from the fact that $\lim_{k \to \infty} \|\bar{\boldsymbol{G}}_k(N_k) - \nabla f(\boldsymbol{X}_k)\| = 0$ wp1,

$$\|\bar{\boldsymbol{G}}_k(N_k)\| > \epsilon_0. \tag{34}$$

Since k is a successful iteration, we know from Step 5 of Algorithm 1 that $\hat{\rho}_k \geq \eta$. Furthermore, Step 5 of Algorithm 1, Assumption 6 and (34) then imply that

$$\begin{aligned} f(\boldsymbol{X}_k) - f(\boldsymbol{X}_{k+1}) + \bar{E}_k^0 - \bar{E}_k^s &= \bar{F}_k^0(N_k) - \bar{F}_k^s(\widetilde{N}_{k+1}) \\ &\geq \eta[M_k(\boldsymbol{X}_k) - M_k(\boldsymbol{X}_k)] \\ &\geq \eta_1 \kappa_{fcd} \|\nabla f(\boldsymbol{X}_k, N_k)\| \min\left\{\frac{\|\nabla f(\boldsymbol{X}_k, N_k)\|}{1 + \|B_k\|}, \Delta_k\right\} \\ &\geq \frac{1}{2}\eta \kappa_{fcd} \epsilon_0 \min\left\{\frac{\epsilon_0}{1 + \kappa_{\mathsf{B}}}, \Delta_k\right\}. \end{aligned} \tag{35}$$

Notice that from Theorem 2 and the fact that f is bounded from below, left-hand side of (35) must tend to zero as k tends to infinity. This implies that

$$\lim_{j\to\infty} \Delta_k \mathbb{I}_{k\in\mathcal{A}_j} = 0. \tag{36}$$

As a consequence of (35) and (36), the minimum operator in (35) becomes binding at Δ_k, yielding that for sufficiently large j and $k \in \mathcal{A}_j$, $\Delta_k \leq \frac{2}{\eta\kappa_{fcd}\epsilon_0}(f(\boldsymbol{X}_k) - f(\boldsymbol{X}_{k+1}) + \bar{E}_k^0 - \bar{E}_k^s)$. From this bound, and using the fact that $\|\boldsymbol{X}_k - \boldsymbol{X}_{k+1}\| \leq \Delta_k$ for all k, we deduce that

$$\|\boldsymbol{X}_{t_j} - \boldsymbol{X}_{\ell_j}\| \leq \sum_{i\in\mathcal{A}_j} \|\boldsymbol{X}_i - \boldsymbol{X}_{i+1}\| \leq \sum_{i\in\mathcal{A}_j} \Delta_i \leq \frac{2(f(\boldsymbol{X}_{t_j}) - f(\boldsymbol{X}_{\ell_j}))}{\eta\kappa_{fcd}\epsilon_0} + \sum_{i\in\mathcal{A}_j} (\bar{E}_i^0 - \bar{E}_i^s). \tag{37}$$

Recall from Theorem 2 and boundedness of f from below that right-hand side of (37) converges to 0 as j goes to infinity. We can conclude from (37) that $\lim_{j\to\infty} \|\boldsymbol{X}_{t_j} - \boldsymbol{X}_{\ell_j}\| = 0$. Consequently, by continuity of the gradient we obtain that $\lim_{j\to\infty} \|\nabla f(\boldsymbol{X}_{t_j}) - \nabla f(\boldsymbol{X}_{\ell_j})\| = 0$. However, this contradicts $\|\nabla f(\boldsymbol{X}_{t_j}) - \nabla f(\boldsymbol{X}_{\ell_j})\| > \epsilon_0$, obtained from (32) and (33). Thus, (20) must hold. Similarly, under any one of the conditions $(b) - (d)$, we can obtain same result by applying the same steps, but this time we use $\widetilde{N}_{k+1} = N_k$ for all $k \in \mathbb{N}$. □

We now shift towards proving the wp1 convergence of ASTRO-DF. Our initial step involves demonstrating that the trust-region radius converges to 0 as k goes to infinity almost surely in Lemma 6. This result can be invoked to deduce that the model gradient error also converges to 0 as k approaches infinity almost surely.

