Stochastic trust-region algorithm in random subspaces with convergence and expected complexity analyses*

Kwassi Joseph Dzahini† Stefan M. Wild†

July 15, 2022

Abstract

This work proposes a framework for large-scale stochastic derivative-free optimization (DFO) by introducing STARS, a trust-region method based on iterative minimization in random subspaces. This framework is both an algorithmic and theoretical extension of an algorithm for stochastic optimization with random models (STORM). Moreover, STARS achieves scalability by minimizing interpolation models that approximate the objective in low-dimensional affine subspaces, thus significantly reducing per-iteration costs in terms of function evaluations and yielding strong performance on large-scale stochastic DFO problems. The user-determined dimension of these subspaces, when the latter are defined, for example, by the columns of so-called Johnson–Lindenstrauss transforms, turns out to be independent of the dimension of the problem. For convergence purposes, both a particular quality of the subspace and the accuracies of random function estimates and models are required to hold with sufficiently high, but fixed, probabilities. Using martingale theory under the latter assumptions, an almost sure global convergence of STARS to a first-order stationary point is shown, and the expected number of iterations required to reach a desired first-order accuracy is proved to be similar to that of STORM and other stochastic DFO algorithms, up to constants.

1 Introduction.

Outstanding growth in the use of computers and sensors has attracted interest in myriad scientific and engineering fields for solving difficult optimization problems involving functions available only through a zeroth-order oracle (i.e., the functions are *black boxes* [3]). Derivative-free optimization (DFO [3, 14, 23]) addresses such situations where closed-form expressions and derivatives are not available. This paper focuses on unconstrained stochastic DFO, wherein the objective function values are accessible only through a blackbox oracle corrupted by stochastic noise. We consider the problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}), \qquad \text{with} \quad f(\boldsymbol{x}) = \mathbb{E}_{\underline{\theta}}\left[f_{\underline{\theta}}(\boldsymbol{x})\right], \tag{1}$$

where the values of the continuously differentiable function $f:\mathbb{R}^n\to\mathbb{R}$ are available only via $f_{\underline{\theta}}$, a stochastically noisy version of f, and where $\underline{\theta}$ is a random variable whose distribution is possibly unknown. Stochastic gradient descent (SGD [30]) is arguably the most used algorithm to solve (1). However, SGD is very sensitive to the choice of its learning rate; and not only does the method use step directions that are not necessarily descent ones [28], but also it exhibits slow convergence [13]. Similar remarks have motivated recent work in stochastic DFO, leading to many algorithms such as model-based methods [9, 13, 33], those using a direct-search approach [1, 2, 12, 16, 17, 29], and others [5, 10, 19, 28] that employ stochastic estimation of the gradient $\nabla f(\boldsymbol{x})$. Because of the unavailability of true function values and the resulting need to use stochastic estimates and/or models, the performance of these methods is highly dependent on how accurate those stochastic quantities are. For example, STORM [9, 13] is a stochastic trust-region algorithm using so-called

*This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research Applied Mathematics Program under Contract No. DE-AC02-06CH11357.

†Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, IL 60439, USA (www.anl.gov/profile/kwassi-joseph-dzahini, www.mcs.anl.gov/~wild/).

β-probabilistically ε_f-accurate estimates of unknown function values as well as α-probabilistically κ-fully linear models of the objective function. For convergence purposes, such estimates and models whose accuracies are dynamically controlled by means of a trust-region radius δ_k need to be sufficiently accurate probabilistically. More precisely, at a given iteration k, the aforementioned estimates and models require, respectively, at least $\mathcal{O}\left(\delta_k^{-4}/(1-\sqrt{\beta})\right)$ and $\mathcal{O}\left((n+1)\max\left\{\delta_k^{-2},\delta_k^{-4}\right\}/(1-\alpha^{1/(n+1)})\right)$ function evaluations in a simple stochastic noise framework where no noisy gradient values are available; these quantities grow rapidly as $\alpha,\beta\in(0,1)$ or the dimension n gets larger or when δ_k tends to zero. On the other hand, ASTRO-DF [33] is a class of stochastic DFO trust-region algorithms where random estimates and polynomial interpolation models are constructed adaptively by using a Monte Carlo sampling whose extent is determined by continuously balancing and monitoring measures of sampling error and model bias. Recalling that the construction of *full space* quadratic interpolation models requires $q(n) := (n+1)(n+2)/2$ ($\ell(n) := n+1$ in the linear case) points [14] and that the accuracy of Monte Carlo-based estimates crucially depends on the sampling size, as suggested by the strong law of large numbers [34, Theorem 2.1.8], ASTRO-DF also does not have low per-iteration costs in terms of function evaluations.

In light of these observations, a key question naturally arises regarding model-based stochastic DFO methods: is there a way to improve the ability of these methods to handle large-scale problems. To answer a question similar to the one above in a context where the objective function is deterministic, Cartis and Roberts [11] recently introduced RSDFO, a general framework of scalable subspace methods for model-based DFO. To achieve scalability, RSDFO approximates the deterministic objective only in subspaces using so-called $\boldsymbol{Q_k}$-fully linear models, with $\boldsymbol{Q_k}\in\mathbb{R}^{n\times p}$ denoting a matrix whose entries are randomly selected and where $p\leq n$ (ideally, $p\ll n$) is a user-determined parameter. The RSDFO framework was then specialized to deterministic nonlinear least-squares problems, and high-probability worst-case complexity bounds of the methods were derived under mild assumptions. As highlighted in [11], another model-based subspace DFO method with similarities to RSDFO but admitting no convergence analysis is the *moving ridge function* approach [18], where an interpolation model is built in an *active space* that is determined by using existing objective function evaluations. Neumaier et al. [27] also proposed the VXQR method with no convergence analysis, where line searches are performed along directions selected in a subspace determined by previous iterates. More recently, a direct-search method based on probabilistic descent in *reduced spaces* was introduced in [31] for the optimization of deterministic objective functions, where the polling directions are constructed in random subspaces; complexity bounds were also derived, making use of probabilistic properties related to both the latter directions and subspaces.

Of the cited methods, the only ones that achieve scalability using random subspace strategies are developed for deterministic objective functions; to our knowledge, no such method exists for stochastic DFO. We address this gap by introducing a stochastic trust-region algorithm in random subspaces (STARS), in which scalability arises from constructing and then minimizing stochastic models that approximate the objective function in low-dimensional random subspaces. By defining these random subspaces through user-provided random matrices $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}\in\mathbb{R}^{n\times p}$ with $p\ll n$, only $q(p)$ and $\ell(p)$ points are required for the construction of quadratic and linear models, respectively, which are cheap models in terms of function evaluations since p can always be chosen independently of n in STARS. While both the convergence and expected complexity analyses of STARS are inspired by those of STORM [9, 13], dealing with the additional difficulties introduced by the randomness stemming from the entries of $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$ is not a trivial task. As an example, no prior work exists showing how so-called β_m-*probabilistically* $(\kappa_{ef},\kappa_{eg};\boldsymbol{Q_k})$-*fully linear models* in random subspaces, crucial for the present analysis, can be made available. Moreover, the theoretical analyses of STORM utilizing random steps $\boldsymbol{s_k}\in\mathbb{R}^n$ do not straightforwardly hold when the latter are replaced with $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}\underline{\boldsymbol{s}}_{\boldsymbol{k}}\in\mathbb{R}^n$ used by STARS. Our main results trivially imply those related to STORM in the full-space case where $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}=\boldsymbol{I_n}\in\mathbb{R}^{n\times n}$ for all $k\in\mathbb{N}$, with $\boldsymbol{I_n}$ denoting the identity matrix.

One of the key contributions of the present work is that it extends the analysis of a STORM-like framework [2, 6, 9, 13, 15–17, 28] to settings where additional randomness stems from internal mechanics (here, the entries of the random matrices $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$) of an algorithm for stochastic DFO. To our knowledge, STARS is the first DFO algorithm for stochastic objectives that achieves scalability using a random subspace strategy; STARS both extends STORM techniques to settings of scalable subspace methods and extends RSDFO techniques to stochastic objective functions. Another key contribution of this work is the results of Theorem 3.3 together with Corollary 3.3, which are, to the best of our knowledge, the first to rigorously provide a detailed strategy for the construction of β_m-probabilistically $(\kappa_{ef},\kappa_{eg};\boldsymbol{Q_k})$-fully linear models for stochastic

objective functions in random subspaces. Moreover, relying on a framework introduced in [9] using results derived from martingale theory, the analysis of STARS demonstrates that while using random subspace models, the expected complexity of the algorithm is surprisingly similar to that of STORM [9], stochastic direct-search [16], and line-search-based [28] methods up to constants.

The manuscript is organized as follows. Section 2 introduces the general framework of STARS and explains how it results in a stochastic process. Section 3 discusses various strategies for the selection of random subspaces, demonstrates how probabilistic estimates and models can be constructed in these subspaces, and provides conditions related to these random quantities that are necessary for the convergence of STARS. Sections 4 and 5, respectively, present the convergence and expected complexity analyses of STARS. Section 6 presents numerical results, followed by a discussion and suggestions for future work.

2 Random subspace trust-region and resulting stochastic process.

This section presents the general framework of STARS and explains how the proposed method results in a stochastic process.

2.1 The random subspace stochastic trust-region method.

Unlike the stochastic trust-region framework STORM [9, 13], which builds full space random models $\underline{m}_k(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, to approximate $f(\boldsymbol{x})$, STARS operates as follows. On iteration k, given a current iterate $\boldsymbol{x_k}$, let $\mathcal{Y}_k \subseteq \mathbb{R}^n$ be the affine space randomly chosen through the range of a matrix $\boldsymbol{Q_k} \in \mathbb{R}^{n\times p}$ whose entries are randomly selected; that is,

$$\mathcal{Y}_k = \{\boldsymbol{x_k} + \boldsymbol{Q_k s} : \boldsymbol{s} \in \mathbb{R}^p\}.$$

Then, given a trust-region radius $\delta_k > 0$, a subspace model $\hat{m}_k(\boldsymbol{s}), \boldsymbol{s} \in \mathbb{R}^p$ is built only on $\mathcal{Y}_k$ using realizations of the stochastically noisy function $f_{\underline{\theta}}$; the model serves as an approximation of $f(\boldsymbol{x})$ in $\mathcal{B}(\boldsymbol{x_k}, \delta_k; \boldsymbol{Q_k}) := \{\boldsymbol{x_k} + \boldsymbol{Q_k s} \in \mathcal{Y}_k : \|\boldsymbol{s}\|_2 \leq \delta_k\}$. For concreteness, here we employ a quadratic subspace model given by

$$f(\boldsymbol{x_k} + \boldsymbol{Q_k s}) \approx \hat{m}_k(\boldsymbol{s}) := f_k + \hat{\boldsymbol{g}}_{\boldsymbol{k}}^\top \boldsymbol{s} + \boldsymbol{s}^\top \hat{\boldsymbol{H}}_k \boldsymbol{s},$$

where $\hat{\boldsymbol{g}}_{\boldsymbol{k}} \in \mathbb{R}^p$ and $\hat{\boldsymbol{H}}_k \in \mathbb{S}^p$ are the low-dimensional model gradient and Hessian, respectively. A *tentative step* $\boldsymbol{Q_k s_k} \in \mathbb{R}^n$ is produced by using the solution $\boldsymbol{s_k} \in \mathbb{R}^p$ obtained by approximately minimizing $\hat{m}_k$ inside the trust region; that is, $\boldsymbol{s_k} \approx \arg\min\{\hat{m}_k(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^p,\ \|\boldsymbol{s}\|_2 \leq \delta_k\}$. Inspired by [11, 13], the *trial step* $\boldsymbol{s_k}$ has to provide a sufficient decrease in $\hat{m}_k$ by satisfying the following standard fraction of the Cauchy decrease condition.

Assumption 1. *For every iteration k, a trial step $\boldsymbol{s_k} \in \mathbb{R}^p$ is computed so that*[1]

$$\hat{m}_k(\boldsymbol{0}) - \hat{m}_k(\boldsymbol{s_k}) \geq \frac{\kappa_{fcd}}{2} \|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \min\left\{\delta_k, \frac{\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2}{\max\left(\left\|\hat{\boldsymbol{H}}_k\right\|, 1\right)}\right\}, \tag{2}$$

for some constant $\kappa_{fcd} \in (0, 1]$.

Estimates f_k^0 and f_k^s of $f(\boldsymbol{x_k})$ and $f(\boldsymbol{x_k} + \boldsymbol{Q_k s_k})$, respectively, are constructed by using evaluations of the noisy function $f_{\underline{\theta}}$. The possible change in f given by $\boldsymbol{s_k}$ is measured by comparing the estimates f_k^0 and f_k^s through the value of a ratio ρ_k. An iteration is called successful if a decrease in the estimates is deemed sufficient, in which case the current solution is updated by $\boldsymbol{x_k} + \boldsymbol{Q_k s_k}$ and the trust-region radius is not decreased. Otherwise, the latter is decreased while the former is not updated. Because of inaccurate estimates, successful iterations qualified as *false* could lead to an increase in f. While the present algorithmic framework is not able to identify such iterations, as was the case in [2, 13, 16, 17, 28], it will be possible to prove later in the proof of Theorem 4.2 combined with (8) that *true* iterations occur sufficiently often for convergence of the proposed method to hold.

[1] Throughout the manuscript, the matrix norm $\|\cdot\|$ is supposed to be *consistent* with the Euclidean norm; that is, $\|\boldsymbol{Ax}\|_2 \leq \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|_2$, which holds for both Frobenius and spectral matrix norms.

2.2 Stochastic process generated by the algorithm.

The random variables considered here are all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω, referred to as the *sample space*, is a nonempty set whose elements ω and subsets are called *sample points* and *events*, respectively. $\mathcal{F}$ is a collection of such events, which is called a *σ-algebra*. $\mathbb{P}$ is a finite measure defined on the measurable space $(\Omega, \mathcal{F})$, satisfying $\mathbb{P}(\Omega) = 1$ and referred to as the *probability measure*. An increasing subsequence $\{\mathcal{S}_k\}_k$ of σ-algebras of $\mathcal{F}$ will be called a *filtration*. Let $\mathscr{B}(\mathbb{R}^n)$ be the σ-algebra generated by the open sets of $\mathbb{R}^n$, also known as the Borel σ-algebra of $\mathbb{R}^n$. A random variable $\underline{z}$ is a measurable map defined on $(\Omega, \mathcal{F}, \mathbb{P})$ into the measurable space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, where measurability means that $\{\underline{z} \in A\} := \{\omega \in \Omega : \underline{z}(\omega) \in A\} =: \underline{z}^{-1}(A) \in \mathcal{F}\ \forall A \in \mathscr{B}(\mathbb{R}^n)$ [7]. For the remainder of the manuscript, vectors will be written in lowercase boldface (e.g., $\boldsymbol{x} \in \mathbb{R}^n, n \geq 2$) while matrices will be written in uppercase boldface (e.g., $\boldsymbol{Q_k} \in \mathbb{R}^{n\times p}$), and underlined letters (e.g., $z, \underline{\boldsymbol{x}}, \underline{\boldsymbol{Q}}_{\boldsymbol{k}}$) will be used to denote random quantities; $z = \underline{z}(\omega)$ will denote a realization of $\underline{z}$.

[0] Initialization
Choose constants $\gamma > 1$, $\eta_1 \in (0,1)$, $\eta_2 > 0$, $j_{\max} \in \mathbb{N}$, $\varepsilon \in (0,1)$, $\beta \in (0,1)$, $c_1 \geq 1$, initial trust-region radius $\delta_0 > 0$, and maximum trust-region radius $\delta_{\max} = \gamma^{j_{\max}}\delta_0$, starting point $\boldsymbol{x_0} \in \mathbb{R}^n$, and dimension $p \in [\![1, n]\!]$.
Set the iteration counter $k \leftarrow 0$.
[1] Construction of subspace model
Generate $\boldsymbol{Q_k}$: a realization of a[2] $WAM(1-\varepsilon, \beta)$, using a distribution $\mathbb{Q}_k$.
Build model $\hat{m}_k : \mathbb{R}^p \to \mathbb{R}$ that is $(\kappa_{ef}, \kappa_{eg}; \boldsymbol{Q_k})$-fully linear in $\mathcal{B}(\boldsymbol{x_k}, c_1\delta_k; \boldsymbol{Q_k})$.
[2] Step calculation
Compute $\boldsymbol{s_k} \approx \arg\min\{\hat{m}_k(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^p : \|\boldsymbol{s}\|_2 \leq \delta_k\}$ satisfying (2).
[3] Estimate computation
Obtain estimates $f_k^0 \approx f(\boldsymbol{x_k})$ and $f_k^s \approx f(\boldsymbol{x_k} + \boldsymbol{Q_k s_k})$ satisfying (6).
[4] Updates
Compute $\rho_k = \frac{f_k^0 - f_k^s}{\hat{m}_k(\boldsymbol{0}) - \hat{m}_k(\boldsymbol{s_k})}$.
If $\rho_k \geq \eta_1$ and $\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \geq \eta_2\delta_k$ (**success**):
set $\boldsymbol{x_{k+1}} = \boldsymbol{x_k} + \boldsymbol{Q_k s_k}$ and $\delta_{k+1} = \min\{\gamma\delta_k, \delta_{\max}\}$.
Otherwise (**failure**): set $\boldsymbol{x_{k+1}} = \boldsymbol{x_k}$ and $\delta_{k+1} = \gamma^{-1}\delta_k$.
Update the iteration counter $k \leftarrow k+1$, and go to **[1]**.

Algorithm 1: STARS

The deterministic models $\hat{m}_k$ and the function estimates f_k^0 and f_k^s are constructed by using evaluations of the noisy function $f_{\underline{\theta}}$ and the randomly selected entries of the deterministic matrix $\boldsymbol{Q_k} = \underline{\boldsymbol{Q}}_{\boldsymbol{k}}(\omega)$. Hence, $\hat{m}_k$, f_k^0, and f_k^s can be considered as realizations of random models and estimates $\hat{\underline{m}}_k$, $\underline{f}_k^0$, and $\underline{f}_k^s$, respectively. Each iteration of Algorithm 1 is therefore influenced by the behavior of these random quantities; and consequently, the algorithm results in a stochastic process. The present work shows that under certain conditions on the sequences $\{\underline{\boldsymbol{Q}}_{\boldsymbol{k}}\}_{k\in\mathbb{N}}$, $\{\underline{f}_k^0, \underline{f}_k^s\}_{k\in\mathbb{N}}$, and $\{\hat{\underline{m}}_k\}_{k\in\mathbb{N}}$, the resulting stochastic process has desirable convergence properties, conditioned on the past, where *past* means the *past history* of the algorithm (see Remark 2.1). In particular, for the needs of convergence and expected complexity analyses presented in Sections 4 and 5, and as will be seen in Section 3, the random models $\hat{\underline{m}}_k$ and estimates $\underline{f}_k^0, \underline{f}_k^s$ are required to be probabilistically sufficiently accurate conditioned on the past, as was the case in [13]. Moreover, $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$ must guarantee a high quality of the selected subspace probabilistically and conditioned on the past, as will be detailed in Section 3.1.

