Adaptive sampling quasi-Newton methods for zeroth-order stochastic optimization

Raghu Bollapragada¹ · Stefan M. Wild²

Received: 15 October 2021 / Accepted: 21 January 2023 / Published online: 14 March 2023
This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2023

Abstract
We consider unconstrained stochastic optimization problems with no available gradient information. Such problems arise in settings from derivative-free simulation optimization to reinforcement learning. We propose an adaptive sampling quasi-Newton method where we estimate the gradients using finite differences of stochastic function evaluations within a common random number framework. We develop modified versions of a norm test and an inner product quasi-Newton test to control the sample sizes used in the stochastic approximations and provide global convergence results to the neighborhood of a locally optimal solution. We present numerical experiments on simulation optimization problems to illustrate the performance of the proposed algorithm. When compared with classical zeroth-order stochastic gradient methods, we observe that our strategies of adapting the sample sizes significantly improve performance in terms of the number of stochastic function evaluations required.

Keywords Derivative-free optimization · Stochastic oracles · Adaptive sampling · Common random numbers

Mathematics Subject Classification 90C56 · 65K05 · 90C15 · 90C30 · 90C53

Stefan M. Wild
wild@lbl.gov
Raghu Bollapragada
raghu.bollapragada@utexas.edu

1 Operations Research and Industrial Engineering, The University of Texas at Austin, Austin, TX 78712, USA
2 Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
1 Introduction

We consider unconstrained stochastic optimization problems of the form

$$\min_{x \in \mathbb{R}^d} F(x) = \mathbb{E}_\zeta [f(x, \zeta)],$$

(1)

where the deterministic objective function $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, $\zeta$ is a random variable with associated probability space $(\mathcal{X}, \mathcal{F}, P)$, and $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$. We consider the settings where one has access only to an oracle or a black-box procedure that, given values of $(x, \zeta)$, outputs the according realization of the stochastic function $f$ and cannot access explicit estimates of the gradient $\nabla F$. Such stochastic optimization problems arise in myriad science and engineering applications, from simulation optimization [12, 31, 40, 53, 54] to reinforcement learning [10, 47, 59]. Several methods have been proposed to solve such derivative-free stochastic optimization problems, and we refer the reader to [3, 44] for surveys of these methods. A popular class of these methods estimate the gradients using stochastic function values and employ standard gradient-based optimization methods using these estimators.

Quasi-Newton methods [52] are recognized as one of the most powerful methods for solving deterministic optimization problems. These methods build quadratic models of the objective information using only gradient information. Recently, researchers have been adapting these methods for stochastic settings when the gradient information is available. The empirical results in [17] indicate that a careful implementation of these methods can be efficient compared with the popular stochastic gradient methods. We adapt these methods to make them suitable for situations where the gradients are estimated using zeroth-order evaluations of the stochastic function $f$.

We propose finite-difference derivative-free stochastic quasi-Newton methods for solving (1) by exploiting common random number (CRN) evaluations of $f$. Common random numbers are a variance reduction technique [57, Chapter 8] increasingly employed for stochastic oracles that permit such evaluation [42]. The CRN setting allows us to define subsampled (i.e., employing a potentially strict subset of $\mathcal{X}$ as samples [15, 56]) gradient estimators for all $x \in \mathbb{R}^d$ and for a given finite-difference parameter $\nu > 0$

$$\nabla_{\text{FD}} F_{\{\zeta_i\}}(x) := \frac{f(x + \nu e_j, \zeta_i) - f(x, \zeta_i)}{\nu}, \quad j = 1, \ldots, d$$

(2)

$$\nabla_{\text{FD}} F_{S_k}(x) := \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \nabla_{\text{FD}} F_{\{\zeta_i\}}(x),$$