Lemma 6. *Let Assumptions 1, 2, 4, 6, and 8 hold. Then, $\Delta_k \xrightarrow{wp1} 0$ as $k \to \infty$ under any one of the conditions $(a)^{df}$-$(d)^{df}$.*

Proof. Let $\mathcal{S} = \{k : \hat{\rho}_k \geq \eta\}$ and $\omega \in \Omega$. We suppress ω in the following statements for ease of notation. Following similar steps as the proof of Theorem 4.5 in [6], we can obtain $\theta \sum_{k\in\mathcal{S}} \Delta_k^2 \leq f(\boldsymbol{x}_0) - f^* + \sum_{k\in\mathcal{S}}(\bar{E}_k^0 - \bar{E}_k^s)$, for any $k \in \mathcal{S}$ where $\theta = (\eta\kappa_{fcd}(2\mu)^{-1}((\mu\kappa_{\mathsf{H}})^{-1} \wedge 1))$, from which we obtain

$$\sum_{k=0}^{\infty} \Delta_k^2 < \frac{\gamma_1^2}{1-\gamma_2^2}\left(\frac{\Delta_0^2}{\gamma_2^2} + \frac{f(\boldsymbol{x}_0) - f^* + \sum_{k=0}^{\infty}(\bar{E}_k^0 - \bar{E}_k^s)}{\theta}\right). \tag{38}$$

The details to obtain (38) can be found in [6]. Moreover, we know from Theorem 2 and 3 that $\mathbb{P}\{|\bar{E}_k - \bar{E}_{k+1}| \geq c_{fd}\Delta_k^2 \text{ i.o.}\} = 0$ for any $c_{fd} > 0$ under any one of the conditions $(a)^{df}$-$(d)^{df}$. It implies that there must exists a sufficiently large K_Δ such that $|\bar{E}_k - \bar{E}_{k+1}| < c_\Delta\Delta_k^2$ for any given $c_\Delta > 0$ and every $k \geq K_\Delta$. Following similar steps outlined in the proof of Theorem 4.5 in [6], we can derive

$$\sum_{k=K_\Delta}^{\infty} \Delta_k^2 < \frac{\gamma_1^2}{1-\gamma_2^2}\left(\frac{\Delta_0^2}{\gamma_2^2} + \frac{f(\boldsymbol{x}_0) - f^* + \widetilde{E}_{0,K_\Delta-1}}{\theta}\right)\left(1 - \frac{\gamma_1^2}{1-\gamma_2^2}\frac{c_\Delta}{\theta}\right)^{-1} < \infty.$$

where $\widetilde{E}_{0,K_\Delta-1} = \sum_{k=0}^{K_\Delta-1}(|\bar{E}_k - \bar{E}_{k+1}|)$. Therefore, $\Delta_k \xrightarrow{wp1} 0$ as $k \to \infty$. As a result, the theorem is satisfied under any one of the condition $(a)^{df}$-$(d)^{df}$. □

Lastly, we present the proof of the strong consistency for ASTRO-DF.

Proof. [of Theorem 4 for ASTRO-DF.] Let $\omega \in \Omega$. We suppress ω in the following statements for ease of notation. We know from Theorem 3 that given any $c_{fd} > 0$, there exists sufficiently large K such that $|\bar{E}_k^i - \bar{E}_k^0| < c_{fd}\Delta_k^2$ for any $k \geq K$ and any $i \in \{1, 2, \ldots, p\}$ under any one of the conditions $(a)^{df}$-$(d)^{df}$. We then obtain from Lemma 4 given small enough $c_{fd} > 0$,

$$\|\nabla f(\boldsymbol{X}_k) - \boldsymbol{G}_k\| < (\kappa_{eg1} + \sqrt{p}c_{fd}\kappa_{eg2})\Delta_k, \tag{39}$$

for any $k \geq K$. Now define the set

$$\mathcal{V} := \left\{\omega : \exists\ \{k_j\} \text{ s.t. } \left(\Delta_{k_j}(\omega) \leq \frac{\|\boldsymbol{G}_{k_j}\|\kappa_{fcd}(1-\eta)}{\kappa_{Lg} + \kappa_{\mathsf{H}} + \kappa_{eg1}}\right) \bigcap \left(\hat{\rho}_{k_j}(\omega) < \eta\right)\right\}.$$

Following the same steps in the proof of Lemma 5, we have that wp1 for sufficiently large k, any $c_{fd} > 0$, and any $c_f > 0$

$$\begin{aligned}
\left|\bar{F}_k^s(N_k) - M_k(\widetilde{\boldsymbol{X}}_{k+1})\right| &\leq \int_0^1 \left\|\nabla f(\boldsymbol{X}_k + \tau\boldsymbol{S}_k) - \nabla f(\boldsymbol{X}_k)\right\|\left\|\boldsymbol{S}_k\right\| d\tau + \frac{1}{2}\|\boldsymbol{S}_k\|^2\kappa_{\mathsf{H}} \\
&\qquad + |\bar{E}_k^s - \bar{E}_k^0| + \|\nabla f(\boldsymbol{X}_k) - \boldsymbol{G}_k\|\|\boldsymbol{S}_k\| \\
&\leq \int_0^1 \kappa_{Lg}\|\boldsymbol{S}_k\|^2\tau d\tau + \left(\frac{1}{2}\kappa_{\mathsf{H}} + c_{fd} + \kappa_{eg1} + \sqrt{p}c_{fd}\kappa_{eg2}\right)\Delta_k^2 \\
&\leq \left(\frac{1}{2}(\kappa_{\mathsf{H}} + \kappa_{Lg}) + c_{fd} + \kappa_{eg1} + \sqrt{p}c_{fd}\kappa_{eg2}\right)\Delta_k^2,
\end{aligned}$$