For all $k \in \mathbb{N}$, denote by $(\underline{\boldsymbol{Q}}_{\boldsymbol{k}})_{ij}, (i,j) \in [\![1,n]\!] \times [\![1,p]\!] =: \mathbb{I}_{n,p}$ the entries of the random matrix $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$, where for $a \leq b$, $[\![a,b]\!] := [a,b] \cap \mathbb{Z}$ throughout the manuscript. To formalize conditioning on the past, we consider

[2] See Definition 3.2 for the meaning of $WAM(1-\varepsilon, \beta)$.

the three filtrations $\{\mathcal{F}_{k-1}\}_{k\in\mathbb{N}}$, $\left\{\mathcal{F}^{Q}_{k-1}\right\}_{k\in\mathbb{N}}$, and $\left\{\mathcal{F}^{\hat{m}\cdot Q}_{k-1}\right\}_{k\in\mathbb{N}}$ defined respectively by

$$\begin{array}{rcl}\mathcal{F}_{k-1} & := & \sigma\left(\underline{f}^0_\ell, \underline{f}^s_\ell, \underline{\hat{m}}_\ell, (\underline{\boldsymbol{Q}}_{\boldsymbol{\ell}})_{ij} \text{ with } \ell \in [\![0, k-1]\!] \ \& \ (i,j) \in \mathbb{I}_{n,p}\right)\\ \mathcal{F}^{Q}_{k-1} & := & \sigma\left(\underline{f}^0_\ell, \underline{f}^s_\ell, \underline{\hat{m}}_\ell, (\underline{\boldsymbol{Q}}_{\boldsymbol{\ell}})_{ij}, (\underline{\boldsymbol{Q}}_{\boldsymbol{k}})_{ij} \text{ with } \ell \in [\![0, k-1]\!] \ \& \ (i,j) \in \mathbb{I}_{n,p}\right)\\ \mathcal{F}^{\hat{m}\cdot Q}_{k-1} & := & \sigma\left(\underline{f}^0_\ell, \underline{f}^s_\ell, \underline{\hat{m}}_\ell, (\underline{\boldsymbol{Q}}_{\boldsymbol{\ell}})_{ij}, (\underline{\boldsymbol{Q}}_{\boldsymbol{k}})_{ij}, \ \underline{\hat{m}}_k \text{ with } \ell \in [\![0, k-1]\!] \ \& \ (i,j) \in \mathbb{I}_{n,p}\right).\end{array}$$

Here, for *completeness* [8], $\mathcal{F}_{-1} = \sigma(\boldsymbol{x_0})$. By construction, $\mathcal{F}_{k-1} \subset \mathcal{F}^{Q}_{k-1} \subset \mathcal{F}^{\hat{m}\cdot Q}_{k-1}$, $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$ is both $\mathcal{F}^{Q}_{k-1}$ and $\mathcal{F}^{\hat{m}\cdot Q}_{k-1}$-measurable, and $\underline{\hat{m}}_k$ is $\mathcal{F}^{\hat{m}\cdot Q}_{k-1}$-measurable, which imply that $\mathbb{E}\left[\underline{\boldsymbol{Q}}_{\boldsymbol{k}} | \mathcal{F}^{Q}_{k-1}\right] = \mathbb{E}\left[\underline{\boldsymbol{Q}}_{\boldsymbol{k}} | \mathcal{F}^{\hat{m}\cdot Q}_{k-1}\right] = \underline{\boldsymbol{Q}}_{\boldsymbol{k}}$ and $\mathbb{E}\left[\underline{\hat{m}}_k | \mathcal{F}^{\hat{m}\cdot Q}_{k-1}\right] = \underline{\hat{m}}_k$.

Remark 2.1. *At a given iteration k, when generating $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$, because no model or estimates were generated since the start of the iteration, it obviously follows from the construction of Algorithm 1 that its past history can be formalized by $\mathcal{F}_{k-1}$. On the other hand, when constructing $\underline{\hat{m}}_k$, the past of the algorithm that includes $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$, given that the latter is already generated since the beginning of iteration k, can therefore be formalized by $\mathcal{F}^{Q}_{k-1}$. A similar observation explains why $\mathcal{F}^{\hat{m}\cdot Q}_{k-1}$ can formalize the past history of Algorithm 1 before the construction of $\underline{f}^0_k$ and $\underline{f}^s_k$. These observations will also play an important role in the formalization of sufficient accuracy of the random models and estimates in Definitions 3.4 and 3.6 and the formalization of high quality of the random subspaces in Assumption 2.*

3 Subspace selection and probabilistic models and estimates in random subspaces.

The random subspaces in which the models $\underline{\hat{m}}_k$ are built are defined by specific random matrices. For efficiency of the proposed method and for convergence needs, not only must these subspaces be probabilistically rich enough, but also the accuracies of the models and estimates of unknown function values must hold with sufficiently high, but fixed, probabilities. This section discusses strategies for such random subspace selection, demonstrates by means of rigorous results how probabilistic estimates and subspace models can be constructed, and provides conditions that are necessary for the convergence of STARS.

3.1 Subspace selection strategies.

To choose a subspace, there must be enough gradient living in the subspace to ensure analogous reduction of f. In other words, the matrix $\boldsymbol{Q_k} \in \mathbb{R}^{n\times p}$ determining the subspace needs to satisfy (probabilistically) $\left\|\boldsymbol{Q_k}^\top \nabla f(\boldsymbol{x_k})\right\|_2 \geq \alpha_Q \|\nabla f(\boldsymbol{x_k})\|_2$ at every iteration, for some constant $\alpha_Q \in (0,1)$ independent of k. This requirement motivates the following modified definition suggested by [11, 32], which basically says that at least some fraction of the gradient of f must be maintained after projecting it into the selected subspace.

Definition 3.1. *For $\alpha_Q \in (0,1)$, a matrix $\boldsymbol{Q}$ is α_Q-well aligned if, for any vector $\boldsymbol{v} \in \mathbb{R}^n$, $\left\|\boldsymbol{Q}^\top \boldsymbol{v}\right\|_2 \geq \alpha_Q \|\boldsymbol{v}\|_2$.*

Recalling Remark 2.1 and inspired by [11], we propose a probabilistic quantification of the quality of the subspace selection as follows.

Definition 3.2. *For fixed $\alpha_Q, \beta_Q \in (0,1)$, a sequence $\left\{\underline{\boldsymbol{Q}}_{\boldsymbol{k}}\right\}_{k\in\mathbb{N}}$ of random matrices is $(1-\beta_Q)$-probabilistically α_Q-well aligned if, for any $\mathcal{F}_{k-1}$-measurable random vector $\underline{\boldsymbol{v}}$ with realizations $\boldsymbol{v} \in \mathbb{R}^n$, the events $\mathcal{A}_k := \left\{\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top \underline{\boldsymbol{v}}\right\|_2 \geq \alpha_Q \|\underline{\boldsymbol{v}}\|_2\right\}$ satisfy*

$$\mathbb{P}\left(\mathcal{A}_k | \mathcal{F}_{k-1}\right) = \mathbb{E}\left[\mathbb{1}_{\mathcal{A}_k} | \mathcal{F}_{k-1}\right] \geq 1 - \beta_Q, \tag{3}$$

where $\mathbb{1}_{\mathcal{A}_k}$ denotes the indicator function of the event $\mathcal{A}_k$; that is, $\mathbb{1}_{\mathcal{A}_k}(\omega) = 1$ if $\omega \in \mathcal{A}_k$ and $\mathbb{1}_{\mathcal{A}_k}(\omega) = 0$ otherwise. Random matrices satisfying (3) will be referred to as WAM(α_Q, β_Q).

For convergence purposes in Sections 4 and 5, the following will be assumed throughout the manuscript.

Assumption 2. *The sequence* $\{\underline{\boldsymbol{Q}}_{\boldsymbol{k}}\}_{k\in\mathbb{N}}$ *of matrices used by Algorithm 1 is* $(1-\beta_Q)$*-probabilistically* α_Q*-well aligned for some fixed* $\alpha_Q \in (0,1)$ *and* $\beta_Q \in (0,1/2)$.

The next result from [24, Lemma 1] provides a slight generalization of an exponential inequality for chi-square distributions. It will be required later for the proof of Theorem 3.1, inspired by [35, Theorem 2.1], providing a random *matrix ensemble* that will be shown in Corollary 3.1 to satisfy Assumption 2.

Lemma 3.1. *Let* $(\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_p)$ *be i.i.d. Gaussian random variables with mean zero and variance one. Let* $a_1, a_2, \ldots, a_p$ *be nonnegative real numbers. Let* $\underline{z}$ *be the random variable defined by* $\underline{z} = \sum_{i=1}^{p} a_i(\underline{y}_i^2 - 1)$. *Then it holds that for any* $t \geq 0$,

$$\mathbb{P}\left(\underline{z} \leq -2\|\boldsymbol{a}\|_2\sqrt{t}\right) \leq e^{-t}.$$

Theorem 3.1. *Let* $\varepsilon > 0$, $\beta < 1$, $p \geq 4\varepsilon^{-2}\log(1/\beta)$, *and* $\underline{\boldsymbol{R}} = \left(\underline{r}_{ij}\right)_{\substack{1\leq i\leq p\\ 1\leq j\leq n}} \in \mathbb{R}^{p\times n}$ *be a random matrix with i.i.d. standard Gaussian entries. Let* $\underline{\boldsymbol{S}} = \left(\underline{s}_{ij}\right)_{\substack{1\leq i\leq p\\ 1\leq j\leq n}} \in \mathbb{R}^{p\times n}$ *be a random matrix defined by* $\underline{\boldsymbol{S}} = \frac{1}{\sqrt{p}}\underline{\boldsymbol{R}}$; *that is,* $\underline{s}_{ij} = \frac{1}{\sqrt{p}}\underline{r}_{ij} \sim \mathcal{N}(0, 1/p)$. *Then for all* $\boldsymbol{v} \in \mathbb{R}^n$, $\mathbb{P}\left[\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2^2 \geq (1-\varepsilon)\|\boldsymbol{v}\|_2^2\right] \geq 1-\beta$.

Proof. The proof is inspired by that of [35, Lemma 2.12]. Consider any deterministic vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top \in \mathbb{R}^n$, and denote by $\underline{\boldsymbol{s}}_{:j} \in \mathbb{R}^p$, $j \in [\![1, n]\!]$ the columns of the random matrix $\underline{\boldsymbol{S}}$. Then the ith component of the p-tuple $\underline{\boldsymbol{S}}\boldsymbol{v} = \sum_{j=1}^{n} v_j \underline{\boldsymbol{s}}_{:j}$ is the Gaussian random variable $(\underline{\boldsymbol{S}}\boldsymbol{v})_i = \sum_{j=1}^{n} v_j \underline{s}_{ij}$ with mean zero and variance $\mathbb{V}\left[(\underline{\boldsymbol{S}}\boldsymbol{v})_i\right] = \sum_{j=1}^{n} v_j^2 \mathbb{V}\left(\underline{s}_{ij}\right) = \frac{\|\boldsymbol{v}\|_2^2}{p}$. It follows from the independence of $(\underline{\boldsymbol{S}}\boldsymbol{v})_i$, $i \in [\![1, p]\!]$ that the random variable $\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2^2 = \sum_{i=1}^{p}(\underline{\boldsymbol{S}}\boldsymbol{v})_i^2$ is equal in distribution to $\frac{\|\boldsymbol{v}\|_2^2}{p}\underline{w}$, where $\underline{w}$ is a χ_p^2 random variable with p degrees of freedom. Applying Lemma 3.1 with $t = \frac{\varepsilon^2 p}{4}$ and $a_i = 1$, $i \in [\![1, p]\!]$, yields $\mathbb{P}\left(\underline{w} - p \leq -\varepsilon p\right) \leq e^{-\frac{\varepsilon^2 p}{4}}$. Hence, $\mathbb{P}\left[\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2^2 \geq (1-\varepsilon)\|\boldsymbol{v}\|_2^2\right] = 1 - \mathbb{P}\left[\underline{w} \leq (1-\varepsilon)p\right] \geq 1 - e^{-\frac{\varepsilon^2 p}{4}}$. The proof is completed by noticing that $1 - 2e^{-\frac{\varepsilon^2 p}{4}} \geq 1-\beta$ if $p \geq 4\varepsilon^{-2}\log(1/\beta)$. □

The next corollary shows that Assumption 2 can be satisfied by using matrices resulting from Theorem 3.1, and whose number p of columns does not depend on n.

Corollary 3.1. *At a given iteration* k, *let the subspace selection matrix* $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top = \underline{\boldsymbol{S}} \in \mathbb{R}^{p\times n}$ *be provided by Theorem 3.1, with* $\varepsilon = 1 - \alpha_Q$ *and* $\beta = \beta_Q$, *for some* $\alpha_Q \in (0,1)$ *and* $\beta_Q \in (0, 1/2)$. *Assume that* $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$ *is independent of* $\underline{\boldsymbol{Q}}_{\boldsymbol{0}}, \underline{\boldsymbol{Q}}_{\boldsymbol{1}}, \ldots, \underline{\boldsymbol{Q}}_{\boldsymbol{k-1}}$ *and* $\underline{\boldsymbol{v}}$. *Then Assumption 2 holds; that is, for any* $\underline{\boldsymbol{v}}$ $\mathcal{F}_{k-1}$*-measurable,*

$$\mathbb{P}\left(\left\{\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top \underline{\boldsymbol{v}}\right\|_2 \geq \alpha_Q \|\underline{\boldsymbol{v}}\|_2\right\} | \mathcal{F}_{k-1}\right) \geq 1 - \beta_Q.$$

Proof. Since $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$ is independent of $\mathcal{F}_{k-1}$ and $\underline{\boldsymbol{v}}$ is $\mathcal{F}_{k-1}$-measurable, it suffices to show that $\mathbb{P}\left(\left\{\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top \boldsymbol{v}\right\|_2 \geq \alpha_Q\|\boldsymbol{v}\|_2\right\}\right) \geq 1 - \beta_Q$ for any deterministic vector $\boldsymbol{v} \in \mathbb{R}^n$, which easily follows from Theorem 3.1 and the inclusion $\left\{\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2^2 \geq (1-\varepsilon)\|\boldsymbol{v}\|_2^2\right\} \subseteq \left\{\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2 \geq (1-\varepsilon)\|\boldsymbol{v}\|_2\right\}$ due to the inequality $1 - \varepsilon \geq (1-\varepsilon)^2$. □

Inspired by [11], a technical assumption needed for convergence and expected complexity analyses later in Sections 4 and 5 is the following.

Assumption 3. *Any realization* $\boldsymbol{Q}_{\boldsymbol{k}}$ *of the random matrix used by Algorithm 1 satisfies* $\|\boldsymbol{Q}_{\boldsymbol{k}}\| \leq Q_{\max}$ *for all* $k \in \mathbb{N}$, *for some constant* $Q_{\max} > 0$ *independent of* k.

The next theorem, partially proved in the Appendix and formulated based on results from [21], provides another technique for constructing $WAM(1-\varepsilon, \beta)$ satisfying Assumption 2 through so-called Johnson–Lindenstrauss (JL) transforms [20, 21], and whose number p of columns does not depend on n. Moreover, one can easily see that the resulting matrix ensemble satisfies Assumption 3 with $Q_{\max} = \sqrt{n}$, unlike the one from Theorem 3.1 where $\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}\right\| \leq Q_{\max}$ holds with high probability with[3] $Q_{\max} = \Theta(\sqrt{n/p})$ [4, Corollary 3.11].

[3]The reader is referred to [22] for details regarding the notation $\Theta(\cdot)$.

Theorem 3.2. *Let $\varepsilon > 0$, $\beta < 1/2$, and $\underline{\boldsymbol{S}} = \left(\underline{s}_{ij}\right)_{\substack{1\le i\le p\\ 1\le j\le n}} \in \mathbb{R}^{p\times n}$ be a real-valued random matrix defined by $\underline{s}_{ij} = \frac{1}{\sqrt{r}}\underline{\eta}_{ij}\underline{\sigma}_{ij}$, where $r \in \mathbb{N}\setminus\{0\}$ and $\underline{\sigma}_{ij}$ are independent Rademacher random variables. The $\underline{\eta}_{ij}$ satisfying $\mathbb{P}\left(\sum_{i=1}^{p}\underline{\eta}_{ij} = r\right) = 1$ for all $j \in [\![1,n]\!]$ are negatively correlated indicator random variables for the events $\left\{\underline{s}_{ij}\neq 0\right\}$; that is, they satisfy $\mathbb{E}\left[\prod_{(i,j)\in T}\underline{\eta}_{ij}\right] \le (r/p)^{|T|}$ for all $T \subseteq [\![1,p]\!]\times[\![1,n]\!]$ with $|T|\le \ell$, where $\ell \ge \ln\left[\beta^{-1}(\ell+1)\ell/2\right]$. Then $\mathbb{P}\left[(1-\varepsilon)\|\boldsymbol{v}\|_2 \le \|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2 \le (1+\varepsilon)\|\boldsymbol{v}\|_2\right] > 1-\beta$ for any $\boldsymbol{v}\in\mathbb{R}^n$, provided $r \ge 8e^4\sqrt{e}(\ell+1)/(2\varepsilon-\varepsilon^2)$ and $p = 2r^2/(e\ell)$.*

The next corollary shows that *WAM*(α_Q,β_Q) satisfying Assumptions 2 and 3 can be provided by specific JL transforms $\underline{\boldsymbol{S}}$ resulting from Theorem 3.2.

Corollary 3.2. *At a given iteration k, let the subspace selection matrix $\boldsymbol{Q}_{\boldsymbol{k}}^{\top} = \underline{\boldsymbol{S}} \in \mathbb{R}^{p\times n}$ be provided by Theorem 3.2, with $\varepsilon = 1-\alpha_Q$ and $\beta = \beta_Q$. Assume that $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$ is independent of $\underline{\boldsymbol{Q}}_{\boldsymbol{0}}, \underline{\boldsymbol{Q}}_{\boldsymbol{1}}, \ldots, \underline{\boldsymbol{Q}}_{\boldsymbol{k-1}}$ and $\underline{\theta}$. Then Assumption 2 holds.*

Proof. As was explained in the proof of Corollary 3.1, it suffices to show that for all deterministic vectors $\boldsymbol{v}\in\mathbb{R}^n$, $\mathbb{P}\left(\left\{\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^{\top}\boldsymbol{v}\right\|_2 \ge \alpha_Q\|\boldsymbol{v}\|_2\right\}\right) \ge 1-\beta_Q$. This immediately follows from the result of Theorem 3.2, together with the inclusion $\{(1-\varepsilon)\|\boldsymbol{v}\|_2 \le \|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2 \le (1+\varepsilon)\|\boldsymbol{v}\|_2\} \subseteq \{\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2 \ge (1-\varepsilon)\|\boldsymbol{v}\|_2\}$. □

3.2 Probabilistic models and estimates in random subspaces.

As mentioned at the beginning of Section 3, the models used by Algorithm 1 need to be sufficiently accurate. Motivated by [13], this sufficient accuracy is formalized for deterministic subspace models by the following measure of accuracy inspired by [11].