(3)

which employ forward differences for the independent and identically distributed (i.i.d.) samples of $\zeta$ in the set $S_k$ along each canonical direction $e_j \in \mathbb{R}^d$. CRN-based gradient estimates possess lower variance than do independent-sample-based gradient estimates. Moreover, CRNs can be employed in many practical settings, including policy optimization problems in reinforcement learning wherein a randomly sampled environmental state can be queried multiple times.
The performance of stochastic quasi-Newton methods is highly dependent on the quality of the gradient approximations. The gradient estimation considered in this work has two sources of error: error due to the finite-difference approximation and error due to the stochastic approximation. The latter error depends on the number of samples $|S_k|$ used in the estimation. Using too few samples affects the stability of a method using the estimates; using a large number of samples results in computational inefficiency. For settings where gradient information is available, researchers have developed practical tests to adaptively increase the sample sizes used in the stochastic approximations and have supported these tests with global convergence results [4, 5, 14, 17, 21, 33] to the optimal solution under the commonly employed assumptions that the variance in the stochastic gradients is bounded.

In this paper we modify these tests to address the challenges associated with the finite-difference approximation errors, and we demonstrate the resulting method on simulation optimization problems. We establish a global linear rate of convergence and global sublinear rate of convergence to a neighborhood of the solution for strongly convex and nonconvex functions $F$, respectively.

The paper is organized into five sections. A brief literature review and notation are provided in the rest of this section. Section 2 describes the components of our algorithm, and Sect. 3 establishes theoretical convergence results. Section 4 describes the algorithmic components for handling nonsmooth stochastic functions. Numerical experiments are provided in Sect. 5, and concluding remarks are provided in Sect. 6.

1.1 Literature review

Finite-difference-based versions of the standard stochastic gradient method (“stochastic approximation”) of Robbins and Monro [55] soon followed that work, in both univariate [39] and multivariate [13] settings. Stochastic approximation methods based on CRNs were analyzed in [41, 45].

Kelley [38] proposed and analyzed quasi-Newton methods for solving noisy problems with noise decaying as the iterates approach the solution. Berahas et al. [6] proposed a quasi-Newton method for solving noisy problems using finite-difference gradient estimators where the finite-difference parameter is carefully chosen based on the mechanism proposed by Moré and Wild [49] to ensure stability in the search directions. They considered the settings where the noise is assumed to be bounded and cannot be controlled. In our settings, the noise is stochastic with bounded variance (see Assumption D), can be unbounded, and is controlled within the CRN framework.

Different forms of gradient estimators [4], in addition to the finite-difference-based estimators, can be employed in solving derivative-free optimization problems. Recently, Berahas et al. [7] analyzed methods that employ various forms of gradient estimators in solving noisy derivative-free optimization problems. They established conditions on the gradient estimation errors that guarantee convergence to a neighborhood of the optimal solution.

Another class of methods that exploit CRN settings is that of two-point (or multipoint) bandit feedback. These methods include variants of mirror descent and random
search and were originally motivated by and analyzed for convex objectives [1, 27, 30, 32, 34, 46, 51, 58, 61, 64].

Related classes of methods for nonconvex stochastic optimization include zeroth-order extensions of both conditional gradient methods [4, 5, 33] and other proximal-point approaches [36, 37].

Model-based trust-region methods [12, 24, 28, 29, 43, 62, 63] and direct search methods [2, 23, 25, 26] are alternative approaches to gradient estimation-based methods.

1.2 Notation and Subsampled Gradient Estimator Preliminaries

Although we focus here on subsampled gradient estimators of the form in (3), our algorithmic framework and analysis extend to other settings, which we formalize here. We note that the source of stochasticity in the sequel is due to the random sampling of a sequence of sets \{S_k\}_k.

Given samples \( S_k = \{\zeta_1, \ldots, \zeta_{|S_k|}\} \), we define a subsampled function by

\[
F_{S_k}(x) := \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} f(x, \zeta_i) \quad \forall x \in \mathbb{R}^d.
\]

(4)

Our primary algorithmic assumption concerns the form of the randomized sampling performed to obtain \{S_k\}_k and hence the subsampled functions \( F_{S_0}, F_{S_1}, \ldots \).