where the second inequality comes from (18). Without loss of generality, we can set $(1 + \sqrt{p}\kappa_{eg2})c_{fd} = 2^{-1}(\kappa_{Lg} + \kappa_{\mathsf{H}})$. We have $\mathbb{P}\{\mathcal{V}\} = 0$ by again following the rest steps in the proof of Lemma 5. We have from Lemma 6 and (39) that $\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| \xrightarrow{wp1} 0$ as $k \to \infty$. We then obtain $\liminf \|\nabla f(\boldsymbol{X}_k)\| \xrightarrow{wp1} 0$ as $k \to \infty$ with $\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| \xrightarrow{wp1} 0$ as $k \to \infty$, Lemma 6, and $\mathbb{P}\{\mathcal{V}\} = 0$. The proof follows from that of Theorem 4.6 in [6]. The wp1 convergence of Algorithm 2 holds by $\liminf_{k\to\infty} \|\nabla f(\boldsymbol{X}_k)\| = 0$ wp1 and Lemma 6. The proof is completed following steps in Theorem 5.5 in [5]. □

6. COMPLEXITY

In this section, we present the iteration and work complexity analysis for ASTRO and ASTRO-DF under different adaptive sampling rules that correspond to the use of CRN and existing properties of the sample paths. We remark that the iteration complexity results of $\mathcal{O}(\epsilon^{-2})$ proven in this paper are consistent with the literature, and in particular, those of STORM (the competing method) with an advantage. ASTRO and ASTRO-DF algorithms make explicit use of the sampling rules to ensure accurate estimates and high-quality models.

6.1 Iteration complexity

We present the wp1 iteration complexity for ASTRO and ASTRO-DF. We denote the iteration stopping at ϵ-optimality as T_ϵ.

Theorem 6 (Iteration Complexity)**.** *Given small enough $\epsilon > 0$, $T_\epsilon \leq c_T \epsilon^{-2}$ almost surely for some well-defined random variable $c_T > 0$ and*

- *for ASTRO: if Assumptions 2-6, and 7 hold and under conditions (a)-(d);*
- *for ASTRO-DF: if Assumptions 1, 2, 4, 6, and 8 hold and under conditions $(a)^{df}$-$(d)^{df}$.*

We make two observations before presenting the proof of Theorem 6.

(*ic*-a) Both algorithms achieve an iteration complexity rate of $\mathcal{O}(\epsilon^{-2})$ under corresponding criteria. This implies that, under certain conditions, the algorithms can achieve the same convergence rate in terms of iterations with a smaller order of sampling, represented by smaller values of β_g and β_f.

(*ic*-b) Theorem 6 shows the wp1 iteration complexity, which is stronger than the claim that the random variable $\epsilon^2 T_\epsilon$ is $\mathcal{O}_p(1)$. Moreover, if we impose specific regularity conditions on the random variable c_T, such as having finite first moments, we can achieve the L_1 results, which aligns with the similar finding presented in [4]. However, we achieve this canonical rate without relying on assumptions such as probabilistically fully linear models, or their independence.

6.2 Proof of Theorem 6

By directly employing Corollary 5, we can establish the wp1 convergence of ASTRO.

Proof. [of Theorem 6 for ASTRO.] Let us first focus on the scenario where any one of the condition (a)-(c) holds. Let $f^* := \min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) > -\infty$ be the optimal function value and $\omega \in \Omega$. We suppress ω in the following statements for ease of notation. We know from Theorem 2 and 3 and Corollary 5 that there exists sufficiently large $K < T_\epsilon$ such that given small enough ϵ and any $c_{fd} > 0$, for all $k \in [K, T_\epsilon)$, $\|\nabla f(\boldsymbol{X}_k)\| > \epsilon$, $\|\bar{\boldsymbol{G}}_k(N_k)\| > \epsilon/2$, $|\bar{E}_k^0 - \bar{E}_k^s| \leq c_{fd}\Delta_k^2$, and

$$\Delta_k \geq \frac{1}{2}\left(\frac{\gamma_2 \kappa_{fcd}(1-\eta)}{1+\kappa_{Lg}+\kappa_{\mathsf{B}}}\right)\epsilon. \tag{40}$$