Definition 3.3. *Assume ∇f is Lipschitz continuous. Given $\boldsymbol{Q}_{\boldsymbol{k}} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{x}_{\boldsymbol{k}} \in \mathbb{R}^n$, a function $\hat{m}_k : \mathbb{R}^p \to \mathbb{R}$ is a $(\kappa_{ef},\kappa_{eg};\boldsymbol{Q}_{\boldsymbol{k}})$-fully linear model of f in $\mathcal{B}(\boldsymbol{x}_{\boldsymbol{k}},\delta_k;\boldsymbol{Q}_{\boldsymbol{k}})$ for some $\delta_k > 0$ if there exist constants $\kappa_{ef},\kappa_{eg} > 0$ independent of k such that for all $\boldsymbol{s}\in\mathbb{R}^p$ with $\|\boldsymbol{s}\|_2 \le \delta_k$,*

$$|f(\boldsymbol{x}_{\boldsymbol{k}}+\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}) - \hat{m}_k(\boldsymbol{s})| \le \kappa_{ef}\delta_k^2 \quad \text{and} \quad \left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}}+\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}) - \nabla\hat{m}_k(\boldsymbol{s})\right\|_2 \le \kappa_{eg}\delta_k. \tag{4}$$

As pointed out in [11], the gradient condition in (4) is due to the fact that given any fixed $\boldsymbol{x}$ and $\boldsymbol{Q}$, if $\hat{f}(\boldsymbol{s}) := f(\boldsymbol{x}+\boldsymbol{Q}\boldsymbol{s})$, then $\nabla_{\boldsymbol{s}}\hat{f}(\boldsymbol{s}) = \boldsymbol{Q}^{\top}\nabla_{\boldsymbol{x}}f(\boldsymbol{x}+\boldsymbol{Q}\boldsymbol{s})$. Moreover, denoting by $\boldsymbol{I}$ the identity matrix in full-dimensional subspaces where $p=n$, then $(\kappa_{ef},\kappa_{eg};\boldsymbol{I})$-fully linear models correspond to standard fully linear models [14, Definition 6.1].

Recalling Remark 2.1, the next definition formalizing sufficient accuracy of probabilistic subspace models $\hat{\underline{m}}_k$ is a stochastic variant of Definition 3.3, which generalizes [13, Definition 3.4] to low-dimensional subspaces. The existence of such models will be rigorously demonstrated in Theorem 3.3 and Corollary 3.3.

Definition 3.4. *A sequence $\{\hat{\underline{m}}_k\}$ of random models is β_m-probabilistically $(\kappa_{ef},\kappa_{eg};\boldsymbol{Q}_{\boldsymbol{k}})$-fully linear with respect to the random sequence $\left\{\mathcal{B}(\underline{\boldsymbol{x}}_{\boldsymbol{k}},\underline{\delta}_k;\underline{\boldsymbol{Q}}_{\boldsymbol{k}})\right\}$ if the events*

$$I_k^Q = \left\{\text{Given } \boldsymbol{Q}_{\boldsymbol{k}} \text{ and } \boldsymbol{x}_{\boldsymbol{k}}, \hat{\underline{m}}_k \text{ is } (\kappa_{ef},\kappa_{eg};\boldsymbol{Q}_{\boldsymbol{k}})\text{-fully linear for } f \text{ in } \mathcal{B}(\underline{\boldsymbol{x}}_{\boldsymbol{k}},\underline{\delta}_k;\underline{\boldsymbol{Q}}_{\boldsymbol{k}})\right\}$$

satisfy the supermartingale-like condition

$$\mathbb{P}\left(I_k^Q|\mathcal{F}_{k-1}^Q\right) = \mathbb{E}\left[\mathbb{1}_{I_k^Q}|\mathcal{F}_{k-1}^Q\right] \ge \beta_m. \tag{5}$$

Recalling that estimates of unknown function values also need to be sufficiently accurate, we introduce a formalization of sufficient accuracy inspired by [2, 13, 16, 17, 28] and motivated by [11] as follows.

Definition 3.5. *Given $\varepsilon_f > 0$ and $\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{k}} \in \mathbb{R}^n$, $\boldsymbol{x}_{\boldsymbol{k}} \in \mathbb{R}^n$, f_k^0 and f_k^s are called ε_f-accurate estimates of $f(\boldsymbol{x}_{\boldsymbol{k}})$ and $f(\boldsymbol{x}_{\boldsymbol{k}}+\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{k}})$, respectively, for a given δ_k if*

$$\left|f_k^0 - f(\boldsymbol{x}_{\boldsymbol{k}})\right| \le \varepsilon_f\delta_k^2 \quad \text{and} \quad |f_k^s - f(\boldsymbol{x}_{\boldsymbol{k}}+\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{k}})| \le \varepsilon_f\delta_k^2. \tag{6}$$

Recalling Remark 2.1, we extend this definition to the following stochastic variant.

Definition 3.6. *A sequence of random estimates* $\{\underline{f}_k^0, \underline{f}_k^s\}$ *is said to be* β_f*-probabilistically* ε_f*-accurate with respect to the corresponding sequence* $\{\underline{\boldsymbol{x}}_{\boldsymbol{k}}, \underline{\delta}_k, \underline{\boldsymbol{Q}}_{\boldsymbol{k}}\underline{\boldsymbol{s}}_{\boldsymbol{k}}\}$ *if the events*

$$J_k^Q \quad = \quad \{\text{Given } \boldsymbol{Q_k} \text{ and } \boldsymbol{x_k}, \underline{f}_k^0 \text{ and } \underline{f}_k^s \text{ are } \varepsilon_f\text{-accurate estimates of } f(\boldsymbol{x_k}) \text{ and } f(\boldsymbol{x_k} + \boldsymbol{Q_k s_k}), \text{ respectively, for } \underline{\delta}_k\}$$

satisfy the supermartingale-like condition $\mathbb{P}\left(J_k^Q | \mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right) \geq \beta_f$.

Based on Definitions 3.2, 3.4, and 3.6, *true* iterations, in other words, those for which $\mathbb{1}_{\mathcal{A}_k}\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q} = 1$, occur with a total probability of at least $\tilde{\beta} := \beta_f\beta_m(1-\beta_Q)$ conditioned on $\mathcal{F}_{k-1}$. Noting that by construction $\mathbb{1}_{\mathcal{A}_k}$ is both $\mathcal{F}_{k-1}^Q$-measurable and $\mathcal{F}_{k-1}^{\hat{m}\cdot Q}$-measurable since $\mathcal{F}_{k-1}^Q \subset \mathcal{F}_{k-1}^{\hat{m}\cdot Q}$, and that $\mathbb{1}_{I_k^Q}$ is $\mathcal{F}_{k-1}^{\hat{m}\cdot Q}$-measurable, then for $F := \left\{\mathcal{A}_k \cap I_k^Q \cap J_k^Q\right\}$, we have

$$\begin{aligned}\mathbb{P}(F|\mathcal{F}_{k-1}) &= \mathbb{E}\left[\mathbb{1}_{\mathcal{A}_k}\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q}|\mathcal{F}_{k-1}\right] = \mathbb{E}\left[\mathbb{1}_{\mathcal{A}_k}\mathbb{1}_{I_k^Q}\,\mathbb{E}\left[\mathbb{1}_{J_k^Q}|\mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right]|\mathcal{F}_{k-1}\right] \\ &\geq \beta_f\,\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_k}\,\mathbb{E}\left[\mathbb{1}_{I_k^Q}|\mathcal{F}_{k-1}^Q\right]|\mathcal{F}_{k-1}\right] \geq \beta_f\beta_m\,\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_k}|\mathcal{F}_{k-1}\right] \geq \tilde{\beta}.\end{aligned} \tag{7}$$

The second equality and the first two inequalities used the *William's Tower Property* for conditional expectations [8, Theorem 34.4.], namely, the fact that if $\underline{z}$ is integrable and the σ-algebras $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy $\mathcal{G}_1 \subset \mathcal{G}_2$, then $\mathbb{E}\left[\underline{z}|\mathcal{G}_1\right] = \mathbb{E}\left[\mathbb{E}\left[\underline{z}|\mathcal{G}_2\right]|\mathcal{G}_1\right]$. Moreover, for a σ-algebra $\mathcal{G}$, $\mathbb{E}\left[\underline{r}\underline{u}|\mathcal{G}\right] = \underline{r}\,\mathbb{E}\left[\underline{u}|\mathcal{G}\right]$ if $\underline{r}$ is $\mathcal{G}$-measurable and $\underline{r}$ and $\underline{r}\underline{u}$ are integrable [8, Theorem 34.3.].

Note that even though the present algorithmic framework does not distinguish true iterations from *false* ones, it holds that $\mathbb{P}\left(\limsup_{k\to\infty} \underline{w}_k = \infty\right) = 1$, where

$$\underline{w}_k := \sum_{i=0}^{k}\left(2\mathbb{1}_{\mathcal{A}_i}\mathbb{1}_{I_i^Q}\mathbb{1}_{J_i^Q} - 1\right), \tag{8}$$

as will be seen later in the proof of Theorem 4.2 provided $\tilde{\beta} > 1/2$. This shows that true iterations occur sufficiently often for the convergence of Algorithm 1 to hold (see, e.g., [17, Theorem 3.6] for a similar result).

Remark 3.1. *Calculations similar to those in* (7) *easily yield*

$$\mathbb{P}\left(I_k^Q \cap J_k^Q|\mathcal{F}_{k-1}^Q\right) \geq \beta_f\beta_m, \qquad \mathbb{P}\left(J_k^Q|\mathcal{F}_{k-1}^Q\right) \geq \beta_f, \tag{9}$$

$$\mathbb{P}\left(I_k^Q|\mathcal{F}_{k-1}\right) \geq \beta_m \quad \text{and} \quad \mathbb{P}\left(J_k^Q|\mathcal{F}_{k-1}\right) \geq \beta_f, \tag{10}$$

which will be useful in the proofs of Theorems 4.1 and 4.2, respectively.

3.3 Construction of probabilistic estimates and models in random subspaces.

The construction of probabilistic estimates of Definition 3.6 trivially follows the strategies described in [2, Section 2.3], [13, Section 5], and [17, Section 5.1] and hence are not presented here again.

Next is stated an extension to subspaces of [3, Lemma 9.4], which will be useful for the proof of Theorem 3.3, one of the main results of the present work. Its proof is presented in the Appendix.

Lemma 3.2. *Let* f *be differentiable with a* L_g*-Lipschitz continuous gradient, and let* $\boldsymbol{s}, \boldsymbol{d} \in \mathbb{R}^p$. *For a given matrix* $\boldsymbol{Q_k}$ *and vector* $\boldsymbol{x_k}$, *let* $\hat{f}(\boldsymbol{s}) = f(\boldsymbol{x_k} + \boldsymbol{Q_k s})$, *and assume that Assumption 3 holds. Then*

$$\left|\hat{f}(\boldsymbol{s}+\boldsymbol{d}) - \hat{f}(\boldsymbol{s}) - \boldsymbol{d}^\top\nabla\hat{f}(\boldsymbol{s})\right| \leq \frac{1}{2}L_g Q_{\max}^2\|\boldsymbol{d}\|_2^2.$$

The next result, which is a stochastic variant of [3, Theorem 9.5], shows, together with Corollary 3.3, how β_m-probabilistically $(\kappa_{ef}, \kappa_{eg}; \boldsymbol{Q_k})$-fully linear models of Definition 3.4 can be made available.

Theorem 3.3. *Let the assumptions of Lemma 3.2 hold. Assume that there exists a finite constant $V_f > 0$ such that $\mathbb{V}_{\underline{\theta}}[f_{\underline{\theta}}(\boldsymbol{x})] \leq V_f$ for all $\boldsymbol{x} \in \mathbb{R}^n$. For all $i = 0, 1, \ldots, p$, let $\underline{\theta}^i_\ell, \ell = 1, \ldots, \pi_k$ be independent random samples of the independent random variables $\underline{\theta}^i$ following the same distribution as $\underline{\theta}$, and define $\underline{f}^s_k(\boldsymbol{s_i}, \underline{\theta}^i) := \frac{1}{\pi_k} \sum_{\ell=1}^{\pi_k} f_{\underline{\theta}^i_\ell}(\boldsymbol{x_k} + \boldsymbol{Q_k s_i})$, where the points $\mathbb{S}^k := \{\boldsymbol{s_0}, \boldsymbol{s_1}, \ldots, \boldsymbol{s_p}\} \subseteq \mathbb{R}^p \cap \mathcal{B}(\boldsymbol{0}, c_1\delta_k)$ are affinely independent with $\boldsymbol{s_0} = \boldsymbol{0}$ and $c_1 \geq 1$ a constant. Let $\rho_k := \max_{\boldsymbol{s_i} \in \mathbb{S}^k} \|\boldsymbol{s_i} - \boldsymbol{s_0}\|_2$ denote the approximate diameter of $\mathbb{S}^k$ and define $\hat{\boldsymbol{L}}_{\boldsymbol{k}} := \frac{1}{\rho_k} [\boldsymbol{s_1} - \boldsymbol{s_0} \quad \cdots \quad \boldsymbol{s_p} - \boldsymbol{s_0}] \in \mathbb{R}^{p \times p}$. Let $\boldsymbol{1} := [1, \ldots, 1]^\top \in \mathbb{R}^{p+1}$, and define $\boldsymbol{Y} := [\boldsymbol{s_0} \ \ \boldsymbol{s_1} \ \cdots \ \boldsymbol{s_p}] \in \mathbb{R}^{p \times (p+1)}$. Consider the random vector $\underline{f}_k(\mathbb{S}^k, \underline{\theta}) := \left[\underline{f}^s_k(\boldsymbol{s_0}, \underline{\theta}^0), \underline{f}^s_k(\boldsymbol{s_1}, \underline{\theta}^1), \ldots, \underline{f}^s_k(\boldsymbol{s_p}, \underline{\theta}^p)\right]^\top \in \mathbb{R}^{p+1}$ and the random linear model $\underline{\hat{m}}_k(\boldsymbol{s}) := \underline{a}_0 + \underline{\boldsymbol{a}}^\top \boldsymbol{s}$, where $(\underline{a}_0, \underline{\boldsymbol{a}}) \in \mathbb{R} \times \mathbb{R}^p$ is the unique solution of $\begin{bmatrix} \boldsymbol{1} & \boldsymbol{Y}^\top \end{bmatrix} \begin{bmatrix} \underline{a}_0 \\ \underline{\boldsymbol{a}} \end{bmatrix} = \underline{f}_k(\mathbb{S}^k, \underline{\theta})$. Defining*[4]

$$\kappa_{ef} := \frac{1}{2}\left(1 + \frac{1}{2}c_1^2\delta_{\max} + 2c_1\sqrt{p}\left\|\hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1}\right\|\right) L_g Q_{\max}^2, \qquad \kappa_{eg} := \left(1 + c_1\sqrt{p}\left\|\hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1}\right\|\right) L_g Q_{\max}^2, \tag{11}$$

consider the events

$$E_f := \left\{\left|\hat{f}(\boldsymbol{s}) - \underline{\hat{m}}_k(\boldsymbol{s})\right| \leq \kappa_{ef}\delta_k^2\right\}, \qquad E_g := \left\{\left\|\nabla \hat{f}(\boldsymbol{s}) - \nabla \underline{\hat{m}}_k(\boldsymbol{s})\right\|_2 \leq \kappa_{eg}\delta_k\right\}.$$

Then

$$\mathbb{P}\left[E_f \cap E_g\right] \geq \beta_m \quad \text{for some } \beta_m \in (0,1) \quad \text{and all} \quad \|\boldsymbol{s}\|_2 \leq \delta_k, \tag{12}$$

provided

$$\pi_k \geq \frac{16 V_f}{c_1^2 L_g^2 Q_{\max}^4 \rho_k^2 \min\{\delta_k^2, \delta_k^4\}\left(1 - \beta_m^{1/(p+1)}\right)}. \tag{13}$$

Proof. We build the random linear model $\underline{\hat{m}}_k(\boldsymbol{s}) = \underline{a}_0 + \underline{\boldsymbol{a}}^\top \boldsymbol{s}$ by seeking values for $(\underline{a}_0, \underline{\boldsymbol{a}}) \in \mathbb{R} \times \mathbb{R}^p$ such that $\underline{\hat{m}}_k(\boldsymbol{s_i}) = \underline{f}^s_k(\boldsymbol{s_i}, \underline{\theta}^i)$ for all $\boldsymbol{s_i}$ in the interpolation set $\mathbb{S}^k$, which is equivalent to the $(p+1) \times (p+1)$ linear system of equations

$$\begin{bmatrix} \boldsymbol{1} & \boldsymbol{Y}^\top \end{bmatrix} \begin{bmatrix} \underline{a}_0 \\ \underline{\boldsymbol{a}} \end{bmatrix} = \underline{f}_k(\mathbb{S}^k, \underline{\theta}). \tag{14}$$

Since the points in $\mathbb{S}^k$ are affinely independent, the matrix $\hat{\boldsymbol{L}}_{\boldsymbol{k}}$ is invertible (see, e.g., [3, Proposition 9.1]). Hence, system (14) has a unique solution. Moreover, $\underline{\boldsymbol{a}} = \nabla \underline{\hat{m}}_k(\boldsymbol{s})$ can also be computed by solving the $p \times p$ linear system

$$\rho_k \hat{\boldsymbol{L}}_{\boldsymbol{k}}^\top \underline{\boldsymbol{a}} = \underline{\boldsymbol{\delta}}^{\boldsymbol{f_k}(\mathbb{S}^k, \underline{\boldsymbol{\theta}})}, \quad \text{where} \quad \underline{\boldsymbol{\delta}}^{\boldsymbol{f_k}(\mathbb{S}^k, \underline{\boldsymbol{\theta}})} = \begin{bmatrix} \underline{f}^s_k(\boldsymbol{s_1}, \underline{\theta}^1) - \underline{f}^s_k(\boldsymbol{s_0}, \underline{\theta}^0) \\ \underline{f}^s_k(\boldsymbol{s_2}, \underline{\theta}^2) - \underline{f}^s_k(\boldsymbol{s_0}, \underline{\theta}^0) \\ \vdots \\ \underline{f}^s_k(\boldsymbol{s_p}, \underline{\theta}^p) - \underline{f}^s_k(\boldsymbol{s_0}, \underline{\theta}^0) \end{bmatrix} \in \mathbb{R}^p. \tag{15}$$

To prove (12), we consider the events

$$B_i := \left\{\left|\underline{f}^s_k(\boldsymbol{s_i}, \underline{\theta}^i) - \hat{f}(\boldsymbol{s_i})\right| \leq \frac{1}{2}\kappa'_{eg}\rho_k \min\{\delta_k, \delta_k^2\}\right\}, \quad i = 0, 1, \ldots, p,$$

where $\kappa'_{eg} := \frac{1}{2}c_1 L_g Q_{\max}^2$; and we assume that

$$\mathbb{P}\left(B_i\right) \geq \beta_m^{1/(p+1)} \quad \text{for some} \quad \beta_m \in (0,1) \text{ and all } i = 0, 1, \ldots, p. \tag{16}$$

[4] While κ_{ef} and κ_{eg} seem to always depend on k, this is not the case since $\left\|\hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1}\right\|$ can be controlled by the geometry of the set $\mathbb{S}^k$, as will be seen later by means of Corollary 3.3.