Assumption A At every iteration \( k \), the sample set \( S_k \) consists of i.i.d. samples of \( \zeta \). As a consequence, for all \( x \in \mathbb{R}^d \) and \( k \in \mathbb{Z}_+ \),

\[
\mathbb{E}_{\zeta_i}[f(x, \zeta_i)] = F(x), \quad \forall \zeta_i \in S_k.
\]

From Assumption A, for any subsampled function \( F_{S_k}(x) \) of the form (4), we have that \( \mathbb{E}_{S_k}[F_{S_k}(x)] = F(x) \). Also from this assumption, we have that for the gradient estimator in (3) and any \( x_k \in \mathbb{R}^d \),

\[
\mathbb{E}_{S_k}\left[\nabla FD F_{S_k}(x_k)\right] = \mathbb{E}_{S_k}\left[\frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \left[f(x_k + \nu e_j, \zeta_i) - f(x_k, \zeta_i)\right]\nu\right]_{j=1}^d = \nabla FD F(x_k),
\]

(5)

where \( \nabla FD F(x) \) is the zeroth-order quantity based on deterministic forward differences:

\[
\nabla FD F(x) := \left[\frac{F(x + \nu e_j) - F(x)}{\nu}\right]_{j=1}^d \quad \forall x \in \mathbb{R}^d.
\]

(6)

We also make assumptions about the smoothness of the expected function \( F \) and the stochastic function \( f \). The first such assumption concerns the smoothness of the
objective function $F$. We note that this assumption is slightly weaker than the next assumption requiring differentiability of the stochastic functions $f(\cdot, \zeta)$.

**Assumption B** The function $F$ in (1) is continuously differentiable and has Lipschitz continuous gradients with Lipschitz constant $L F > 0$.

When combined with Assumption A, Assumption B implies that $\nabla F(x_k)$ is an unbiased estimator of $\nabla F(x_k)$ but $\nabla F_S(x_k)$ is a biased estimator of the gradient $\nabla F(x_k)$ and that the bias can be deterministically bounded by

$$
\left\| \nabla F(x_k) - \nabla F_S(x_k) \right\|^2 = \sum_{j=1}^d \left( \frac{F(x_k + v e_j) - F(x_k)}{\nu} - \nabla F(x_k)_j \right)^2 \\
\leq \sum_{j=1}^d \left( \frac{L F \nu}{2} \right)^2 \\
= \left( \frac{L F \nu}{d} \right)^2,
$$

(7)

where the inequality follows from the following result, which holds for functions $F$ with $L F$-Lipschitz continuous gradients.

**Lemma 1** (Descent Lemma [11]) If $F : \mathbb{R}^d \mapsto \mathbb{R}$ is continuously differentiable with a $L F$-Lipschitz continuous gradient on $\mathbb{R}^d$, then

$$
F(y) \leq F(x) + (y - x)^T \nabla F(x) + \frac{L F}{2} \|y - x\|^2 \quad \text{for all } x, y \in \mathbb{R}^d.
$$

The bias term in (7) is a direct result of the absence of gradient information (and thus the derivative-free estimation), and we design the components of our proposed algorithm accordingly.

Our sample size selection techniques in Sect. 2.1 will rely on Assumption A and thus do not require the subsampled gradients to exist. That is, the sampling procedure works even when the individual or subsampled functions are nondifferentiable as long as the expected function $F$ is differentiable.

For deriving the remaining components of the algorithm, we will make use of the additional assumption that the subsampled gradients exist and are Lipschitz continuous.

**Assumption C** For every $\zeta$, the function $f(\cdot, \zeta)$ in (1) is continuously differentiable and has Lipschitz continuous gradients with Lipschitz constant $L f \zeta > 0$.

Assumption C implies that any subsampled gradient

$$
\nabla F_{S_k}(x) := \frac{1}{|S_k|} \sum_{\zeta \in S_k} \nabla x f(x, \zeta) \quad \forall x \in \mathbb{R}^d,
$$
is Lipschitz continuous with Lipschitz constant $L_{\nabla f}$. Assumption C is strictly stronger than Assumption B since the former ensures that $L_{\nabla F} = L_{\nabla f}$ is a Lipschitz constant for $\nabla F$. In Sect. 4, we employ the weaker Assumption B and modify the algorithmic components accordingly.

Our final general-purpose assumption concerns the variance in the stochastic functions $f$. We note that this assumption is weaker than requiring that the variance be bounded uniformly.