Then Assumption 6 and the condition for a successful iteration yeild for any $k \in \mathcal{S}_{T_\epsilon}(\omega)$,

$$\bar{F}_k^0(N_k) - \bar{F}_k^s(\widetilde{N}_{k+1}) \geq \frac{\eta\kappa_{fcd}}{2\mu}\min\left\{\frac{1}{\mu\kappa_{\mathsf{B}}}, 1\right\}\Delta_k^2, \tag{41}$$

where $\mathcal{S}_{T_\epsilon} := \{T_\epsilon > k \geq K : \hat{\rho}_k \geq \eta\}$. Let us denote the RHS of (41) as $c_{fcd}\Delta_k^2$ and the RHS of (40) as $c_{lb}\epsilon$. We then obtain

$$|\mathcal{S}_{T_\epsilon}| c_{lb} c_{fcd} \epsilon^2 \leq \sum_{k\in\mathcal{S}_{T_\epsilon}} c_{fcd}\Delta_k^2 \leq \sum_{k\in\mathcal{S}_{T_\epsilon}} (\bar{F}_k(N_k) - \bar{F}_{k+1}(\widetilde{N}_k)) \leq f(\boldsymbol{x}_0) - f^* + \sum_{k\in\mathcal{S}_{T_\epsilon}} (|\bar{E}_k^0 - \bar{E}_k^s|).$$

Then we obtain from the definition of K that given any $c_{fd} > 0$, $|\bar{E}_k^0 - \bar{E}_k^s| < c_{fd}\Delta_k^2$ for any $k \in \mathcal{S}_{T_\epsilon}$. Hence, we obtain $|\mathcal{S}_{T_\epsilon}|(c_{lb}c_{fcd} - c_{fd})\epsilon^2 \leq f(\boldsymbol{x}_0) - f^*$. Due to (40), a success event occurs after a finite number of unsuccessful iterations, allowing us to define the maximum portion of the unsuccessful iterations between K and T_ϵ as $c_{usp} \in (0,1)$, i.e., $|\mathcal{S}_{T_\epsilon}| \geq (1-c_{usp})(T_\epsilon - K)$. Then we obtain $(1-c_{usp})(T_\epsilon - K)(c_{lb}c_{fcd} - c_{fd})\epsilon^2 \leq f(\boldsymbol{x}_0) - f^*$. As a result, the assertion of the theorem has

to hold under any one of (a)-(c). Now let us consider the scenario where the condition (d) holds. We know from Theorem 3 and 3 and Corollary 5 that there exists sufficiently large $K_{sf} < T_\epsilon$ such that given small enough ϵ and any $c_{gd} > 0$, for all $k \in [K_{sf}, T_\epsilon)$, $\|\nabla f(\boldsymbol{X}_k)\| > \epsilon$, $\|\bar{\boldsymbol{G}}_k(N_k)\| > \epsilon/2$, (40), and $\|\bar{\boldsymbol{G}}(\boldsymbol{x}, N_k) - \bar{\boldsymbol{G}}(\boldsymbol{x} + \boldsymbol{S}_k, N_k)\| \leq c_{gd}\|\boldsymbol{S}_k\|$. We note that

$$\bar{F}_k^s(N_k) - \bar{F}_k^0(N_k) = 2\left[\bar{\boldsymbol{G}}_k^0(N_k) - \frac{1}{2}\bar{\boldsymbol{G}}_k^s(N_k)\right]^\intercal \boldsymbol{S}_k + 2\int_0^1 \left[\bar{\boldsymbol{G}}(\boldsymbol{X}_k + \tau\boldsymbol{S}_k, N_k) - \bar{\boldsymbol{G}}_k^0(N_k)\right]^\intercal \boldsymbol{S}_k \, d\tau - \int_0^1 \left[\bar{\boldsymbol{G}}(\boldsymbol{X}_k + \tau\boldsymbol{S}_k, N_k) - \bar{\boldsymbol{G}}_k^s(N_k)\right]^\intercal \boldsymbol{S}_k \, d\tau.$$

Then, we obtain from the definition of K_{sf} that given any $c_{gd1} > 0$, $c_{gd2} > 0$, and c_{gd3},

$$2[\bar{\boldsymbol{G}}_k^0(N_k) - 2^{-1}\bar{\boldsymbol{G}}_k^s(N_k)]^\intercal \boldsymbol{S}_k < c_{gd1}\Delta_k^2, [\bar{\boldsymbol{G}}(\boldsymbol{X}_k + \tau\boldsymbol{S}_k, N_k) - \bar{\boldsymbol{G}}_k^0(N_k)]^\intercal \boldsymbol{S}_k < c_{gd2}\tau\Delta_k^2,$$

and

$$[\bar{\boldsymbol{G}}(\boldsymbol{X}_k + \tau\boldsymbol{S}_k, N_k) - \bar{\boldsymbol{G}}_k^s(N_k)]^\intercal \boldsymbol{S}_k < 2c_{gd3}\tau\Delta_k^2,$$

for sufficiently large k. Then following similar steps for conditions (a)-(c), and setting $\mathcal{S}_{T_\epsilon}^{sg} := \{T_\epsilon > k \geq K : \hat{\rho}_k \geq \eta\}$, we obtain

$$\sum_{k \in \mathcal{S}_{T_\epsilon}^{sg}} c_{fcd2}\Delta_k^2 \leq \sum_{k \in \mathcal{S}_{T_\epsilon}^{sg}} \bar{F}_k^s(N_k) - \bar{F}_k^0(N_k) \leq \sum_{k \in \mathcal{S}_{T_\epsilon}^{sg}} (c_{gd1} + c_{gd2} + c_{gd3})\Delta_k^2.$$