It follows from the independence of the random variables $\underline{f}_k^s(\boldsymbol{s_i}, \underline{\theta}^i)$, $i = 0, 1, \ldots, p$, that the event $B := \bigcap_{i=0}^p B_i$ satisfies $\mathbb{P}(B) = \prod_{i=0}^p \mathbb{P}(B_i) \geq \beta_m$. Then the remainder of the proof considers three parts. The first two show respectively that $B \subseteq E_g$ and $B \subseteq E_f$, which imply that $B \subseteq E_f \cap E_g$ and hence $\mathbb{P}(E_f \cap E_g) \geq \beta_m$. Part 3 provides the condition under which (16) holds.

Part 1 ($B \subseteq E_g$). To demonstrate the latter inclusion, we will show that

$$B \subseteq E_1 := \left\{ \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^\top \left[\underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0}) \right] \right\|_2 \leq 2\sqrt{p}\kappa'_{eg}\delta_k \right\}. \tag{17}$$

Then, since the matrix norm is consistent with the Euclidean norm, the inequality

$$\left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-\top} \hat{\boldsymbol{L}}_{\boldsymbol{k}} \left(\underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0}) \right) \right\|_2 \leq \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-\top} \right\| \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}} \left(\underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0}) \right) \right\|_2$$

implies that $E_1 \subseteq E_2 := \left\{ \left\| \underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0}) \right\|_2 \leq 2\sqrt{p}\kappa'_{eg} \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1} \right\| \delta_k \right\}$, where we used the fact that $\left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-\top} \right\| = \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1} \right\|$. Using the L_g-Lipschitz continuity of ∇f and the fact that $\nabla \hat{f}(\boldsymbol{s}) = \boldsymbol{Q}_{\boldsymbol{k}}^\top \nabla f(\boldsymbol{x_k} + \boldsymbol{Q_k s})$ with $\|\boldsymbol{Q_k}\| \leq Q_{\max}$, we get

$$\begin{aligned}
\left\| \nabla \hat{f}(\boldsymbol{s}) - \underline{\boldsymbol{a}} \right\|_2 &\leq \left\| \underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0}) \right\|_2 + \left\| \nabla \hat{f}(\boldsymbol{s_0}) - \nabla \hat{f}(\boldsymbol{s}) \right\|_2 \\
&\leq 2\sqrt{p}\kappa'_{eg} \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1} \right\| \delta_k + L_g \left\| \boldsymbol{Q}_{\boldsymbol{k}}^\top \right\| \|\boldsymbol{Q_k}\| \|\boldsymbol{s_0} - \boldsymbol{s}\|_2 \\
&\leq \left(2\sqrt{p}\kappa'_{eg} \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1} \right\| + L_g Q_{\max}^2 \right) \delta_k = \left(1 + c_1\sqrt{p} \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1} \right\| \right) L_g Q_{\max}^2 \delta_k,
\end{aligned}$$

which implies that $E_2 \subseteq E_g := \left\{ \left\| \nabla \hat{f}(\boldsymbol{s}) - \underline{\boldsymbol{a}} \right\|_2 \leq \left(1 + c_1\sqrt{p} \left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1} \right\| \right) L_g Q_{\max}^2 \delta_k \right\}$.

To show (17), we first notice, using (15), that $\hat{\boldsymbol{L}}_{\boldsymbol{k}}^\top \underline{\boldsymbol{a}} = \frac{1}{\rho_k} \underline{\boldsymbol{\delta}}^{\boldsymbol{f_k}(\mathbb{S}^{\boldsymbol{k}}, \underline{\boldsymbol{\theta}})}$, and then

$$\hat{\boldsymbol{L}}_{\boldsymbol{k}}^\top \nabla \hat{f}(\boldsymbol{s_0}) = \frac{1}{\rho_k} \left[(\boldsymbol{s_1} - \boldsymbol{s_0})^\top \nabla \hat{f}(\boldsymbol{s_0}), \ \ldots, \ (\boldsymbol{s_p} - \boldsymbol{s_0})^\top \nabla \hat{f}(\boldsymbol{s_0}) \right]^\top \in \mathbb{R}^p.$$

Thus, the ith component of the vector $\hat{\boldsymbol{L}}_{\boldsymbol{k}}^\top \left(\underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0}) \right)$ is given by

$$\underline{\Psi}_k(\boldsymbol{s_i}, \boldsymbol{s_0}) := \frac{1}{\rho_k} \left(\underline{f}_k^s(\boldsymbol{s_i}, \underline{\theta}^i) - \underline{f}_k^s(\boldsymbol{s_0}, \underline{\theta}^0) - (\boldsymbol{s_i} - \boldsymbol{s_0})^\top \nabla \hat{f}(\boldsymbol{s_0}) \right).$$

We notice that $|\underline{\Psi}_k(\boldsymbol{s_i}, \boldsymbol{s_0})| \leq \underline{\Psi}_k^{1,i} + \psi_k^{2,i}$, where

$$\begin{aligned}
\underline{\Psi}_k^{1,i} &:= \frac{1}{\rho_k} \left| \underline{f}_k^s(\boldsymbol{s_i}, \underline{\theta}^i) - \hat{f}(\boldsymbol{s_i}) \right| + \frac{1}{\rho_k} \left| \underline{f}_k^s(\boldsymbol{s_0}, \underline{\theta}^0) - \hat{f}(\boldsymbol{s_0}) \right| \\
\text{and} \quad \psi_k^{2,i} &:= \frac{1}{\rho_k} \left| \hat{f}(\boldsymbol{s_i}) - \hat{f}(\boldsymbol{s_0}) - (\boldsymbol{s_i} - \boldsymbol{s_0})^\top \nabla \hat{f}(\boldsymbol{s_0}) \right|.
\end{aligned}$$

It follows from Lemma 3.2 that

$$\psi_k^{2,i} \leq \frac{1}{2\rho_k} L_g Q_{\max}^2 \|\boldsymbol{s_i} - \boldsymbol{s_0}\|_2^2 \leq \frac{1}{2} L_g Q_{\max}^2 \rho_k \leq \frac{1}{2} L_g Q_{\max}^2 c_1 \delta_k. \tag{18}$$

Now assume that the event B occurs. Then $\underline{\Psi}_k^{1,i} \leq \kappa'_{eg}\delta_k$ for all $i = 0, 1, \ldots, p$. It follows from (18) and the inequality $|\underline{\Psi}_k(\boldsymbol{s_i}, \boldsymbol{s_0})|^2 \leq 2\left(\left(\underline{\Psi}_k^{1,i} \right)^2 + \left(\psi_k^{2,i} \right)^2 \right)$ that

$$\begin{aligned}
\left\| \hat{\boldsymbol{L}}_{\boldsymbol{k}}^\top \left(\underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0}) \right) \right\|_2^2 &= \sum_{i=1}^p \underline{\Psi}_k(\boldsymbol{s_i}, \boldsymbol{s_0})^2 \\
&\leq 2p\kappa'^2_{eg}\delta_k^2 + \frac{1}{2} p L_g^2 Q_{\max}^4 c_1^2 \delta_k^2 = p L_g^2 Q_{\max}^4 c_1^2 \delta_k^2,
\end{aligned}$$

where the last equality follows from the definition of κ'_{eg}. This means that $B \subseteq E_1$.

Part 2 ($B \subseteq E_f$). To show the latter inclusion, we recall that $\hat{\underline{m}}_k(\boldsymbol{s_0}) = \underline{f}_k^s(\boldsymbol{s_0}, \underline{\theta}^0)$ and $\hat{\underline{m}}_k(\boldsymbol{s_0}) - \hat{\underline{m}}_k(\boldsymbol{s}) = (\boldsymbol{s_0} - \boldsymbol{s})^\top \underline{\boldsymbol{a}}$. Then the following holds:

$$\begin{aligned}
\left|\hat{f}(\boldsymbol{s}) - \hat{\underline{m}}_k(\boldsymbol{s})\right| &= \left|\hat{f}(\boldsymbol{s}) - \underline{f}_k^s(\boldsymbol{s_0}, \underline{\theta}^0) + \hat{\underline{m}}_k(\boldsymbol{s_0}) - \hat{\underline{m}}_k(\boldsymbol{s})\right| \\
&= \left|\hat{f}(\boldsymbol{s}) - \underline{f}_k^s(\boldsymbol{s_0}, \underline{\theta}^0) + (\boldsymbol{s_0} - \boldsymbol{s})^\top \underline{\boldsymbol{a}}\right| \\
&= \left|\hat{f}(\boldsymbol{s}) - \hat{f}(\boldsymbol{s_0}) - (\boldsymbol{s} - \boldsymbol{s_0})^\top \nabla \hat{f}(\boldsymbol{s_0}) - \left(\underline{f}_k^s(\boldsymbol{s_0}, \underline{\theta}^0) - \hat{f}(\boldsymbol{s_0})\right) + (\boldsymbol{s} - \boldsymbol{s_0})^\top \left(\nabla \hat{f}(\boldsymbol{s_0}) - \underline{\boldsymbol{a}}\right)\right| \\
&\leq \left|\hat{f}(\boldsymbol{s}) - \hat{f}(\boldsymbol{s_0}) - (\boldsymbol{s} - \boldsymbol{s_0})^\top \nabla \hat{f}(\boldsymbol{s_0})\right| + \left|\underline{f}_k^s(\boldsymbol{s_0}, \underline{\theta}^0) - \hat{f}(\boldsymbol{s_0})\right| + \|\boldsymbol{s} - \boldsymbol{s_0}\|_2 \left\|\underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0})\right\|_2.
\end{aligned}$$

As before, the first term on the right-hand side of the last inequality is bounded by using Lemma 3.2 as follows:

$$\left|\hat{f}(\boldsymbol{s}) - \hat{f}(\boldsymbol{s_0}) - (\boldsymbol{s} - \boldsymbol{s_0})^\top \nabla \hat{f}(\boldsymbol{s_0})\right| \leq \frac{1}{2} L_g Q_{\max}^2 \|\boldsymbol{s} - \boldsymbol{s_0}\|_2^2 \leq \frac{1}{2} L_g Q_{\max}^2 \delta_k^2. \tag{19}$$

To bound the last two terms, we note that the inclusions $B \subseteq B_0$ and $B \subseteq E_1 \subseteq E_2$ yield $\left|\underline{f}_k^s(\boldsymbol{s_0}, \underline{\theta}^0) - \hat{f}(\boldsymbol{s_0})\right| \leq \frac{1}{2}\kappa'_{eg}\rho_k\delta_k^2$ and $\left\|\underline{\boldsymbol{a}} - \nabla \hat{f}(\boldsymbol{s_0})\right\|_2 \leq 2\sqrt{p}\kappa'_{eg}\left\|\hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1}\right\|\delta_k$, respectively. Thus, these inequalities combined with (19) and the inequalities $\rho_k \leq c_1\delta_{\max}$ and $\|\boldsymbol{s} - \boldsymbol{s_0}\|_2 \leq \delta_k$ lead to

$$\begin{aligned}
\left|\hat{f}(\boldsymbol{s}) - \hat{\underline{m}}_k(\boldsymbol{s})\right| &\leq \frac{1}{2} L_g Q_{\max}^2 \delta_k^2 + \frac{1}{2}\kappa'_{eg}\rho_k\delta_k^2 + 2\sqrt{p}\kappa'_{eg}\left\|\hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1}\right\|\delta_k^2 \\
&= \frac{1}{2}\left(1 + \frac{1}{2}c_1^2\delta_{\max} + 2c_1\sqrt{p}\left\|\hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1}\right\|\right) L_g Q_{\max}^2 \delta_k^2,
\end{aligned}$$

which shows that $B \subseteq E_f$.

Part 3 To complete the proof, we show (16) using the Chebyshev inequality as follows:

$$\begin{aligned}
\mathbb{P}\left(B_i^c\right) &= \mathbb{P}\left(\left|\underline{f}_k^s(\boldsymbol{s_i}, \underline{\theta}^i) - \mathbb{E}\left[\underline{f}_k^s(\boldsymbol{s_i}, \underline{\theta}^i)\right]\right| > \frac{1}{2}\kappa'_{eg}\rho_k \min\{\delta_k, \delta_k^2\}\right) \\
&\leq \frac{4\mathbb{V}\left[\underline{f}_k^s(\boldsymbol{s_i}, \underline{\theta}^i)\right]}{\kappa'^2_{eg}\rho_k^2 \min\{\delta_k^2, \delta_k^4\}} \leq \frac{4V_f}{\pi_k\kappa'^2_{eg}\rho_k^2 \min\{\delta_k^2, \delta_k^4\}} \leq 1 - \beta_m^{1/(p+1)},
\end{aligned}$$

provided (13) holds; that is, π_k is chosen according to

$$\pi_k \geq \frac{4V_f}{\kappa'^2_{eg}\rho_k^2 \min\{\delta_k^2, \delta_k^4\}\left(1 - \beta_m^{1/(p+1)}\right)} = \frac{16V_f}{c_1^2 L_g^2 Q_{\max}^4 \rho_k^2 \min(\delta_k^2, \delta_k^4)\left(1 - \beta_m^{1/(p+1)}\right)}.$$

□

The next result shows that for a particular geometry of the interpolation set $\mathbb{S}^k$ of Theorem 3.3, the resulting model gradient is a forward finite-difference stochastic gradient estimator of $\boldsymbol{Q}_{\boldsymbol{k}}^\top \nabla f(\boldsymbol{x_k})$ and its corresponding parameters κ_{ef} and κ_{eg} do not depend on k, as mentioned above.

Corollary 3.3. *Under all the assumptions of Theorem 3.3, assume further that $\boldsymbol{s_i} = h\boldsymbol{e_i} \in \mathbb{R}^p, i \in [\![1, p]\!]$, where $\boldsymbol{e_i}$ is the ith standard basis vector of $\mathbb{R}^p$ and $h = \min\{h_{\text{opt}}, \delta_k\}$ for some $h_{\text{opt}} > 0$ sufficiently small. Let $\underline{f}_k^0(\boldsymbol{x_k}) := \underline{f}_k^s(0, \underline{\theta}^0)$ and $\underline{f}_k^s(\boldsymbol{x_k} + h(\boldsymbol{Q_k})_{:i}) := \underline{f}_k^s(\boldsymbol{s_i}, \underline{\theta}^i)$ for all $i \in [\![1, p]\!]$, where $(\boldsymbol{Q_k})_{:i}$ denotes the ith column of $\boldsymbol{Q_k}$. Then $\boldsymbol{Q_k}^\top \nabla f(\boldsymbol{x_k}) \approx \hat{\boldsymbol{g}}_{\boldsymbol{k}} = \hat{\underline{\boldsymbol{g}}}_{\boldsymbol{k}}(\omega)$, where $\hat{\underline{\boldsymbol{g}}}_{\boldsymbol{k}}$ is defined by the forward finite-difference scheme $(\hat{\underline{\boldsymbol{g}}}_{\boldsymbol{k}})_i = \frac{\underline{f}_k^s(\boldsymbol{x_k} + h(\boldsymbol{Q_k})_{:i}) - \underline{f}_k^0(\boldsymbol{x_k})}{h}$, $i \in [\![1, p]\!]$. Moreover, the corresponding model parameters κ_{ef} and κ_{eg} given by (11) are independent of k.*

Proof. Note that $\mathbb{S}^k = \{\boldsymbol{0}, h\boldsymbol{e_1}, h\boldsymbol{e_2}, \ldots, h\boldsymbol{e_p}\} \subseteq \mathbb{R}^p \cap \mathcal{B}(0; c_1\delta_k)$, $\rho_k = h$, and hence $\hat{\boldsymbol{L}}_{\boldsymbol{k}} = \boldsymbol{I_p}$. The proof immediately follows from (15) by replacing $\underline{\boldsymbol{a}}$ with $\hat{\underline{\boldsymbol{g}}}_{\boldsymbol{k}}$. The nondependence of κ_{ef} and κ_{eg} on k trivially follows from the fact that $\hat{\boldsymbol{L}}_{\boldsymbol{k}}^{-1} = \boldsymbol{I_p}$. □

4 Convergence analysis.

This section presents convergence results of Algorithm 1 using ideas inspired by [2, 8, 13, 16, 17]. Section 4.1 presents preliminary results necessary for the proof of the main results. Section 4.2 proves that the sequence of random trust-region radii converges to zero almost surely. Section 4.3 demonstrates the existence of a subsequence of random iterates generated by the proposed method, which drives the norm of ∇f to zero almost surely. Proofs that are not presented in this section can be found in the Appendix.

4.1 Preliminary results.

In the remainder of the manuscript, the following inspired by [13, Assumptions 4.1 and 4.3], respectively, will be assumed.

Assumption 4. *For given $\boldsymbol{x_0}$, $\delta_{\max}$, and $\boldsymbol{Q_0}$, let $\mathscr{L}(\boldsymbol{x_0};\boldsymbol{Q_0}) \subset \mathbb{R}^n$ be the set containing all iterates of Algorithm 1. Define the region $\mathscr{L}^\star(\boldsymbol{x_0};\boldsymbol{Q_0})$ considered by the algorithm realizations as $\mathscr{L}^\star(\boldsymbol{x_0};\boldsymbol{Q_0}) = \bigcup_{\{k\geq 0,\, \boldsymbol{x_k}\in\mathscr{L}(\boldsymbol{x_0};\boldsymbol{Q_0})\}} \mathcal{B}(\boldsymbol{x_k},\delta_{\max};\boldsymbol{Q_k})$. Then $f(\boldsymbol{x}) \geq f_{\min}$ for all $\boldsymbol{x}\in\mathbb{R}^n$ and for some constant $f_{\min} > -\infty$. Moreover, f and its gradient are Lipschitz continuous on $\mathscr{L}^\star(\boldsymbol{x_0};\boldsymbol{Q_0})$.*

Assumption 5. *There exists some $\kappa_h \geq 1$ such that for all $k\geq 0$, the Hessian $\hat{\boldsymbol{H}}_k$ of all realizations of $\hat{m}_k$ satisfies $\left\|\hat{\boldsymbol{H}}_k\right\| \leq \kappa_h$.*

The following subspace variant of [13, Lemma 4.5] shows that if δ_k is small enough compared with the size $\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2$ of a $(\kappa_{ef},\kappa_{eg};\boldsymbol{Q_k})$-fully linear model gradient, then the trial step $\boldsymbol{s_k}$ provides a decrease in f proportional to $\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2$.