**Assumption D** The variance in the stochastic functions is bounded by the norm of the gradient of the expected function. That is, there exist scalars $\omega_1, \omega_2 \geq 0$ such that

$$
\mathbb{E}_\zeta \left[ (f(x, \zeta) - F(x))^2 \right] \leq \omega_1^2 + \omega_2^2 \| \nabla F(x) \|^2 \quad \forall x \in \mathbb{R}^d.
$$

Before proceeding, we note that the generated $x_{k+1}$ is a random variable for $k \in \mathbb{Z}_+$; however, when conditioned on $x_k$ (i.e., conditioned on the filtration $\mathbb{T}_k = \sigma(x_0, S_1, S_2, \ldots, S_{k-1})$), the only remaining source of randomness is from the sample set $S_k$. For ease of exposition, we let $\mathbb{E}_{S_k}[\cdot]$ denote the conditional expectation given $x_k$ (i.e., $\mathbb{E}[\cdot|\mathbb{T}_k]$).

## 2 A zeroth-order stochastic quasi-Newton algorithm

The update form of a finite-difference, zeroth-order stochastic quasi-Newton method is given by

$$
x_{k+1} = x_k - \alpha_k H_k \nabla^{FD} F_{S_k}(x_k),
$$

where $\alpha_k > 0$ is the step length, $H_k$ is a positive-definite quasi-Newton matrix, and $\nabla^{FD} F_{S_k}(x_k)$ is a finite-difference, subsampled (or batch) gradient estimate defined by (3). While we consider here forward finite differences to estimate the subsampled gradient, we note that other derivative-free techniques (e.g., central finite differences, polynomial interpolation; see [44]) can be employed to estimate the gradient.

We now discuss the algorithmic components consisting of sample size selection (Sect. 2.1), finite-difference parameter and step-length selection (Sects. 2.2 and 2.3, respectively), and quasi-Newton updates (Sect. 2.4). The complete algorithm is formally stated as Algorithm 1.

### 2.1 Sample size selection

We propose to control the sample sizes $|S_k|$ used in the gradient estimation in order to achieve fast convergence. We explore two different strategies to control the sample sizes in settings where no gradient information is available (i.e., based only on zeroth-order information). We note that the resulting strategies are useful in settings beyond derivative-free ones; they can be applied in any setting where biased gradient estimators are found.
2.1.1 Norm test

A popular deterministic condition (see, e.g., Equation (3.2) in [21], Equation (15) in [22]) for gradient estimators \( g_k \) to satisfy is the norm condition given by

\[
\| g_k - \nabla F(x_k) \|^2 \leq \theta^2 \| \nabla F(x_k) \|^2, \quad \theta > 0.
\] (9)

Satisfying (9) in expectation is the basis for controlling the sample sizes used in subsampled gradient methods; that is,

\[
\mathbb{E}_{S_k} \left[ \| g_k - \nabla F(x_k) \|^2 \right] \leq \theta^2 \| \nabla F(x_k) \|^2, \quad \theta > 0.
\]

One can employ this condition on a finite-difference subsampled gradient estimator such as (3); that is,

\[
\mathbb{E}_{S_k} \left[ \| \nabla_{FD} F_{S_k}(x_k) - \nabla F(x_k) \|^2 \right] \leq \theta^2 \| \nabla F(x_k) \|^2, \quad \theta > 0.
\] (10)

However, it is not always possible to satisfy this condition because of the inherent bias in the finite-difference subsampled gradient estimator:

\[
\nabla_{FD} F_{S_k}(x_k) - \nabla F(x_k) = \frac{\nabla_{FD} F_{S_k}(x_k) - \nabla_{FD} F(x_k)}{\text{sampling error}} + \frac{\nabla_{FD} F(x_k) - \nabla F(x_k)}{\text{bias}},
\] (11)

where \( \nabla_{FD} F \) is the deterministic finite-difference estimator in (6).