It implies that $(1 - c_{usp})(T_\epsilon - K_g)(c_{fcd} - (c_{gd1} + c_{gd2} + c_{gd3}))\epsilon^2 = 0$. As a result, the assertion of the theorem has to hold under condition (d). □

Now we will focus on proving the iteration complexity of ASTRO-DF. As a first step we need the derivative free counterpart of Corollary 5. It guarantees that as the gradient is constrained to a lower bound, the TR radius is also ensured to have a lower bound, thereby facilitating the occurrence of successful iterations over time.

Lemma 7. *Let Assumptions 1, 2, 4, 6, and 8 hold and $\epsilon > 0$ be given. Then there exists $c_{lb} > 0$ where $\mathbb{P}\{\Delta_k < c_{lb}\epsilon$ for large $k < T_\epsilon \Rightarrow k \in \mathcal{S}\} = 1$ under any one of $(a)^{df}$-$(d)^{df}$.*

Proof. Let ω be fixed and we suppress ω in the following statements for ease of notation. For the case under condition $(a)^{df}$, the theorem is satisfied trivially. The proof is completed by trivially following steps in the proof of Lemma 4.8 in [6]. Now let us focus on the case under any one of the conditions $(b)^{df}$-$(d)^{df}$. We first obtain from the proof of ASTRO-DF convergence that for sufficiently large k, any $c_{fd} > 0$, and any $c_f > 0$

$$\begin{aligned} \left|\bar{F}_k^s(N_k) - M_k(\widetilde{\boldsymbol{X}}_{k+1})\right| &\leq \int_0^1 \left\|\nabla f(\boldsymbol{X}_k + \tau\boldsymbol{S}_k) - \nabla f(\boldsymbol{X}_k)\right\|\left\|\boldsymbol{S}_k\right\| d\tau + \frac{1}{2}\|\boldsymbol{S}_k\|^2\kappa_{\mathsf{H}} \\ &\quad + |\bar{E}_k^s - \bar{E}_k^0| + \|\nabla f(\boldsymbol{X}_k) - \boldsymbol{G}_k\|\|\boldsymbol{S}_k\| \\ &\leq \int_0^1 \kappa_{Lg}\|\boldsymbol{S}_k\|^2\tau d\tau + \left(\frac{1}{2}\kappa_{\mathsf{H}} + c_f + \kappa_{eg1} + \sqrt{p}c_{fd}\kappa_{eg2}\right)\Delta_k^2 \\ &\leq \left(\frac{1}{2}(\kappa_{\mathsf{H}} + \kappa_{Lg}) + c_f + \kappa_{eg1} + \sqrt{p}c_{fd}\kappa_{eg2}\right)\Delta_k^2 \;\; wp1, \end{aligned} \tag{42}$$

and $\|\nabla f(\boldsymbol{X}_k) - \boldsymbol{G}_k\| < (\kappa_{eg1} + \sqrt{p}c_{fd}\kappa_{eg2})\Delta_k$ wp1 Then, if $\Delta_k(\omega) < c_{lb}\epsilon$ for large $k < T_\epsilon(\omega)$ and some $c_{lb} > 0$, we get

$$\|\boldsymbol{G}_k\| \geq \|\nabla f(\boldsymbol{X}_k)\| - \|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\| > \left(\frac{1}{c_{lb}} - \kappa_{eg1} - \sqrt{p}c_{fd}\kappa_{eg2}\right)\Delta_k, \tag{43}$$

where the last inequality comes from $\|\nabla f(\boldsymbol{X}_k(\omega))\| > \epsilon$ since $k < T_\epsilon$. To complete the proof we need to show the model will lead to success.