Lemma 4.1. *Suppose that the function $\hat{m}_k$ is a $(\kappa_{ef},\kappa_{eg};\boldsymbol{Q_k})$-fully linear model for f in $\mathcal{B}(\boldsymbol{x_k};\delta_k;\boldsymbol{Q_k})$. Then the trial step $\boldsymbol{s_k}$ leads to an improvement in f such that*

$$f(\boldsymbol{x_k}+\boldsymbol{Q_k}\boldsymbol{s_k}) - f(\boldsymbol{x_k}) \leq -\frac{\kappa_{fcd}}{4}\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2\delta_k, \quad \text{if } \delta_k \leq \min\left\{\frac{1}{\kappa_h}, \frac{\kappa_{fcd}}{8\kappa_{ef}}\right\}\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2. \tag{20}$$

The following subspace variant of [13, Lemma 4.6] shows that the guaranteed decrease in f provided by $\boldsymbol{s_k}$ is proportional to $\left\|\boldsymbol{Q_k}^\top\nabla f(\boldsymbol{x_k})\right\|_2$ if δ_k is small enough compared with $\left\|\boldsymbol{Q_k}^\top\nabla f(\boldsymbol{x_k})\right\|_2$.

Lemma 4.2. *Let Assumption 5 hold, and suppose the model $\hat{m}_k$ is $(\kappa_{ef},\kappa_{eg};\boldsymbol{Q_k})$-fully linear. Define $C_1 = \frac{\kappa_{fcd}}{4}\max\left\{\frac{\kappa_h}{\kappa_h+\kappa_{eg}}, \frac{8\kappa_{ef}}{8\kappa_{ef}+\kappa_{fcd}\kappa_{eg}}\right\}$. Then the trial step $\boldsymbol{s_k}$ leads to an improvement in f such that*

$$f(\boldsymbol{x_k}+\boldsymbol{Q_k}\boldsymbol{s_k}) - f(\boldsymbol{x_k}) \leq -C_1\left\|\boldsymbol{Q_k}^\top\nabla f(\boldsymbol{x_k})\right\|_2\delta_k, \tag{21}$$

$$\text{if} \quad \delta_k \leq \min\left\{\frac{1}{\kappa_h+\kappa_{eg}}, \frac{1}{\frac{8\kappa_{ef}}{\kappa_{fcd}}+\kappa_{eg}}\right\}\left\|\boldsymbol{Q_k}^\top\nabla f(\boldsymbol{x_k})\right\|_2. \tag{22}$$

The next subspace variant of [13, Lemma 4.7] shows that if the model and the estimates are sufficiently accurate and δ_k is small enough compared with $\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2$, then the iteration is successful.

Lemma 4.3. *Let Assumption 5 hold. Assume that $\hat{m}_k$ is $(\kappa_{ef},\kappa_{eg};\boldsymbol{Q_k})$-fully linear and the estimates $\{f_k^0, f_k^s\}$ are ε_f-accurate with $\varepsilon_f \leq \kappa_{ef}$. Then the kth iteration is successful if*

$$\delta_k \leq \min\left\{\frac{1}{\kappa_h}, \frac{1}{\eta_2}, \frac{\kappa_{fcd}(1-\eta_1)}{8\kappa_{ef}}\right\}\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2. \tag{23}$$

Proof. The proof immediately follows from that of [13, Lemma 4.7] with minor modifications and is not presented here again. □

The next result, which is a subspace variant of [13, Lemma 4.8], guarantees an amount of decrease in f on true successful iterations.

Lemma 4.4. *Suppose that Assumption 5 holds and the estimates* $\{f_k^0, f_k^s\}$ *are* ε_f*-accurate with* $\varepsilon_f < \frac{1}{4}\eta_1\eta_2\kappa_{fcd}\min\left\{\frac{\eta_2}{\kappa_h}, 1\right\}$*. If the* k*th iteration is successful, then the improvement in* f *is such that*

$$f(\boldsymbol{x_{k+1}}) - f(\boldsymbol{x_k}) \leq -C_2\delta_k^2, \tag{24}$$

where $C_2 = \frac{1}{2}\eta_1\eta_2\kappa_{fcd}\min\left\{\frac{\eta_2}{\kappa_h}, 1\right\} - 2\varepsilon_f > 0$.

Proof. Again, the proof immediately follows from that of [13, Lemma 4.8] with minor modifications and is not presented here. □

Inspired by a result from the proof of [13, Theorem 4.11], the next lemma quantifies the maximum possible amount of increase in f on false successful iterations.

Lemma 4.5. *Under Assumptions 3 and 4, assume that Algorithm 1 erroneously accepts a step* $\boldsymbol{s_k}$ *leading to an increase in the objective function when* $\left\|\boldsymbol{Q_k}^\top \nabla f(\boldsymbol{x_k})\right\| \geq \zeta\delta_k$ *for some constant* $\zeta > 0$*, and let* $C_3(\zeta) = 1 + \frac{3L_g}{2\zeta}Q_{\max}^2$*. Then the increase in* f *is such that*

$$f(\boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s_k}) - f(\boldsymbol{x_k}) \leq C_3(\zeta)\left\|\boldsymbol{Q_k}^\top \nabla f(\boldsymbol{x_k})\right\|\delta_k. \tag{25}$$

Later, to prove the key result of Theorem 4.1, we favor techniques derived in [28], unlike [9, 13]. These techniques, also used in [2, 16, 17], make use of event indicator functions. Here, by means of Lemma 4.7, we introduce a measurability result that is related to one of these indicator functions and is crucial in the present analysis. But first, the following result is required.

Lemma 4.6. *Any partial derivative* $\frac{\partial f}{\partial x^i} : (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, $i \in [\![1, n]\!]$*, of a differentiable function* f *is a Borel measurable function.*

Proof. Let $(\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_n})$ be the canonical basis of $\mathbb{R}^n$. For any $i \in [\![1, n]\!]$ and $\boldsymbol{x} = (x^1, x^2, \ldots, x^n)$, define, for all $j \in \mathbb{N}$, $D_{ij}(\boldsymbol{x}) = \frac{f(\boldsymbol{x}+h_j\boldsymbol{e_i}) - f(\boldsymbol{x})}{h_j}$, where $\{h_j\}_{j\in\mathbb{N}}$ is a sequence of positive real numbers converging to 0; f is differentiable and hence continuous, and thus is Borel measurable [8, Theorem 13.2]. Therefore, since for all $i \in [\![1, n]\!]$, $\lim_{j\to\infty} D_{ij}(\boldsymbol{x}) = \frac{\partial f}{\partial x^i}(\boldsymbol{x})$, then $\frac{\partial f}{\partial x^i}$ is also Borel measurable [8, Theorem 13.3 and Theorem 13.4-(ii)]. □

Lemma 4.7. *Let* $\zeta \geq 0$ *be a constant, and consider the event* $\Gamma_k := \{\underline{g}_k \geq 0\}$*, where* $\underline{g}_k := \left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top \nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\right\|_2 - \zeta\underline{\delta}_k$*. Then* $\underline{g}_k$*,* $\mathbb{1}_{\Gamma_k}$*, and* $\mathbb{1}_{\bar{\Gamma}_k}$ *are* $\mathcal{F}_{k-1}^Q$*-measurable.*

Proof. By construction, all the entries $\left(\underline{\boldsymbol{Q}}_{\boldsymbol{k}}\right)_{ij}$, $(i, j) \in [\![1, n]\!] \times [\![1, p]\!]$ of $\underline{\boldsymbol{Q}}_{\boldsymbol{k}}$ are $\mathcal{F}_{k-1}^Q$-measurable, while $\underline{\delta}_k$ and the components $\underline{\boldsymbol{x}}_{\boldsymbol{k}}^i$, $i \in [\![1, n]\!]$ of $\underline{\boldsymbol{x}}_{\boldsymbol{k}}$ are $\mathcal{F}_{k-1}$-measurable and hence $\mathcal{F}_{k-1}^Q$-measurable since $\mathcal{F}_{k-1} \subseteq \mathcal{F}_{k-1}^Q$. Thus, for any $i \in [\![1, n]\!]$, $\frac{\partial f}{\partial x^i}(\underline{\boldsymbol{x}}_{\boldsymbol{k}})$ is $\mathcal{F}_{k-1}^Q$-measurable as the composition of the Borel measurable function $\frac{\partial f}{\partial x^i}$ (from Lemma 4.6) and the random map $\underline{\boldsymbol{x}}_{\boldsymbol{k}} : (\Omega, \mathcal{F}_{k-1}^Q) \to (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ [8, Theorem 13.3 and Theorem 13.1-(ii)]. Recall that sums and products of measurable functions are also measurable [8, Theorem 13.3], and note that the random variable

$$\underline{u}_k := \left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top \nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\right\|_2^2 = \sum_{j=1}^{p}\left(\sum_{i=1}^{n}\left(\underline{\boldsymbol{Q}}_{\boldsymbol{k}}\right)_{ij}\frac{\partial f}{\partial x^i}(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\right)^2 \tag{26}$$

is consequently $\mathcal{F}_{k-1}^Q$-measurable. Thus $\underline{g}_k = \sqrt{\underline{u}_k} - \zeta\underline{\delta}_k$ is also $\mathcal{F}_{k-1}^Q$-measurable. The $\mathcal{F}_{k-1}^Q$-measurability of $\underline{g}_k$ ensures in particular that $\Gamma_k = \underline{g}_k^{-1}([0, \infty)) \in \mathcal{F}_{k-1}^Q$ since $[0, \infty)$ is a Borel set of $\mathbb{R}$, whence the *simple* real functions $\mathbb{1}_{\Gamma_k}$ and $\mathbb{1}_{\bar{\Gamma}_k}$ [8, Equation (13.3)] are also $\mathcal{F}_{k-1}^Q$-measurable. □

4.2 Zeroth-order convergence result.

To prove in Theorem 4.1 that the sequence of random trust-region radii converges to zero almost surely, we assume the following.

Assumption 6. *The sequences of estimates* $\{\underline{f}_k^0, \underline{f}_k^s\}$ *and models* $\{\hat{\underline{m}}_k\}$ *generated by Algorithm 1 are, respectively,* β_f*-probabilistically* ε_f*-accurate for* $\varepsilon_f < \min\left\{\kappa_{ef}, \frac{1}{4}\eta_1\eta_2\kappa_{fcd}\min\left\{\frac{\eta_2}{\kappa_h}, 1\right\}\right\}$ *and* β_m*-probabilistically* $(\kappa_{ef}, \kappa_{eg}; \boldsymbol{Q_k})$*-fully linear for some* $\beta_f, \beta_m \in (0,1)$.

Recall Definitions 3.1 and 3.2. The result presented next shows that the sequence of trust-region radii converges to zero whether the sequence of subspace selection matrices is well aligned or not. The corresponding proof improves on the analyses presented in [9, 13], for example, by reducing the number of subcases of **Case 1**, discussed below, from four to three. This improvement is the outcome of a simple motivation to circumvent the derivation of a lower bound on the probability of the event that "either the model is good and the estimates are bad, or the model is bad and the estimates are good," as was done in [9, 13]. Another remarkable difference is the fact that both subcases do not straightforwardly use the same σ-algebras for conditioning on the past, as is the case in [9, 13], which results from the need to introduce a random variable $\mathbb{1}_{G_k}$ related to the model gradient in **Case 2**. We also note that unlike prior similar works, the remainder of the present analysis explicitly emphasizes the way all the σ-algebras $\mathcal{F}_{k-1}$, $\mathcal{F}_{k-1}^Q$, and $\mathcal{F}_{k-1}^{\hat{m}\cdot Q}$ work together for the proofs of the proposed results.

Theorem 4.1. *Let all assumptions that were made in Lemmas 4.1-4.4 hold with the same constants* C_1, C_2*, and* $C_3 := C_3(\zeta)$*, for some fixed* $\zeta > 0$ *satisfying*

$$\zeta \geq \kappa_{eg} + \max\left\{\eta_2, \kappa_h, \frac{8\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)}\right\}. \tag{27}$$

Let $\nu \in (0,1)$ *be chosen according to*

$$\frac{\nu}{1-\nu} \geq \frac{8\gamma^2}{\min\{C_2, \zeta C_1\}}. \tag{28}$$

Assume further that Assumption 6 holds with $\beta_f, \beta_m \in (0,1)$ *satisfying*[5]

$$\frac{\beta_f\beta_m - \frac{1}{2}}{(1-\beta_f)+(1-\beta_m)} \geq \frac{C_3}{C_1} \quad \text{and} \quad \frac{\beta_f}{1-\beta_f} \geq \frac{2\nu\zeta\gamma^2 C_3}{(1-\nu)(\gamma^2-1)} + 2\gamma^2. \tag{29}$$

Define $\varrho = \frac{1}{2}\beta_f(1-\nu)(1-\frac{1}{\gamma^2}) > 0$*. Then, the random function* $\underline{\phi}_k := \nu(f(\underline{\boldsymbol{x}}_{\boldsymbol{k}}) - f_{\min}) + (1-\nu)\underline{\delta}_k^2$ *satisfies*

$$\mathbb{E}\left[\underline{\phi}_{k+1} - \underline{\phi}_k | \mathcal{F}_{k-1}^Q\right] \leq -\varrho\underline{\delta}_k^2 \quad \text{for all } k \in \mathbb{N}, \tag{30}$$

which implies that

$$\sum_{k=0}^{\infty} \underline{\delta}_k^2 < \infty \quad \text{almost surely.} \tag{31}$$

Proof. The proof is inspired by those of [9, Theorem 3] and [13, Theorem 4.11]. The overall goal is to prove (30). Indeed noticing that $\underline{\phi}_k > 0$ and then taking expectations on both sides of the inequality in (30), we obtain (31) (see, e.g., [16, Theorem 3] for details). Let ϕ_k denote realizations of $\underline{\phi}_k$, and recall that on all successful iterations, $\boldsymbol{x_{k+1}} = \boldsymbol{x_k} + \boldsymbol{Q_k s_k}$ and $\delta_{k+1} = \min\{\gamma\delta_k, \delta_{\max}\}$, implying

$$\phi_{k+1} - \phi_k \leq \nu\left(f(\boldsymbol{x_{k+1}}) - f(\boldsymbol{x_k})\right) + (1-\nu)(\gamma^2-1)\delta_k^2, \tag{32}$$

while on unsuccessful iterations, $\boldsymbol{x_{k+1}} = \boldsymbol{x_k}$ and $\delta_{k+1} = \gamma^{-1}\delta_k$, in which case

$$\phi_{k+1} - \phi_k \leq (1-\nu)\left(\frac{1}{\gamma^2} - 1\right)\delta_k^2 =: b_1 < 0. \tag{33}$$

[5] In (29), the result is intentionally presented with $(1-\beta_f)+(1-\beta_m)$ instead of $2-\beta_f-\beta_m$ in order to emphasize one of its differences compared with [9, 13] in which a term similar to $(1-\beta_f)(1-\beta_m)$ was derived.

Recall the event $\Gamma_k = \left\{ \left\| \boldsymbol{Q}_{\boldsymbol{k}}^\top \nabla f(\boldsymbol{x_k}) \right\|_2 \geq \zeta \underline{\delta}_k \right\}$ of Lemma 4.7 with ζ satisfying (27). The proof considers two cases: $\mathbb{1}_{\Gamma_k} = 1$ and $\mathbb{1}_{\bar{\Gamma}_k} = 1$. Inspired by [13, proof of Theorem 4.11], Case 1 aims to show that

$$\mathbb{E}\left[\mathbb{1}_{\Gamma_k} \left(\underline{\phi}_{k+1} - \underline{\phi}_k \right) | \mathcal{F}_{k-1}^Q \right] \leq -2\mathbb{1}_{\Gamma_k}(1-\nu)(\gamma^2-1)\underline{\delta}_k^2 \leq -\frac{1}{2}\mathbb{1}_{\Gamma_k}\beta_f(1-\nu)(1-\frac{1}{\gamma^2})\underline{\delta}_k^2, \tag{34}$$

where the last inequality follows from $1 - \frac{1}{\gamma^2} < \gamma^2 - 1$ and $\beta_f \leq 1$. On the other hand, inspired by [9, proof of Theorem 3], Case 2 shows that

$$\mathbb{E}\left[\mathbb{1}_{\bar{\Gamma}_k} \left(\underline{\phi}_{k+1} - \underline{\phi}_k \right) | \mathcal{F}_{k-1}^Q \right] \leq -\frac{1}{2}\mathbb{1}_{\bar{\Gamma}_k}\beta_f(1-\nu)(1-\frac{1}{\gamma^2})\underline{\delta}_k^2. \tag{35}$$

Thus, combining (34) and (35) leads to (30).

Case 1: $\mathbb{1}_{\Gamma_k} = 1$

Unlike the proof of [13, Theorem 4.11, Case 1] where four subcases were considered, only three are analyzed next.