For any finite-difference parameter \( \nu > 0 \), the second term in (11) can be nonzero, and thus condition (10) may not be satisfied (e.g., at points where \( \nabla F(x_k) \) is close to zero). Moreover, sample selection will affect only the first term in (11). Therefore, we propose to look at the norm condition on the finite-difference subsampled gradient estimation error. In particular, we use the condition

\[
\mathbb{E}_{S_k} \left[ \| \nabla_{FD} F_{S_k}(x_k) - \nabla_{FD} F(x_k) \|^2 \right] \leq \theta^2 \| \nabla_{FD} F(x_k) \|^2, \quad \theta > 0.
\] (12)

This condition relaxes the right-hand side of (10). That is,

\[
\begin{align*}
\mathbb{E}_{S_k} \left[ \| \nabla_{FD} F_{S_k}(x_k) - \nabla F(x_k) \|^2 \right] & \leq \mathbb{E}_{S_k} \left[ \| \nabla_{FD} F_{S_k}(x_k) - \nabla_{FD} F(x_k) \|^2 \right] + \| \nabla_{FD} F(x_k) - \nabla F(x_k) \|^2 \\
& \leq \theta^2 \| \nabla_{FD} F(x_k) \|^2 + \| \nabla_{FD} F(x_k) - \nabla F(x_k) \|^2 \\
& \leq 2\theta^2 \| \nabla F(x_k) \|^2 + \left( 1 + 2\theta^2 \right) \| \nabla_{FD} F(x_k) - \nabla F(x_k) \|^2
\end{align*}
\]
\[ \leq 2\theta^2 \| \nabla F(x_k) \|^2 + \left(1 + 2\theta^2\right) \frac{L^2}{4} \nabla F \nu^2 d, \]

where the first inequality is due to expansion of the square term and (5), the second inequality is due to (12), the third inequality is due to the fact that \((a+b)^2 \leq 2(a^2+b^2)\), and the last inequality is due to (7). Therefore, our condition (12) is less restrictive than (10) and can be satisfied at all \(x_k\).

The left-hand side of (12) is difficult to compute but can be bounded by the true variance of individual finite-difference gradient estimators \(\nabla \text{FD}_F\{\zeta_i\};\) recall (2)). That is,

\[ E_{S_k} \left[ \| \nabla \text{FD}_F S_k(x_k) - \nabla \text{FD}_F F(x_k) \|^2 \right] \leq \frac{E_{\zeta_i} \left[ \| \nabla \text{FD}_F F\{\zeta_i\}(x_k) - \nabla \text{FD}_F F(x_k) \|^2 \right]}{|S_k|}. \tag{13} \]

To be meaningful, such a bound requires that the true variance be bounded, which is guaranteed by Assumption D; the proof is given in Online Appendix A of the Supplementary Material \[18\]. Consequently, the condition

\[ \frac{E_{\zeta_i} \left[ \| \nabla \text{FD}_F F\{\zeta_i\}(x_k) - \nabla \text{FD}_F F(x_k) \|^2 \right]}{|S_k|} \leq \theta^2 \| \nabla \text{FD}_F F(x_k) \|^2 \tag{14} \]

is sufficient for ensuring that (12) holds. The condition (14) involves the true expected gradient and variance, but these can be approximated with sample gradient and sample variance estimates, respectively, yielding the practical finite-difference norm test

\[ \frac{\text{Var}_{\zeta_i \in S_k^v} \left[ \nabla \text{FD}_F F\{\zeta_i\}(x_k) \right]}{|S_k|} \leq \theta^2 \| \nabla \text{FD}_F S_k(x_k) \|^2, \tag{Norm} \]

where \(S_k^v \subseteq S_k\) is a subset of the current sample with \(|S_k^v| > 1\) and the variance term is defined as

\[ \text{Var}_{\zeta_i \in S_k^v} \left[ \nabla \text{FD}_F F\{\zeta_i\}(x_k) \right] := \frac{1}{|S_k^v| - 1} \sum_{\zeta_i \in S_k^v} \left\| \nabla \text{FD}_F F\{\zeta_i\}(x_k) - \nabla \text{FD}_F S_k(x_k) \right\|^2. \]

In our algorithm, we test condition (Norm); and whenever it is not satisfied, we increase \(|S_k|\) until (Norm) is satisfied. Section 2.5 provides a complete description of the algorithm and addresses the complications that arise because of the relevant approximations.