$$\begin{aligned}
|1 - \hat{\rho}_k| &= \left|\frac{\bar{F}_k^s(N_k) - M_k(\widetilde{\boldsymbol{X}}_{k+1})}{M_k(\boldsymbol{X}_k) - M_k(\widetilde{\boldsymbol{X}}_{k+1})}\right| \\
&\leq \frac{\left(\frac{1}{2}(\kappa_{\mathsf{H}} + \kappa_{Lg}) + c_f + \sqrt{p}\kappa_{eg2}c_{fd} + \kappa_{eg1}\right)\Delta_k^2}{\frac{\kappa_{fcd}}{2}\|\boldsymbol{G}_k\|\left(|\boldsymbol{G}_k\|\kappa_{\mathsf{H}}^{-1} \wedge \Delta_k\right)} \\
&< \frac{\left(\frac{1}{2}(\kappa_{\mathsf{H}} + \kappa_{Lg}) + c_f + \sqrt{p}\kappa_{eg2}c_{fd} + \kappa_{eg1}\right)\Delta_k^2}{\frac{\kappa_{fcd}}{2}\left(c_{lb}^{-1} - \kappa_{eg1} - \sqrt{p}c_{fd}\kappa_{eg2}\right)\left(\left(c_{lb}^{-1} - \kappa_{eg1} - \sqrt{p}c_{fd}\kappa_{eg2}\right)\kappa_{\mathsf{H}}^{-1} \wedge 1\right)\Delta_k^2},
\end{aligned}$$

where the first inequality comes from (42) and the second inequality comes from (43). Then there must exists sufficiently small $c_{lb} > 0$ such that $|1 - \hat{\rho}_k| < 1 - \eta$ for large $k < T_\epsilon$. □

Proof. [of Theorem 6 for ASTRO-DF.] For a fixed ω and WLOG, we obtain $\Delta_k(\omega) \geq c_{lb}(\omega)\epsilon$ for all $k < T_\epsilon(\omega)$. This is true by Lemma 7 with the choice of $c_{lb}(\omega) = \gamma_2 c_{lb}$ for large $k < T_\epsilon(\omega)$. For small k, we can make $c_{lb}(\omega) > 0$ small enough to be notwithstanding this lower bound. Then, we can write $\sum_{k=0}^{\infty}\Delta_k^2(\omega) > \sum_{k=0}^{T_\epsilon(\omega)-1}\Delta_k^2(\omega) > c_{lb}^2(\omega)\epsilon^2 T_\epsilon(\omega)$. We know from Lemma 6 that $\sum_{k=0}^{T_\epsilon(\omega)-1}\Delta_k^2(\omega)$ is finite wp1, which then implies that $\epsilon^2 T_\epsilon(\omega) < c_T^{df}(\omega)$ for all $\epsilon \leq \epsilon_0(\omega)$ and some finite $c_T^{df}(\omega) > 0$ under any one of the conditions $(a)^{df}$-$(d)^{df}$. □

6.3 Work Complexity

We show the wp1 work complexity for Algorithm 1 and 2. We denote the work complexity, i.e., total number of oracle calls until T_ϵ, for Algorithm 1 as $W_\epsilon := \sum_{k=0}^{T_\epsilon}(N_k + \widetilde{N}_k)$, and for Algorithm 2 as $W_\epsilon := \sum_{k=0}^{T_\epsilon}(\sum_{i=0}^{p} N_k^i + \widetilde{N}_k)$.

Theorem 7 (Work Complexity)**.** *Given small enough $\epsilon > 0$,*

$$W_\epsilon \leq c_W \epsilon^{(-2-2\max\{\beta_f,\beta_g\})}(\log 1/\epsilon)^{-1}, \tag{44}$$

wp1 for some well-defined random variable $c_W > 0$ and

- *for ASTRO: if Assumptions 2-6, and 7 hold and under conditions (a)-(d);*
- *for ASTRO-DF: Assumptions 1, 2, 4, 6, and 8 hold and under conditions $(a)^{df}$-$(d)^{df}$.*

We make some observations before providing the proof of Theorem 7.

(*wc*-a) ASTRO's work complexity (Table 1) depends on β_f, not β_g. Precise gradients from the first-order oracle (smaller β_g) differ from interpolation models, but Step 5 needs accurate function estimates for strong consistency.

(*wc*-b) Recently, the first-order STORM has revealed to have a work complexity as $\mathcal{O}(\epsilon^{-6})$ [16]. In contrast, ASTRO under condition (*a*) achieves $\widetilde{\mathcal{O}}(\epsilon^{-6})$. The reason is that the increasing sequence $\{\lambda_k\}$ in (11) and (12) gives rise to $(\log 1/\epsilon)^{-1}$ in the RHS of (44). However, the sequence $\{\lambda_k\}$ enables us to achieve strong consistency and the wp1 complexities without requiring any additional assumptions as stated in (*ic*-b).

(*wc*-c) In situations where the sample path exhibits smoothness, the work complexity for ASTRO is $\widetilde{\mathcal{O}}(\epsilon^{-2})$. However, when considering ASTRO-DF, the complexity scales as $\widetilde{\mathcal{O}}(\epsilon^{-4})$. This implies that when the sample-path gradient remains continuous and CRN is implemented, the first-order oracle gains an advantageous position. The crux of this advantage lies in the fact that (26) can replace (22) so that (16) holds with $\beta_g = 0$.