(Subcase 1i) Good model ($\mathbb{1}_{I_k^Q} = 1$) and good estimates ($\mathbb{1}_{J_k^Q} = 1$). By (27),

$$\left\| \boldsymbol{Q}_{\boldsymbol{k}}^\top \nabla f(\boldsymbol{x_k}) \right\|_2 \geq \max\left\{ \eta_2 + \kappa_{eg}, \kappa_h + \kappa_{eg}, \frac{8\kappa_{ef}}{\kappa_{fcd}} + \kappa_{eg} \right\} \delta_k,$$

implying in particular (22) and hence a decrease in f according to (21). Moreover, by $(\kappa_{ef}, \kappa_{eg}; \boldsymbol{Q_k})$-full linearity,

$$\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \geq \left\| \boldsymbol{Q}_{\boldsymbol{k}}^\top \nabla f(\boldsymbol{x_k}) \right\|_2 - \kappa_{eg}\delta_k \geq \max\left\{ \eta_2, \kappa_h, \frac{8\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)} \right\} \delta_k,$$

thus implying (23); and since $\varepsilon_f \leq \kappa_{ef}$ per assumptions of Lemma 4.3, the iteration is successful. Consequently, (32) together with (21) yields

$$\begin{aligned} \phi_{k+1} - \phi_k &\leq -\nu C_1 \left\| \boldsymbol{Q}_{\boldsymbol{k}}^\top \nabla f(\boldsymbol{x_k}) \right\|_2 \delta_k + (1-\nu)(\gamma^2-1)\delta_k^2 =: b_2 \\ &\leq \left(-\nu C_1 \zeta + (1-\nu)(\gamma^2-1) \right) \delta_k^2 < 0, \end{aligned} \tag{36}$$

where the last inequality follows from (28). Denoting by $\underline{b}_2(\underline{u}_k, \underline{\delta}_k)$ the random variable with realizations $b_2 < 0$, with $\underline{u}_k$ defined in (26), we have from (36) that

$$\mathbb{1}_{\Gamma_k}\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q}\left(\underline{\phi}_{k+1} - \underline{\phi}_k \right) \leq \mathbb{1}_{\Gamma_k}\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q}\underline{b}_2(\underline{u}_k, \underline{\delta}_k). \tag{37}$$

(Subcase 1ii) Bad model ($\mathbb{1}_{\bar{I}_k^Q} = 1$) and good estimates ($\mathbb{1}_{J_k^Q} = 1$). In this case, regardless of the iteration type (i.e., successful or unsuccessful), the change in ϕ_k can always be bounded by using (32) and (25) as follows:

$$\phi_{k+1} - \phi_k \leq \nu C_3 \left\| \boldsymbol{Q}_{\boldsymbol{k}}^\top \nabla f(\boldsymbol{x_k}) \right\| \delta_k + (1-\nu)(\gamma^2-1)\delta_k^2 =: b_3. \tag{38}$$

Denoting by $\underline{b}_3(\underline{u}_k, \underline{\delta}_k)$ the random variable with realizations $b_3 > 0$, from (38) we have that

$$\mathbb{1}_{\Gamma_k}\mathbb{1}_{\bar{I}_k^Q}\mathbb{1}_{J_k^Q}\left(\underline{\phi}_{k+1} - \underline{\phi}_k \right) \leq \mathbb{1}_{\Gamma_k}\mathbb{1}_{\bar{I}_k^Q}\mathbb{1}_{J_k^Q}\underline{b}_3(\underline{u}_k, \underline{\delta}_k) \leq \mathbb{1}_{\Gamma_k}\mathbb{1}_{\bar{I}_k^Q}\underline{b}_3(\underline{u}_k, \underline{\delta}_k). \tag{39}$$

(Subcase 1iii) Bad estimates ($\mathbb{1}_{\bar{J}_k^Q} = 1$). From *Subcase 1ii*, it always holds that

$$\mathbb{1}_{\Gamma_k}\mathbb{1}_{\bar{J}_k^Q}\left(\underline{\phi}_{k+1} - \underline{\phi}_k \right) \leq \mathbb{1}_{\Gamma_k}\mathbb{1}_{\bar{J}_k^Q}\underline{b}_3(\underline{u}_k, \underline{\delta}_k). \tag{40}$$

With the subcases thus complete, since the random variables $\mathbb{1}_{\Gamma_k}, \underline{b}_2(\underline{u}_k, \underline{\delta}_k)$, and $\underline{b}_3(\underline{u}_k, \underline{\delta}_k)$ are $\mathcal{F}_{k-1}^Q$-measurable thanks to the proof of Lemma 4.7, combining (37), (39), and (40) and taking expectations with

respect to $\mathcal{F}_{k-1}^Q$, we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\Gamma_k}\left(\underline{\phi}_{k+1}-\underline{\phi}_k\right)|\mathcal{F}_{k-1}^Q\right] &\leq \mathbb{1}_{\Gamma_k}\underline{b}_2(\underline{u}_k,\underline{\delta}_k)\,\mathbb{E}\left[\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q}|\mathcal{F}_{k-1}^Q\right]\\ &\quad+\mathbb{1}_{\Gamma_k}\underline{b}_3(\underline{u}_k,\underline{\delta}_k)\left(\mathbb{E}\left[\mathbb{1}_{\bar{I}_k^Q}|\mathcal{F}_{k-1}^Q\right]+\mathbb{E}\left[\mathbb{1}_{\bar{J}_k^Q}|\mathcal{F}_{k-1}^Q\right]\right)\\ &\leq \mathbb{1}_{\Gamma_k}\left(\beta_f\beta_m\underline{b}_2(\underline{u}_k,\underline{\delta}_k)+\underline{b}_3(\underline{u}_k,\underline{\delta}_k)(1-\beta_m+1-\beta_f)\right)=:\mathbb{1}_{\Gamma_k}\underline{b}_{2,3}(\underline{u}_k,\underline{\delta}_k),\end{aligned}\tag{41}$$

where the last inequality follows from (5) and (9) and the fact that $\underline{b}_2(\underline{u}_k,\underline{\delta}_k)<0$ while $\underline{b}_3(\underline{u}_k,\underline{\delta}_k)>0$. Note that

$$\begin{aligned}\underline{b}_{2,3}(\underline{u}_k,\underline{\delta}_k) &= \nu\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top\nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\right\|_2\underline{\delta}_k\left[-\beta_f\beta_mC_1+(2-\beta_f-\beta_m)C_3\right]\\ &\quad+(1-\nu)(\gamma^2-1)(2-\beta_f-\beta_m+\beta_f\beta_m)\underline{\delta}_k^2\\ &\leq \nu\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top\nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\right\|_2\underline{\delta}_k\left[-\beta_f\beta_mC_1+(2-\beta_f-\beta_m)C_3\right]+2(1-\nu)(\gamma^2-1)\underline{\delta}_k^2,\end{aligned}$$

where the last inequality follows from $2-\beta_f-\beta_m+\beta_f\beta_m=1+(1-\beta_f)(1-\beta_m)\leq 2$.

For β_f and β_m chosen according to the first condition in (29), and for ν satisfying (28), the following holds:

$$\beta_f\beta_mC_1-(2-\beta_f-\beta_m)C_3\geq\frac{1}{2}C_1\geq 2\frac{2(1-\nu)(\gamma^2-1)}{\nu\zeta}.$$

Hence, since $\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top\nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\right\|_2\geq\zeta\underline{\delta}_k$, then

$$\begin{aligned}\underline{b}_{2,3}(\underline{u}_k,\underline{\delta}_k) &\leq -\frac{1}{2}\left[\beta_f\beta_mC_1-(2-\beta_f-\beta_m)C_3\right]\nu\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top\nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\right\|_2\underline{\delta}_k\\ &\leq -\frac{1}{4}C_1\nu\left\|\underline{\boldsymbol{Q}}_{\boldsymbol{k}}^\top\nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\right\|_2\underline{\delta}_k\leq-2(1-\nu)(\gamma^2-1)\underline{\delta}_k^2,\end{aligned}$$

which, together with (41), leads to (34).

Case 2: $\mathbb{1}_{\bar{\Gamma}_k}=1$

Consider the event $G_k:=\left\{\left\|\underline{\hat{\boldsymbol{g}}}_{\boldsymbol{k}}\right\|_2\geq\eta_2\underline{\delta}_k\right\}$, and note that since $\hat{m}_k:=\underline{f}_k+\underline{\hat{\boldsymbol{g}}}_{\boldsymbol{k}}^\top\underline{\boldsymbol{s}}+\frac{1}{2}\underline{\boldsymbol{s}}^\top\underline{\hat{H}}_k\underline{\boldsymbol{s}}$ is $\mathcal{F}_{k-1}^{\hat{m}\cdot Q}$-measurable by construction, then in particular $\left\|\underline{\hat{\boldsymbol{g}}}_{\boldsymbol{k}}\right\|_2$ and hence $\mathbb{1}_{G_k},\mathbb{1}_{\bar{G}_k}$ are also $\mathcal{F}_{k-1}^{\hat{m}\cdot Q}$-measurable. Observe that if $\mathbb{1}_{\bar{G}_k}=1$, then the iteration is unsuccessful since $\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2<\eta_2\delta_k$, leading to (33). Hence,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\bar{\Gamma}_k}\mathbb{1}_{\bar{G}_k}(\underline{\phi}_{k+1}-\underline{\phi}_k)|\mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right] &\leq -(1-\nu)\left(1-\frac{1}{\gamma^2}\right)\mathbb{E}\left[\mathbb{1}_{\bar{\Gamma}_k}\mathbb{1}_{\bar{G}_k}\underline{\delta}_k^2|\mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right]\\ &\leq -\frac{1}{2}\mathbb{1}_{\bar{\Gamma}_k}\mathbb{1}_{\bar{G}_k}\beta_f(1-\nu)\left(1-\frac{1}{\gamma^2}\right)\underline{\delta}_k^2,\end{aligned}\tag{42}$$

where the last inequality is due to the $\mathcal{F}_{k-1}^{\hat{m}\cdot Q}$-measurability of $\mathbb{1}_{\bar{\Gamma}_k}$ and $\underline{\delta}_k$, given that $\mathcal{F}_{k-1}\subset\mathcal{F}_{k-1}^Q\subset\mathcal{F}_{k-1}^{\hat{m}\cdot Q}$.

Inspired by [9, Theorem 3], which improved on the proof of [13, Theorem 4.11, Case 2] in which four subcases were considered, only two are analyzed next. The overall goal is to prove that

$$\mathbb{E}\left[\mathbb{1}_{\bar{\Gamma}_k}\mathbb{1}_{G_k}(\underline{\phi}_{k+1}-\underline{\phi}_k)|\mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right]\leq-\frac{1}{2}\mathbb{1}_{\bar{\Gamma}_k}\mathbb{1}_{G_k}\beta_f(1-\nu)\left(1-\frac{1}{\gamma^2}\right)\underline{\delta}_k^2,\tag{43}$$

which, combined with (42), leads to

$$\mathbb{E}\left[\mathbb{1}_{\bar{\Gamma}_k}(\underline{\phi}_{k+1}-\underline{\phi}_k)|\mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right]\leq-\frac{1}{2}\mathbb{1}_{\bar{\Gamma}_k}\beta_f(1-\nu)\left(1-\frac{1}{\gamma^2}\right)\underline{\delta}_k^2$$

and consequently

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\bar{\Gamma}_k}(\underline{\phi}_{k+1}-\underline{\phi}_k)|\mathcal{F}_{k-1}^Q\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\bar{\Gamma}_k}(\underline{\phi}_{k+1}-\underline{\phi}_k)|\mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right]|\mathcal{F}_{k-1}^Q\right]\\ &\leq \mathbb{E}\left[-\frac{1}{2}\mathbb{1}_{\bar{\Gamma}_k}\beta_f(1-\nu)\left(1-\frac{1}{\gamma^2}\right)\underline{\delta}_k^2|\mathcal{F}_{k-1}^Q\right]=-\frac{1}{2}\mathbb{1}_{\bar{\Gamma}_k}\beta_f(1-\nu)\left(1-\frac{1}{\gamma^2}\right)\underline{\delta}_k^2,\end{aligned}$$

showing (35), where the last equality follows from the $\mathcal{F}_{k-1}^{Q}$-measurability of $\mathbb{1}_{\bar{\Gamma}_k}$ and $\underline{\delta}_k$. What remains to be shown is (43) as achieved next, assuming that $\mathbb{1}_{G_k} = 1$.
(Subcase 2i) Good estimates ($\mathbb{1}_{J_k^Q} = 1$) and $\mathbb{1}_{G_k} = 1$. While f is decreased on successful iterations because of good estimates thanks to Lemma 4.4 and in this case (24) holds, δ_k is reduced on unsuccessful iterations, leading to (33). However, combining (24) and (32) yields $\phi_{k+1} - \phi_k \leq \left[-\nu C_2 + (1-\nu)(\gamma^2 - 1)\right] \delta_k^2 \leq b_1$, where the last inequality is due to (28), which shows that (33) holds in any case. Consequently,

$$\mathbb{1}_{\bar{\Gamma}_k} \mathbb{1}_{G_k} \mathbb{1}_{J_k^Q} (\underline{\phi}_{k+1} - \underline{\phi}_k) \leq -\mathbb{1}_{\bar{\Gamma}_k} \mathbb{1}_{G_k} \mathbb{1}_{J_k^Q} (1-\nu) \left(1 - \frac{1}{\gamma^2}\right) \underline{\delta}_k^2. \tag{44}$$

(Subcase 2ii) Bad estimates ($\mathbb{1}_{\bar{J}_k^Q} = 1$) and $\mathbb{1}_{G_k} = 1$. In this case, a successful step can lead to an increase in f according to (25), which combined with (32) yield (38) and hence

$$\mathbb{1}_{\bar{\Gamma}_k} \mathbb{1}_{G_k} \mathbb{1}_{\bar{J}_k^Q} (\underline{\phi}_{k+1} - \underline{\phi}_k) \leq -\mathbb{1}_{\bar{\Gamma}_k} \mathbb{1}_{G_k} \mathbb{1}_{\bar{J}_k^Q} \left(\nu C_3 \zeta + (1-\nu)(\gamma^2 - 1)\right) \underline{\delta}_k^2. \tag{45}$$

With the subcases complete, recall that $\mathbb{E}\left[\mathbb{1}_{J_k^Q} | \mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right] \geq \beta_f$. Then combining (44) and (45) and taking expectations with respect to $\mathcal{F}_{k-1}^{\hat{m}\cdot Q}$ lead to

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\bar{\Gamma}_k} \mathbb{1}_{G_k} (\underline{\phi}_{k+1} - \underline{\phi}_k) | \mathcal{F}_{k-1}^{\hat{m}\cdot Q}\right] &\leq \mathbb{1}_{\bar{\Gamma}_k} \mathbb{1}_{G_k} \left(-\beta_f (1-\nu)(1 - 1/\gamma^2)\right. \\
&\quad \left. + (1-\beta_f)\left(\nu C_3 \zeta + (1-\nu)(\gamma^2 - 1)\right)\right) \underline{\delta}_k^2 \\
&\leq -\frac{1}{2} \mathbb{1}_{\bar{\Gamma}_k} \mathbb{1}_{G_k} \beta_f (1-\nu) \left(1 - \frac{1}{\gamma^2}\right) \underline{\delta}_k^2,
\end{aligned}$$

thus proving (43), where the last inequality follows from the second condition in (29), and the proof is complete. □

4.3 Liminf-type convergence.

The next result demonstrates the existence of a subsequence of random iterates generated by Algorithm 1, which drives ∇f to zero almost surely. While the corresponding proof is inspired by that of [13, Theorem 4.16], we point out the additional difficulty introduced in the present work by the random matrices $\boldsymbol{Q}_{\boldsymbol{k}}$. Unlike [13, Section 4] where "auxiliary lemmas" similar to those of Section 4.1 are directly related to $\|\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\|_2$ and the size $\|\boldsymbol{g}_{\boldsymbol{k}}\|_2$ of a full space model gradient, thus easing the proof of the liminf-type result, this is not the case here. Instead, recovering $\|\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\|_2$ through $\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top} \nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2$ by means of the well-alignment assumption becomes crucial.

Theorem 4.2. *Let all the assumptions made in Theorem 4.1 hold. Assume further that Assumption 2 holds with $0 \leq \beta_Q < 1 - \frac{1}{2\beta_f \beta_m}$. Then, the sequence $\{\underline{\boldsymbol{x}}_{\boldsymbol{k}}\}_{k \in \mathbb{N}}$ of random iterates generated by Algorithm 1 satisfies*

$$\liminf_{k \to \infty} \|\nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\|_2 = 0 \quad \textit{almost surely.}$$

Proof. The result is proved by contradiction conditioned on the almost sure event $E_0 := \{\underline{\delta}_k \to 0\}$ (thanks to Theorem 4.1), inspired by the proof of [13, Theorem 4.16] and also using ideas from [2, 17]. Assume that with nonzero probability there exists a random variable $\underline{\varepsilon}' > 0$ such that $\|\nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\|_2 \geq \underline{\varepsilon}'$ for all $k \in \mathbb{N}$.

Let $\{\boldsymbol{x}_{\boldsymbol{k}}\}, \{\delta_k\}$, and ε' be realizations of $\{\underline{\boldsymbol{x}}_{\boldsymbol{k}}\}, \{\underline{\delta}_k\}$, and $\underline{\varepsilon}'$, respectively, for which $\|\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\|_2 \geq \varepsilon'$ for all k. Since $\delta_k \to 0$, there exists $k_0 \in \mathbb{N}$ such that

$$\delta_k < b := \alpha_Q \cdot \min\left\{ \frac{\varepsilon'}{2(\kappa_h + \kappa_{eg})}, \frac{\varepsilon'}{2(\eta_2 + \kappa_{eg})}, \frac{\varepsilon'}{\frac{16\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)} + 2\kappa_{eg}}, \frac{\delta_{\max}}{\gamma} \right\} \tag{46}$$

for all $k \geq k_0$. Let $\underline{r}_k$ be the random variable with realizations $r_k = \log_\gamma\left(\frac{\delta_k}{b}\right)$. Then $r_k < 0$ for all $k \geq k_0$. The main idea is to show that such realizations occur only with probability zero, leading to a contradiction.

To prove that $\{\underline{r}_k\}$ is a submartingale, recall the events I_k^Q and J_k^Q satisfying (10) and $\mathcal{A}_k$ satisfying (3). Consider some iteration $k \geq k_0$ for which $\mathcal{A}_k$, I_k^Q, and J_k^Q all occur, which happens with probability at least $\beta_f \beta_m (1-\beta_Q)$ conditioned on $\mathcal{F}_{k-1}$, as was shown in (7). It follows from (46) that

$$\|\nabla f(\boldsymbol{x_k})\|_2 \geq \frac{1}{\alpha_Q} \max\left\{\kappa_h + \kappa_{eg}, \eta_2 + \kappa_{eg}, \frac{8\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)} + \kappa_{eg}\right\} \delta_k, \tag{47}$$

which, together with the $(\kappa_{ef}, \kappa_{eg}; \boldsymbol{Q_k})$-full linearity and α_Q-well alignment, yield

$$\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \geq \left\|\boldsymbol{Q}_{\boldsymbol{k}}^\top \nabla f(\boldsymbol{x_k})\right\|_2 - \kappa_{eg}\delta_k \geq \alpha_Q \|\nabla f(\boldsymbol{x_k})\|_2 - \kappa_{eg}\delta_k \geq \max\left\{\kappa_h, \eta_2, \frac{8\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)}\right\} \delta_k. \tag{48}$$

Thus, the kth iteration is successful according to Lemma 4.3. Hence, $\delta_{k+1} = \gamma \delta_k$, and hence $r_{k+1} = r_k + 1$. For all other outcomes of $\mathcal{A}_k$, I_k^Q, and J_k^Q, which occur with a total probability of at most $1 - \beta_f \beta_m (1-\beta_Q)$, it always holds that $\delta_{k+1} \geq \gamma^{-1}\delta_k$, which implies that $r_{k+1} \geq r_k - 1$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_k \cap I_k^Q \cap J_k^Q} (\underline{r}_{k+1} - \underline{r}_k) \,|\mathcal{F}_{k-1}\right] &\geq \beta_f \beta_m (1-\beta_Q) \qquad \text{and} \\
\mathbb{E}\left[\mathbb{1}_{\overline{\mathcal{A}_k \cap I_k^Q \cap J_k^Q}} (\underline{r}_{k+1} - \underline{r}_k) \,|\mathcal{F}_{k-1}\right] &\geq \beta_f \beta_m (1-\beta_Q) - 1,
\end{aligned}$$

which shows that $\{\underline{r}_k\}$ is a submartingale if $0 \leq \beta_Q < 1 - \frac{1}{2\beta_f \beta_m}$.

Using the random walk defined in (8), one can easily show (following, e.g., [2, Theorem 4], [13, Theorem 4.16] and [17, Theorem 3.6 and Lemmas 4.7 and 4.12]) that $\{\underline{w}_k\}$ is also a submartingale with bounded increments and, as such, cannot converge to a finite value, and therefore $\mathbb{P}\left(\limsup_{k\to\infty} \underline{w}_k = \infty\right) = 1$ (thanks to [13, Theorem 4.4]). However, since $r_k - r_{k_0} \geq w_k - w_{k_0}$ by construction, the sequence of realizations r_k such that $r_k < 0 \ \forall k \ \geq k_0$ occurs with probability zero, leading to a contradiction, which achieves the proof. □

5 Expected complexity analysis.

Section 5.1 introduces relevant assumptions, definitions, and theorems derived in the analysis of a general renewal-reward stochastic process and its associated stopping time introduced in [9] for the expected complexity analysis of a stochastic trust-region method. That renewal-reward process was also used in [16, 28] for the analyses of stochastic direct-search and line-search methods. We show in Section 5.2 how the aforementioned assumptions are satisfied for Algorithm 1 and we bound the expected number of iterations required to achieve $\|\nabla f(\underline{\boldsymbol{x}}_{\boldsymbol{k}})\| \leq \varepsilon$.

5.1 A renewal-reward martingale process.

A formal definition from [9] of the stopping time related to a discrete time stochastic process is as follows.

Definition 5.1. *A random variable $\underline{\tau}_\epsilon$ is called a stopping time with respect to a given discrete time stochastic process $\{\underline{x}_k\}_{k\in\mathbb{N}}$ if the event $\{\underline{\tau}_\epsilon = k\}$ belongs to the σ-algebra $\sigma(\underline{x}_1, \ldots, \underline{x}_k)$ generated by $\underline{x}_1, \ldots, \underline{x}_k$, for each $k \in \mathbb{N}$.*

Consider a stochastic process $\left\{\left(\underline{\phi}_k, \underline{\delta}_k\right)\right\}_{k\in\mathbb{N}}$, where $\underline{\phi}_k, \underline{\delta}_k \in [0, \infty)$, and introduce on the same probability space as $\left\{\left(\underline{\phi}_k, \underline{\delta}_k\right)\right\}_{k\in\mathbb{N}}$ a biased random walk process $\{\underline{\tilde{w}}_k\}_{k\in\mathbb{N}}$ obeying the following dynamics:

$$\mathbb{P}\left(\underline{\tilde{w}}_{k+1} = 1 | \mathcal{F}_k\right) = q \qquad \text{and} \qquad \mathbb{P}\left(\underline{\tilde{w}}_{k+1} = -1 | \mathcal{F}_k\right) = 1 - q, \tag{49}$$

where $q \in (1/2, 1)$ and $\mathcal{F}_k = \sigma\left(\left(\underline{\phi}_0, \underline{\delta}_0, \underline{\tilde{w}}_0\right), \ldots, \left(\underline{\phi}_k, \underline{\delta}_k, \underline{\tilde{w}}_k\right)\right)$, with $\underline{\tilde{w}}_0 = 1$.

Let $\{\underline{\tau}_\epsilon\}_{\epsilon>0}$ be a family of stopping times with respect to $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$, parameterized by ϵ. A bound on $\mathbb{E}[\underline{\tau}_\epsilon]$ is derived in [9, 16, 28] under the following assumption.

Assumption 7. *The following hold for the stochastic process $\left\{\left(\underline{\phi}_k, \underline{\delta}_k, \underline{\tilde{w}}_k\right)\right\}_{k\in\mathbb{N}}$.*

(i) *There exist* $\lambda \in (0,\infty)$ *and* $\delta_{\max} = \delta_0 e^{\lambda j_{\max}}$ *for some* $j_{\max} \in \mathbb{Z}$ *such that* $\underline{\delta}_k \leq \delta_{\max}$ *for all* k.

(ii) *There exists* $\delta_\epsilon = \delta_0 e^{\lambda j_\epsilon}$, *for some* $j_\epsilon \in \mathbb{Z}$ *with* $j_\epsilon \leq 0$ *such that for all* k

$$\mathbb{1}_{\{\underline{\tau}_\epsilon > k\}} \underline{\delta}_{k+1} \geq \mathbb{1}_{\{\underline{\tau}_\epsilon > k\}} \min\left\{\underline{\delta}_k e^{\lambda \tilde{\underline{w}}_{k+1}}, \delta_\epsilon\right\}, \tag{50}$$

where $\tilde{\underline{w}}_{k+1}$ *satisfies* (49).

(iii) *There exists a nondecreasing function* $h : [0,\infty) \to (0,\infty)$ *and a constant* $\varrho > 0$ *such that for all* k

$$\mathbb{E}\left[\underline{\phi}_{k+1} - \underline{\phi}_k | \mathcal{F}_k\right] \mathbb{1}_{\{\underline{\tau}_\epsilon > k\}} \leq -\varrho h\left(\underline{\delta}_k\right) \mathbb{1}_{\{\underline{\tau}_\epsilon > k\}}.$$

Noticing that the event $\{\underline{\delta}_k \geq \delta_\epsilon\}$ occurs sufficiently frequently on average, often $\mathbb{E}\left[\underline{\phi}_{k+1} - \underline{\phi}_k\right]$ can be bounded by some negative fixed constant [9], thus allowing a bound on the expected stopping time $\mathbb{E}\left[\underline{\tau}_\epsilon\right]$, as stated next.

Theorem 5.1. *Under Assumption 7,* $\mathbb{E}\left[\underline{\tau}_\epsilon\right] \leq \frac{q}{2q-1} \cdot \frac{\underline{\phi}_0}{\varrho h(\delta_\epsilon)} + 1$.

5.2 Expected complexity result.

Consider the process $\left\{\left(\underline{\phi}_k, \underline{\delta}_k, \tilde{\underline{w}}_k\right)\right\}_{k\in\mathbb{N}}$, where $\underline{\phi}_k$ is defined in Theorem 4.1, $\underline{\delta}_k$ is the trust-region radius, and $\tilde{\underline{w}}_k$ is defined by $\tilde{\underline{w}}_k = 2\left(\mathbb{1}_{\mathcal{A}_k} \mathbb{1}_{I_k^Q} \mathbb{1}_{J_k^Q} - \frac{1}{2}\right)$. Given $\epsilon \in (0,1)$, let $\underline{\tau}_\epsilon$ be the number of iterations required by Algorithm 1 to first drive the norm of the gradient of f below ϵ:

$$\underline{\tau}_\epsilon := \inf\left\{k \in \mathbb{N} : \|\nabla f(\boldsymbol{\underline{x}_k})\|_2 \leq \epsilon\right\}.$$

Then $\underline{\tau}_\epsilon$ is a stopping time for the stochastic process generated by Algorithm 1, and hence for $\left\{\left(\underline{\phi}_k, \underline{\delta}_k, \tilde{\underline{w}}_k\right)\right\}_{k\in\mathbb{N}}$ [9, 16, 28]. More precisely, the occurrence of $\{\underline{\tau}_\epsilon = k\}$ can be determined by observing $\left(\underline{\phi}_0, \underline{\delta}_0, \tilde{\underline{w}}_0\right)$, $\ldots$ $\left(\underline{\phi}_{k-1}, \underline{\delta}_{k-1}, \tilde{\underline{w}}_{k-1}\right)$, which means that $\underline{\tau}_\epsilon$ is a stopping time with respect to the filtration $\{\mathcal{F}_{k-1}\}_{k\in\mathbb{N}}$. To bound $\mathbb{E}\left[\underline{\tau}_\epsilon\right]$ using Theorem 5.1, we show next that Assumption 7 holds for $\left\{\left(\underline{\phi}_k, \underline{\delta}_k, \tilde{\underline{w}}_k\right)\right\}_{k\in\mathbb{N}}$.

From (30), since $\mathcal{F}_{k-1} \subset \mathcal{F}_{k-1}^Q$, we observe that[6] for all $k \in \mathbb{N}$:

$$\mathbb{E}\left[\underline{\phi}_{k+1} - \underline{\phi}_k | \mathcal{F}_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\underline{\phi}_{k+1} - \underline{\phi}_k | \mathcal{F}_{k-1}^Q\right] | \mathcal{F}_{k-1}\right] \leq \mathbb{E}\left[-\varrho \underline{\delta}_k^2 | \mathcal{F}_{k-1}\right] = -\varrho \underline{\delta}_k^2. \tag{51}$$

Note that (51) holds for any realizations of the random variables $\mathbb{E}\left[\underline{\phi}_{k+1} - \underline{\phi}_k | \mathcal{F}_{k-1}\right]$ and $\underline{\delta}_k^2$ and hence on the event $\{\underline{\tau}_\epsilon > k\}$ in particular, which shows that Assumption 7-(iii) holds with the constant $\varrho = \frac{1}{2}\beta_f(1-\nu)(1-\frac{1}{\gamma^2})$ of Theorem 4.1 and $h(t) = t^2$. Assumption 7-(i) automatically follows from the initialization strategy in Algorithm 1 with $\lambda = \ln(\gamma)$. Before showing by means of Lemma 5.1, inspired by [9, Lemma 7], that (50) is satisfied, we first define the constant

$$\delta_\epsilon := \frac{\epsilon}{\xi}, \quad \text{with } \xi \geq \frac{1}{\alpha_Q} \max\left\{\kappa_h + \kappa_{eg}, \eta_2 + \kappa_{eg}, \frac{8\kappa_{ef}}{\kappa_{fcd}(1-\eta_1)} + \kappa_{eg}\right\} =: \hat{\xi}, \tag{52}$$

inspired by the proof of Theorem 4.2, especially (47). Then, following [9] exactly, one can assume without loss of generality that $\delta_\epsilon = \gamma^i \delta_0$ for some integer $i =: j_\epsilon \leq 0$, whence $\underline{\delta}_k = \gamma^{\underline{i}_k} \delta_\epsilon$ for any k and some integer $\underline{i}_k$.

Lemma 5.1. *Let all the assumptions made in Theorem 4.2 hold. Then* (50) *is satisfied for* $\tilde{\underline{w}}_k = 2\left(\mathbb{1}_{\mathcal{A}_k} \mathbb{1}_{I_k^Q} \mathbb{1}_{J_k^Q} - \frac{1}{2}\right)$, *with* $\lambda = \ln(\gamma)$ *and some fixed* $q \in (\tilde{\beta}, 1)$ *with* $\tilde{\beta} = \beta_f \beta_m (1-\beta_Q)$.

[6] Unlike the analysis in [9], the first equality of (51) is essential and due to the fact that in the present framework $\underline{\tau}_\epsilon$ is a stopping time with respect to σ-algebras $\mathcal{F}_{k-1}$ that are smaller than $\mathcal{F}_{k-1}^Q$ with respect to which $\{\underline{\phi}_k\}_{k\in\mathbb{N}}$ was proved in Theorem 4.1 Eq. (30) to be a supermartingale.

Proof. The result is proved by suitably adapting the proof of [9, Lemma 7] as was done in [16]. First, we notice that (50) trivially holds when $\mathbb{1}_{\{\tau_\epsilon > k\}} = 0$. Next, we show that when $\mathbb{1}_{\{\tau_\epsilon > k\}} = 1$, then

$$\underline{\delta}_{k+1} \geq \min\left\{\delta_\epsilon, \min\left\{\delta_{\max}, \gamma\underline{\delta}_k\right\}\mathbb{1}_{\mathcal{A}_k}\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q} + \gamma^{-1}\underline{\delta}_k\left(1 - \mathbb{1}_{\mathcal{A}_k}\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q}\right)\right\}.$$

Recall that $\delta_k = \gamma^{i_k}\delta_\epsilon$ for some integer i_k. Therefore, if $\delta_k > \delta_\epsilon$, then $\delta_k \geq \gamma\delta_\epsilon$, which implies that $\delta_{k+1} \geq \gamma^{-1}\delta_k \geq \delta_\epsilon$. Now assume that $\delta_k \leq \delta_\epsilon$. Since $\tau_\epsilon > k$, we have $\|\nabla f(\boldsymbol{x_k})\|_2 > \epsilon = \xi\delta_\epsilon \geq \xi\delta_k$. Thus, it follows from the definition of ξ that (47) holds. If $\mathbb{1}_{\mathcal{A}_k} = 1$, $\mathbb{1}_{I_k^Q} = 1$, and $\mathbb{1}_{J_k^Q} = 1$, then (48) holds, and consequently the kth iteration is successful, as was explained in the proof of Theorem 4.2. Hence, $\boldsymbol{x_{k+1}} = \boldsymbol{x_k} + \boldsymbol{Q_k s_k}$ and $\delta_{k+1} = \min\{\delta_{\max}, \gamma\delta_k\}$. If $\mathbb{1}_{\mathcal{A}_k}\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q} = 0$, then it always holds that $\delta_{k+1} \geq \gamma^{-1}\delta_k$. The proof is completed by observing that $\mathbb{P}\left(\mathbb{1}_{\mathcal{A}_k}\mathbb{1}_{I_k^Q}\mathbb{1}_{J_k^Q} = 1\right) = q$ for some fixed $q \geq \tilde{\beta}$. □

The main complexity result is provided by the next theorem.

Theorem 5.2. *Let all the assumptions made in Theorem 4.2 hold. Then*

$$\mathbb{E}\left[\tau_\epsilon\right] \leq \frac{\tilde{\beta}}{2\tilde{\beta} - 1} \cdot \frac{\underline{\phi}_0\tilde{\xi}^2}{\varrho\epsilon^2} + 1$$

for some $\tilde{\xi} \geq \hat{\xi}$, where ϱ is the constant of Theorem 4.1, $\tilde{\beta}$ is the same constant of Lemma 5.1, and $\hat{\xi}$ is defined in (52).

Proof. Recall the choice of q in Lemma 5.1, and pick some $\tilde{\xi} \geq \hat{\xi}$ such that $\delta_\epsilon = \frac{\epsilon}{\tilde{\xi}} = \gamma^i\delta_0$ for some integer $i \leq 0$ (without loss of generality), as discussed above. The proof follows by employing Theorem 5.1 with $h(t) = t^2$, which yields

$$\mathbb{E}\left[\tau_\epsilon\right] \leq \frac{q}{2q - 1} \cdot \frac{\underline{\phi}_0\tilde{\xi}^2}{\varrho\epsilon^2} + 1 \leq \frac{\tilde{\beta}}{2\tilde{\beta} - 1} \cdot \frac{\underline{\phi}_0\tilde{\xi}^2}{\varrho\epsilon^2} + 1.$$

□

6 Numerical results.

We now illustrate the performance of STARS for various random subspace and full space (i.e., STORM-like) forms. We consider stochastically noisy variants of 40 deterministic unconstrained problems ranging in dimension from $n = 98$ to $n = 125$; see Table 1. All objective functions are sums of squares, that is, $f(\boldsymbol{x}) = \sum_{i=1}^m f_i(\boldsymbol{x})^2$, and are corrupted with either *additive* or *multiplicative* stochastic noise. In the former case, the noisy $f_{\underline{\theta}}$ is given by $f_{\underline{\theta}}(\boldsymbol{x}) = f(\boldsymbol{x}) + \underline{\theta}$; in the latter, $f_{\underline{\theta}}(\boldsymbol{x}) = f(\boldsymbol{x})(1 + \underline{\theta})$. In both cases, the centered random variable $\underline{\theta}$ with standard deviation $\sigma = 10^{-3}$ is either normally or uniformly distributed. In order to take into account the variability due to the stochastic noise or random subspaces, 20 replications (corresponding to random seeds common across the tests) were performed for each of the 40 problems; combined with the two distributions of $\underline{\theta}$, we thus have a total of 1,600 *problem instances*.

For all tested variants, we use p-dimensional linear interpolation models $\hat{m}_k$, where $\nabla\hat{m}_k = \hat{\boldsymbol{g}}_{\boldsymbol{k}}$ is obtained from the forward finite-difference approximation from Corollary 3.3. We employ the forward finite-difference parameter $h = \min\{h_{\text{opt}}, \delta_k\}$, with h_{opt} obtained following [26] and provided in Table 1 and δ_k denoting the current trust-region radius. We employ such linear interpolation models so as to focus on differences due to the size and form of $\boldsymbol{Q_k^\top}$ and since the corresponding trust-region subproblems can be solved exactly: $\boldsymbol{s_k^\star} = -\delta_k\frac{\hat{\boldsymbol{g}}_{\boldsymbol{k}}}{\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2}$ if $\hat{\boldsymbol{g}}_{\boldsymbol{k}} \neq \boldsymbol{0}$ and $\boldsymbol{s_k^\star} = \boldsymbol{0}$ otherwise. When computing estimates of unknown function values at each iteration by means of a Monte Carlo approach using $n_k = 25$ noisy function evaluations for all k, available samples from previous iterations are reused, following the strategy described in the last paragraph of [2, Section 2.3], which was also used in [17, Section 5.2]. For $\boldsymbol{Q_k^\top}$, motivated by Theorems 3.1 and 3.2, we tested both Gaussian and $(r = 1)$-Hashing strategies. We found these to perform comparably for the tested settings and hence report the results for the Gaussian case (labeled G-STARS-p) whereby the entries of

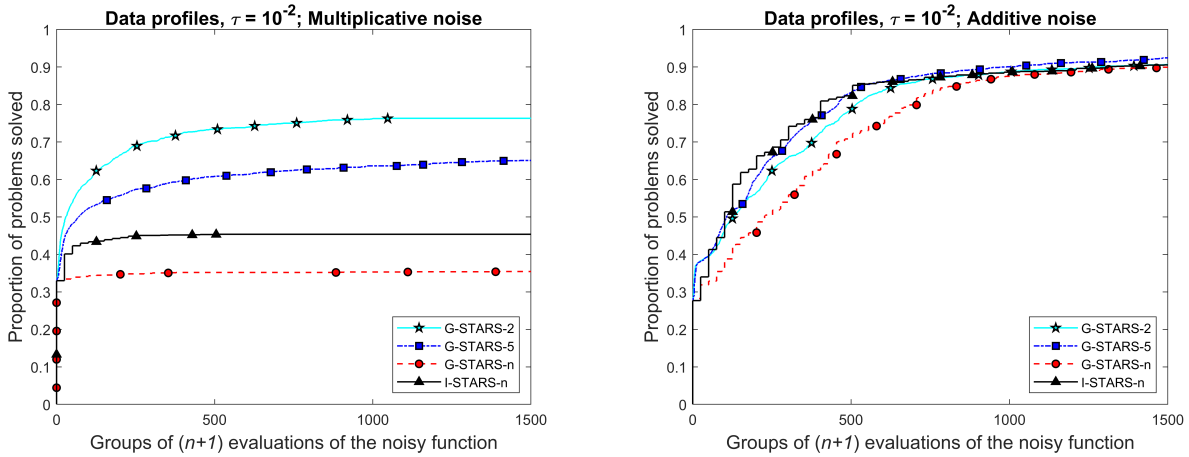

Figure 1: Data profiles for convergence tolerance $\tau = 10^{-2}$ on 1,600 problem instances for multiplicative noise (left) and additive noise (right) with standard deviation $\sigma = 10^{-3}$.

$\boldsymbol{Q}_{\boldsymbol{k}}^{\top} \in \mathbb{R}^{p \times n}$ are distributed as $\mathcal{N}(0, 1/p)$. We also test a STORM-like instance of STARS (labeled I-STARS-n) where $\boldsymbol{Q}_{\boldsymbol{k}} = \boldsymbol{I}_{\boldsymbol{n}} \in \mathbb{R}^{n \times n}$. All algorithmic variants used the parameters $\gamma = 2$, $\eta_2 = 90\eta_1 = 0.9$, $\delta_{\max} = 5$, $c_1 = 1$ and $\delta_0 = 1$.