6.4 Proof of Theorem 7

Proof. [of Theorem 7 for ASTRO.] Let ω be any sample path in Ω. We suppress ω in the following statements for ease of notation. We first know from Theorem 2.8 in [5] that, $\hat{\sigma}_F(\boldsymbol{x}, N) \to \sigma_F(\boldsymbol{x})$ almost surely as N goes to ∞. Since λ_k is lower bound for N and converges to ∞ as k goes to ∞, $\hat{\sigma}_F(\boldsymbol{x}, \lambda_k) \to \sigma_F(\boldsymbol{x})$ almost surely as k goes to ∞. As a result, there exists sufficiently large K_f such that $\hat{\sigma}_F(\boldsymbol{x}, \lambda_k) \leq 2\sigma_f$ for any $k \geq K_f$ and $\boldsymbol{x} \in \mathbb{R}^d$. Moreover, we know from Theorem 2.8 in [5] and Assumption 3 that $\mathrm{Tr}(\hat{\sigma}^2_{\boldsymbol{G}}(\boldsymbol{x}, n)) \to \mathrm{Tr}(\sigma^2_{\boldsymbol{G}}(\boldsymbol{x}))$ almost surely as n goes to ∞. Then there exists sufficiently large K_g such that $\mathrm{Tr}(\hat{\sigma}^2_{\boldsymbol{G}}(\boldsymbol{x}, \lambda_k)) \leq 2d\sigma_g^2$ for any $k \geq K_g$ and $\boldsymbol{x} \in \mathbb{R}^d$. Lastly, let K and c_{lb} be the ones defined in the proof of 6. Without loss of generality, we now assume that ϵ is small enough such that $K_\sigma < T_\epsilon$, where $K_\sigma := \max\{K_g, K_f, K\}$. Without loss of generality, we now assume that ϵ is small enough such that $K_\sigma < T_\epsilon$. Then, we obtain from (11) for all $k \in \{K_\sigma, K_\sigma + 1, \ldots, T_\epsilon\}$

$$\max\{N_k, \widetilde{N}_{k+1}\} \leq \max\left\{\frac{2\mathrm{Tr}(\sigma^2_{\boldsymbol{G}})}{\kappa^2(c^{db}_{lb}(\omega))^{2\beta_g}}, \frac{2\sigma_f^2}{\kappa^2(c^{db}_{lb}(\omega))^{2\beta_f}}\right\}\epsilon^{2\max\{\beta_f,\beta_g\}}\lambda_k, \tag{45}$$

Let us denote the RHS of (45) as $c_{ub}\epsilon^{-2\max\{\beta_f,\beta_g\}}\lambda_k$. Then we obtain

$$\begin{aligned}\sum_{k=0}^{T_\epsilon}(N_k + \widetilde{N}_{k+1}) &\leq \sum_{k=0}^{K_\sigma}(N_k + \widetilde{N}_{k+1}) + \sum_{k=K_\sigma+1}^{T_\epsilon} 2c_{ub}\epsilon^{-2\max\{\beta_f,\beta_g\}}\lambda_k \\ &\leq \sum_{k=0}^{K_\sigma}(N_k + \widetilde{N}_{k+1}) + 2c_{ub}\epsilon^{-2\max\{\beta_f,\beta_g\}}T_\epsilon\lambda_{T_\epsilon},\end{aligned} \tag{46}$$

where the first inequality comes from (45). As a result, we obtain from Theorem 6 that the RHS of (46) can be upper bounded by $c_W\epsilon^{(-2-2\max\{\beta_f,\beta_g\})}(\log 1/\epsilon)^{-1}$ almost surely, where c_W is a finite positive random variable. □

Proof. [of Theorem 7 for ASTRO-DF.] Let ω be any sample path in Ω. Let $K_f(\omega)$ be the same as before. Without loss of generality, we now assume that ϵ is small enough such that $K_f(\omega) < T_\epsilon(\omega)$. We also know from the proof of Theorem 6 that there exists small enough $c^{df}_{lb}(\omega) > 0$ such that $\Delta_k(\omega) \geq c^{df}_{lb}(\omega)\epsilon$ for all $k < T_\epsilon(\omega)$. Then, we obtain for all $k \in \{K_f(\omega), K_f(\omega) + 1, \ldots, T_\epsilon(\omega)\}$,

$$\max\left\{\max_{i\in\{0,1,\ldots,p\}} N^i_k, \widetilde{N}_{k+1}\right\} \leq \max\left\{1, \frac{2\sigma_f}{\kappa^2_{af}(c^{df}_{lb}(\omega))^{2\beta_f}}\right\}\epsilon^{-2\beta_f}\lambda_k. \tag{47}$$