All STARS variants are assessed by using data profiles [25]. For each of the 1,600 noisy problem instances, let $\boldsymbol{x}_{\boldsymbol{N}}$ be the point with the best f function value obtained by an algorithm after N evaluations of $f_{\underline{\theta}}$, denote by $f^\star$ the least such value found by all the algorithms, and let $\boldsymbol{x}_{\boldsymbol{0}}$ be the starting point. A problem is considered successfully solved within a convergence tolerance $\tau \in [0, 1]$ after N evaluations if

$$f(\boldsymbol{x}_{\boldsymbol{N}}) \leq f^\star + \tau(f(\boldsymbol{x}_{\boldsymbol{0}}) - f^\star).$$

The horizontal and vertical axes of the data profiles show, respectively, the number of noisy function evaluations divided by $(n + 1)$ and the proportion of problems solved. During the experiments a budget of $1{,}500(n + 1)$ noisy function evaluations is allocated to all the algorithms. For multiplicative noise, Figure 1 (left) shows that when the overall noise (i.e., $f(\boldsymbol{x})\underline{\theta}$) is large (which is the case for many of these problems: for nearly half of these problem instances $f(\boldsymbol{x}_0) \geq 10^3$), the G-STARS variants with small values of p perform well relative to the full space I-STARS-n variant. On the other hand, for this sampling budget and additive noise level, Figure 1 (right) shows that I-STARS-n is competitive with G-STARS with $p = 2, 5$. Unsurprisingly, in both cases, G-STARS-n (which uses $p = \min\{n, 100\}$) is outperformed by the I-STARS-n, which is using a better conditioned, deterministic matrix throughout.

Conclusion.

This work introduces STARS, the first DFO algorithm developed for stochastic objective functions that achieves scalability using random models constructed in low-dimensional random subspaces. The analysis of STARS extends an existing framework of model-based stochastic DFO (where randomness comes only from the stochasticity of the objective function), to settings where additional randomness stems from the mechanics of the algorithm. Making use of a supermartingale-based framework, we prove that the expected complexity of STARS, which uses subspace models, is similar to that of stochastic DFO algorithms in a smooth nonconvex setting. Numerical experiments demonstrate the performance of STARS on large-scale problems using linear interpolation models in subspaces of various dimensions.

We note that for ease of exposition we have focused in (1) on $f_{\underline{\theta}}$ being an unbiased estimator of f, but this can be relaxed in all ways that the STORM framework can address. Similarly, for concreteness, the stated algorithm and numerical results have focused respectively on linear and quadratic models, but the analysis readily applies to more general probabilistically fully linear models in subspaces. Apart from the use of particular random interpolation models, the present work also employs a classical Monte Carlo sampling

strategy for computing estimates. Future works can improve this sampling strategy while also using more general random models.

References

[1] S. K. Anagnostidis, A. Lucchi, and Y. Diouane, *Direct-search for a class of stochastic min-max problems*, in International Conference on Artificial Intelligence and Statistics, 2021, pp. 3772–3780.

[2] C. Audet, K. J. Dzahini, M. Kokkolaras, and S. Le Digabel, *Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates*, Computational Optimization and Applications, 79 (2021), pp. 1–34, https://doi.org/10.1007/s10589-020-00249-0.

[3] C. Audet and W. Hare, *Derivative-Free and Blackbox Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, Cham, Switzerland, 2017, https://doi.org/10.1007/978-3-319-68913-5.

[4] A. S. Bandeira and R. Van Handel, *Sharp nonasymptotic bounds on the norm of random matrices with independent entries*, The Annals of Probability, 44 (2016), pp. 2479–2506, https://doi.org/10.1214/15-AOP1025.

[5] A. S. Berahas, L. Cao, and K. Scheinberg, *Global convergence rate analysis of a generic line search algorithm with noise*, SIAM Journal on Optimization, 31 (2021), pp. 1489–1518, https://doi.org/10.1137/19M1291832.

[6] E. Bergou, Y. Diouane, V. Kungurtsev, and C. W. Royer, *A stochastic Levenberg–Marquardt method using random models with complexity results*, SIAM/ASA Journal on Uncertainty Quantification, 10 (2022), pp. 507–536, https://doi.org/10.1137/20M1366253.

[7] R. N. Bhattacharya and E. C. Waymire, *A Basic Course in Probability Theory*, vol. 69, Springer, 2007, https://doi.org/10.1007/978-3-319-47974-3.

[8] P. Billingsley, *Probability and Measure*, Wiley, New York, USA, third ed., 1995.

[9] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg, *Convergence rate analysis of a stochastic trust region method via submartingales*, INFORMS Journal on Optimization, 1 (2019), pp. 92–119, https://doi.org/10.1287/ijoo.2019.0016.

[10] R. Bollapragada and S. M. Wild, *Adaptive sampling quasi-Newton methods for zeroth-order stochastic optimization*, Tech. Report 2109.12213, ArXiv, 2021, https://arxiv.org/abs/2109.12213.

[11] C. Cartis and L. Roberts, *Scalable subspace methods for derivative-free nonlinear least-squares optimization*, Preprint 2102.12016, arXiv, 2021, https://arxiv.org/abs/2102.12016.

[12] K. H. Chang, *Stochastic Nelder–Mead simplex method - a new globally convergent direct search method for simulation optimization*, European Journal of Operational Research, 220 (2012), pp. 684–694, https://doi.org/10.1016/j.ejor.2012.02.028.

[13] R. Chen, M. Menickelly, and K. Scheinberg, *Stochastic optimization using a trust-region method and random models*, Mathematical Programming, 169 (2018), pp. 447–487, https://doi.org/10.1007/s10107-017-1141-8.

[14] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*, SIAM, Philadelphia, 2009, https://doi.org/10.1137/1.9780898718768.

[15] F. E. Curtis and K. Scheinberg, *Adaptive stochastic optimization: A framework for analyzing stochastic optimization algorithms*, IEEE Signal Processing Magazine, 37 (2020), pp. 32–42, https://doi.org/10.1109/MSP.2020.3003539.

[16] K. J. Dzahini, *Expected complexity analysis of stochastic direct-search*, Computational Optimization and Applications, 81 (2022), pp. 179–200, https://doi.org/10.1007/s10589-021-00329-9.

[17] K. J. Dzahini, M. Kokkolaras, and S. Le Digabel, *Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates*, Mathematical Programming, (2022), https://doi.org/10.1007/s10107-022-01787-7.

[18] J. C. Gross and G. T. Parks, *Optimization by moving ridge functions: derivative-free optimization for computationally intensive functions*, Engineering Optimization, 54 (2021), pp. 553–575, https://doi.org/10.1080/0305215X.2021.1886286.

[19] B. Jin, K. Scheinberg, and M. Xie, *High probability complexity bounds for line search based on stochastic oracles*, in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 9193–9203.

[20] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Contemporary mathematics, 26 (1984), pp. 189–206.

[21] D. M. Kane and J. Nelson, *Sparser Johnson–Lindenstrauss transforms*, Journal of the ACM, 61 (2014), pp. 1–23, https://doi.org/10.1145/2559902.

[22] D. E. Knuth, *Big omicron and big omega and big theta*, ACM Sigact News, 8 (1976), pp. 18–24, https://doi.org/10.1145/1008328.1008329.

[23] J. Larson, M. Menickelly, and S. M. Wild, *Derivative-free optimization methods*, Acta Numerica, 28 (2019), pp. 287–404, https://doi.org/10.1017/s0962492919000060.

[24] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Annals of Statistics, 28 (2000), pp. 1302–1338, https://www.jstor.org/stable/2674095.

[25] J. J. Moré and S. M. Wild, *Benchmarking derivative-free optimization algorithms*, SIAM Journal on Optimization, 20 (2009), pp. 172–191, https://doi.org/10.1137/080724083.

[26] J. J. Moré and S. M. Wild, *Estimating derivatives of noisy simulations*, ACM Transactions on Mathematical Software, 38 (2012), pp. 19:1–19:21, https://doi.org/10.1145/2168773.2168777.

[27] A. Neumaier, H. Fendl, H. Schilly, and T. Leitner, *VXQR: derivative-free unconstrained optimization based on QR factorizations*, Soft Computing, 15 (2011), pp. 2287–2298, https://doi.org/10.1007/s00500-010-0652-5.

[28] C. Paquette and K. Scheinberg, *A stochastic line search method with expected complexity analysis*, SIAM Journal on Optimization, 30 (2020), pp. 349–376, https://doi.org/10.1137/18M1216250.

[29] F. Rinaldi, L. N. Vicente, and D. Zeffiro, *A weak tail-bound probabilistic condition for function estimation in stochastic derivative-free optimization*, arXiv, (2022), https://arxiv.org/abs/2202.11074.

[30] H. Robbins and S. Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics, 22 (1951), pp. 400–407, http://www.jstor.org/stable/2236626.

[31] L. Roberts and C. W. Royer, *Direct search based on probabilistic descent in reduced spaces*, arXiv, (2022), https://arxiv.org/abs/2204.01275.

[32] Z. Shao, *On Random Embeddings and their Applications to Optimization*, PhD thesis, University of Oxford, 2022.

[33] 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, 28 (2018), pp. 3145–3176, https://doi.org/10.1137/15M1042425.

[34] T. Tao, *Topics in Random Matrix Theory*, AMS, 2012, https://doi.org/10.1090/gsm/132.

[35] D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Foundations and Trends in Theoretical Computer Science, 10 (2014), pp. 1–157, https://doi.org/10.1561/0400000060.

Appendix.

This appendix presents the proofs of a series of results in the main body of the manuscript.

Proof of Theorem 3.2.

Proof. Since a detailed proof is provided in [21], only its main idea is presented here. We note as mentioned in [21, Section 1.1] that one can assume without any loss of generality that $\|\boldsymbol{v}\|_2 = 1$, in which case the result follows by showing that $\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2^2 \in \left[(1-\varepsilon)^2, (1+\varepsilon)^2\right]$, which is implied by $\left|\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2^2 - 1\right| \leq 2\varepsilon - \varepsilon^2$. Thus, it suffices to show that for any unit norm $\boldsymbol{v}$, $\mathbb{P}\left(\left|\|\underline{\boldsymbol{S}}\boldsymbol{v}\|_2^2 - 1\right| > 2\varepsilon - \varepsilon^2\right) < \beta$, which is proved in [21, Theorem 4.3] for the above choices of ℓ, r, and p. □

Proof of Lemma 3.2.

Proof. First, we note that

$$\begin{aligned}\left|\hat{f}(\boldsymbol{s}+\boldsymbol{d}) - \hat{f}(\boldsymbol{s}) - \boldsymbol{d}^\top \nabla \hat{f}(\boldsymbol{s})\right| &= \left|f(\boldsymbol{x_k} + \boldsymbol{Q_k}(\boldsymbol{s}+\boldsymbol{d})) - f(\boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s}) - \boldsymbol{d}^\top \boldsymbol{Q_k}^\top \nabla f(\boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s})\right| \\ &= \left|f(\boldsymbol{y}+\boldsymbol{d_k}) - f(\boldsymbol{y}) - \boldsymbol{d_k}^\top \nabla f(\boldsymbol{y})\right|,\end{aligned}$$

where $\boldsymbol{y} = \boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s} \in \mathbb{R}^n$ and $\boldsymbol{d_k} = \boldsymbol{Q_k}\boldsymbol{d} \in \mathbb{R}^n$. It follows from the Fundamental Theorem of Calculus in $\mathbb{R}^n$ (see, e.g., [3, Lemma 9.3]) that for all $\boldsymbol{y}, \boldsymbol{d_k} \in \mathbb{R}^n$, $f(\boldsymbol{y}+\boldsymbol{d_k}) - f(\boldsymbol{y}) = \int_0^1 \boldsymbol{d_k}^\top \nabla f(\boldsymbol{y} + \tau \boldsymbol{d_k}) d\tau$. Thus,

$$\begin{aligned}\left|\hat{f}(\boldsymbol{s}+\boldsymbol{d}) - \hat{f}(\boldsymbol{s}) - \boldsymbol{d}^\top \nabla \hat{f}(\boldsymbol{s})\right| &= \left|\int_0^1 \boldsymbol{d_k}^\top \left[\nabla f(\boldsymbol{y} + \tau \boldsymbol{d_k}) - \nabla f(\boldsymbol{y})\right] d\tau\right| \leq \\ \int_0^1 \|\boldsymbol{d_k}\|_2 \|\nabla f(\boldsymbol{y} + \tau \boldsymbol{d_k}) - \nabla f(\boldsymbol{y})\|_2 d\tau &\leq L_g \|\boldsymbol{d_k}\|_2^2 \int_0^1 \tau d\tau \leq \frac{1}{2} L_g Q_{\max}^2 \|\boldsymbol{d}\|_2^2.\end{aligned}$$

□

Proof of Lemma 4.1.

Proof. The proof is similar to that of [13, Lemma 4.5] and is not detailed here. It follows from (2) together with the inequalities $\kappa_h \geq \max\left\{\left\|\hat{\boldsymbol{H}}_k\right\|, 1\right\}$ and $\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \geq \kappa_h \delta_k$ that $\hat{m}_k(\boldsymbol{0}) - \hat{m}_k(\boldsymbol{s_k}) \geq \frac{\kappa_{fcd}}{2} \|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \delta_k$. Since the model is $(\kappa_{ef}, \kappa_{eg}; \boldsymbol{Q_k})$-fully linear, then $f(\boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s_k}) - f(\boldsymbol{x_k}) \leq 2\kappa_{ef}\delta_k^2 + \hat{m}_k(\boldsymbol{s_k}) - \hat{m}_k(\boldsymbol{0}) \leq -\frac{\kappa_{fcd}}{4} \|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \delta_k$, where the last inequality follows from the fact that $\delta_k \leq \frac{\kappa_{fcd}}{8\kappa_{ef}} \|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2$. □

Proof of Lemma 4.2.

Proof. The proof is inspired by (and similar to) that of [13, Lemma 4.6]. By $(\kappa_{ef}, \kappa_{eg}; \boldsymbol{Q_k})$-full linearity and (22), it holds

$$\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \geq \left\|\boldsymbol{Q_k}^\top \nabla f(\boldsymbol{x_k})\right\|_2 - \kappa_{eg}\delta_k \geq \max\left\{\kappa_h, \frac{8\kappa_{ef}}{\kappa_{fcd}}\right\} \delta_k, \tag{53}$$

which shows that condition (20) is satisfied. Consequently f decreases as in (20). The first inequality in (53), together with (22), yields $\|\hat{\boldsymbol{g}}_{\boldsymbol{k}}\|_2 \geq \frac{4C_1}{\kappa_{fcd}} \left\|\boldsymbol{Q_k}^\top \nabla f(\boldsymbol{x_k})\right\|_2$, which, combined with (20), implies (21), and the proof is complete. □

Proof of Lemma 4.5.

Proof. According to the Fundamental Theorem of Calculus in $\mathbb{R}^n$ (see, e.g., [3, Lemma 9.3]), $\left|f(\boldsymbol{x_k}) - f(\boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s_k}) - \nabla f(\boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s_k})^\top (-\boldsymbol{Q_k}\boldsymbol{s_k})\right| \leq \frac{1}{2} L_g Q_{\max}^2 \delta_k^2$, with the inequality also following from $\|\boldsymbol{Q_k}\boldsymbol{s_k}\|_2 \leq Q_{\max}\delta_k$, which implies that

$$f(\boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s_k}) - f(\boldsymbol{x_k}) \leq \boldsymbol{s_k}^\top \left[\boldsymbol{Q_k}^\top \nabla f(\boldsymbol{x_k} + \boldsymbol{Q_k}\boldsymbol{s_k})\right] + \frac{1}{2} L_g Q_{\max}^2 \delta_k^2. \tag{54}$$

Table 1: The 40 problems considered.

No.	Problem	n	m	h_{opt}	No.	Problem	n	m	h_{opt}
1	ARGLALE	100	200	1e-03	21	EIGENC	110	110	1e-04
2	ARGLBLE	100	200	5e-04	22	EXTROSNB	100	100	1e-04
3	ARGLCLE	100	200	5e-04	23	FREUROTH	100	198	2e-04
4	ARTIF	100	100	4e-05	24	INTEGREQ	100	100	1e-05
5	ARWHDNE	100	198	1e-04	25	MANCINO	100	100	2e-03
6	BDQRTIC	100	192	9e-05	26	MOREBV	100	100	1e-07
7	BDVALUES	100	100	1e-02	27	MSQRTA	100	100	2e-04
8	BRATU2D	100	100	4e-06	28	MSQRTB	100	100	2e-04
9	BRATU2DT	100	100	5e-06	29	OSCIGRNE	100	100	9e-05
10	BRATU3D	125	125	1e-05	30	Penalty2	100	200	5e-05
11	BROWNALE	100	100	4e-04	31	PENLT1NE	100	101	8e-03
12	BROYDN3D	100	100	4e-05	32	POWELLSE	100	100	3e-04
13	BROYDNBD	100	100	4e-05	33	POWELLSG	100	100	2e-04
14	CBRATU2D	98	98	4e-06	34	ROSENBR	100	198	1e-04
15	CHANDHEQ	100	100	6e-05	35	SPMSQRT	100	164	3e-04
16	CHEBYQAD	100	100	4e-06	36	SROSENBR	100	100	4e-05
17	ConnBand	100	100	5e-05	37	VARDIMNE	100	102	1e-04
18	CUBE	100	100	4e-05	38	VarTrig	100	100	3e-03
19	EIGENA	110	110	2e-04	39	YATP1SQ	99	99	9e-04
20	EIGENB	110	110	6e-05	40	YATP2SQ	99	99	9e-04

Per Lipschitz continuity of ∇f, $\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}}+\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{k}})-\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2 \leq L_g Q_{\max}^2\delta_k$, which implies that

$$\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}}+\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{k}})\right\|_2 \leq L_g Q_{\max}^2\delta_k+\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2. \tag{55}$$

Thus, using (54) and (55), together with the inequality $\delta_k \leq \zeta^{-1}\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2$, yields

$$\begin{aligned} f(\boldsymbol{x}_{\boldsymbol{k}}+\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{k}}) \quad - \quad & f(\boldsymbol{x}_{\boldsymbol{k}}) \leq \left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}}+\boldsymbol{Q}_{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{k}})\right\|_2\delta_k+\frac{L_g}{2\zeta}Q_{\max}^2\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2\delta_k \\ \leq \quad & \left(\frac{L_g}{\zeta}Q_{\max}^2\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2+\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2\right)\delta_k \\ + \quad & \frac{L_g}{2\zeta}Q_{\max}^2\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2\delta_k=\left(1+\frac{3L_g}{2\zeta}Q_{\max}^2\right)\left\|\boldsymbol{Q}_{\boldsymbol{k}}^{\top}\nabla f(\boldsymbol{x}_{\boldsymbol{k}})\right\|_2\delta_k, \end{aligned}$$

which completes the proof. □