Let us denote the RHS of (47) as $c_{ub}^{df}(\omega)\epsilon^{-2\beta_f}\lambda_k$. Then we obtain

$$\begin{aligned}\sum_{k=0}^{T_\epsilon}\left(\sum_{i=0}^{p}N_k^i+\widetilde{N}_k\right) &\leq \sum_{k=0}^{K_f}\left(\sum_{i=0}^{p}N_k^i+\widetilde{N}_k\right)+\sum_{k=K_f+1}^{T_\epsilon}(p+2)c_{ub}^{df}(\omega)\epsilon^{-2\beta_f}\lambda_k \\ &\leq \sum_{k=0}^{K_f}\left(\sum_{i=0}^{p}N_k^i+\widetilde{N}_k\right)+(p+2)c_{ub}^{df}(\omega)\epsilon^{-2\beta_f}T_\epsilon\lambda_{T_\epsilon},\end{aligned} \tag{48}$$

where the first inequality comes from (47). As a result, we obtain from Theorem 6 that the RHS of (46) can be upper bounded by $c_W^{df}\epsilon^{(-2-2\beta_f)}(\log 1/\epsilon)^{-1}$ almost surely, where c_W^{df} is a finite positive random variable. □

7. CONCLUDING REMARKS

We make three remarks in closing.

1. In simulation folklore, CRN is crucial for the implementation efficiency of any SO algorithm. The complexity results in this paper seem to corroborate such folklore, suggesting that CRN may be remarkably important especially in adaptive-sampling TR algorithms operating in contexts with smooth sample-paths.

2. The heavy discrepancy between the complexity of ASTRO/ASTRO-DF and SGD in the non-CRN context suggests simple alterations in the sufficient reduction test used within standard TR algorithms.

3. We anticipate that the insights obtained from the complexity analysis of ASTRO/ASTRO-DF will transfer to other adaptive sampling TR algorithms because the bulk of our complexity calculations arise out of a generic step within TR rather than algorithmic mechanics specific to ASTRO/ASTRO-DF.

References

[1] K. Chang, L. Hong, and H. Wan, "Stochastic trust-region response-surface method STRONG—a new response-surface framework for simulation optimization," *INFORMS Journal on Computing*, vol. 25, no. 2, pp. 230–243, 2013.

[2] K. Chang and H. Wan, "Stochastic trust region response surface convergent method for generally-distributed response surface," in *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pp. 563–573, 2009.

[3] R. Chen, M. Menickelly, and K. Scheinberg, "Stochastic optimization using a trust-region method and random models," *Mathematical Programming*, vol. 169, no. 2, pp. 447–487, 2018.

[4] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg, "Convergence rate analysis of a stochastic trust-region method via supermartingales," *INFORMS Journal on Optimization*, vol. 1, no. 2, pp. 92–119, 2019.

[5] S. Shashaani, F. S Hashemi, and R. Pasupathy, "ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3145–3176, 2018.

[6] Y. Ha and S. Shashaani, "Iteration complexity and finite-time efficiency of adaptive sampling trust-region methods for stochastic derivative-free optimization," *ArXiv 2305.10650*, 2023.

[7] F. E. Curtis, K. Scheinberg, and R. Shi, "A stochastic trust region algorithm based on careful step normalization," *INFORMS Journal on Optimization*, vol. 1, no. 3, pp. 200–220, 2019.

[8] P. K. Ragavan, S. R. Hunter, R. Pasupathy, and M. R. Taaffe, "Adaptive sampling line search for local stochastic optimization with integer variables," *Mathematical Programming*, vol. 196, no. 1-2, pp. 775–804, 2022.

[9] J. Nocedal, "Updating quasi-newton matrices with limited storage," *Mathematics of Computation*, vol. 35, no. 151, pp. 773–782, 1980.

[10] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[11] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to derivative-free optimization*. Society for Industrial and Applied Mathematics, 1st ed., 2009.

[12] A. R. Conn, K. Scheinberg, and L. N. Vicente, "Global convergence of general derivative-free trust-region algorithms to first-and second-order critical points," *SIAM Journal on Optimization*, vol. 20, no. 1, pp. 387–415, 2009.

[13] B. Ko and Q. Tang, "Sums of dependent nonnegative random variables with subexponential tails," *Journal of Applied Probability*, vol. 45, no. 1, pp. 85–94, 2008.

[14] Y. Yang, K. Wang, R. Leipus, and J. Šiaulys, "Tail behavior of sums and maxima of sums of dependent subexponential random variables," *Acta Applicandae Mathematicae*, vol. 114, no. 3, pp. 219–231, 2011.

[15] P. Glasserman, *Monte Carlo methods in financial engineering*, vol. 53. Springer, 2004.

[16] B. Jin, K. Scheinberg, and M. Xie, "Sample complexity analysis for adaptive optimization algorithms with stochastic oracles," *arXiv preprint arXiv:2303.06838*, 2023.