### 2.1.2 Inner product quasi-Newton test

The norm condition (Norm) controls the variance in the gradient estimation but does not utilize observed quasi-Newton information to control the sample sizes. Bollapragada et al. \[17\] proposed to control the sample sizes used in the gradient estimation.
by ensuring that the stochastic quasi-Newton directions make an acute angle with the true quasi-Newton direction with high probability. That is,

$$\left( H_k \nabla^\text{FD} F_{S_k}(x_k) \right)^T H_k \nabla F(x_k) > 0$$ \hspace{1cm} (15)

holds with high probability. However, one cannot always satisfy this condition, even in expectation, because of the inherent bias in the gradient estimator. We observe that the left-hand side of (15) is

$$\left( H_k \nabla^\text{FD} F_{S_k}(x_k) \right)^T H_k \nabla^\text{FD} F(x_k) + \left( H_k \nabla^\text{FD} F_{S_k}(x_k) \right)^T \left( H_k \nabla F(x_k) - H_k \nabla^\text{FD} F(x_k) \right),$$ \hspace{1cm} (16)

and, taking an expectation, we obtain

$$\mathbb{E}_{S_k} \left[ \left( H_k \nabla^\text{FD} F_{S_k}(x_k) \right)^T H_k \nabla F(x_k) \right] \leq \| H_k \nabla^\text{FD} F(x_k) \|^2 + \left( H_k \nabla^\text{FD} F(x_k) \right)^T \left( H_k \nabla F(x_k) - H_k \nabla^\text{FD} F(x_k) \right)$$

$$\geq \| H_k \nabla^\text{FD} F(x_k) \|^2 - \| H_k \nabla^\text{FD} F(x_k) \| \| H_k \nabla F(x_k) - H_k \nabla^\text{FD} F(x_k) \|$$

$$\geq \| H_k \nabla^\text{FD} F(x_k) \| \left( \| H_k \nabla F(x_k) \| - 2 \| H_k \| \| \nabla F(x_k) - \nabla^\text{FD} F(x_k) \| \right)$$

$$\geq \| H_k \nabla^\text{FD} F(x_k) \| \left( \| H_k \nabla F(x_k) \| - 2 \| H_k \| \| \nabla F(x_k) - \nabla^\text{FD} F(x_k) \| \right)$$

$$\geq \| H_k \nabla^\text{FD} F(x_k) \| \left( \| H_k \nabla F(x_k) \| - \| H_k \| L_{\nabla^\text{FD} F} \sqrt{d} \right),$$

where the second inequality is due to the fact that $\| a \| \geq \| b \| - \| a - b \|$ and the last inequality is due to (7).

When $x_k$ is nearly stationary in the sense that $\| \nabla F(x_k) \| < \frac{\lambda_{\text{max}}(H_k)L_{\nabla^\text{FD} F} \sqrt{d}}{\lambda_{\text{min}}(H_k)}$, where $\lambda_{\text{max}}(H_k)$ and $\lambda_{\text{min}}(H_k) > 0$ are the largest and smallest eigenvalues of $H_k$, respectively, it is not guaranteed that the inequality in (15) can be satisfied in expectation since the term, $\| H_k \nabla F(x_k) \| - \| H_k \| L_{\nabla^\text{FD} F} \sqrt{d}$, could be negative for such near stationary points. Moreover, in the derivative-free setting we do not have access to direct estimates of $\nabla F(x_k)$ to control the quantity (15). Therefore, we propose to consider only the first term in (16)—the inner product between the finite-difference stochastic quasi-Newton direction and the true finite-difference quasi-Newton direction—to control the sample sizes. We ensure that this quantity is close to its expected value by controlling the variance in this quantity. That is, the condition is given by

$$\mathbb{E}_{S_k} \left[ \left( \left( H_k \nabla^\text{FD} F_{S_k}(x_k) \right)^T H_k \nabla^\text{FD} F(x_k) - \| H_k \nabla^\text{FD} F(x_k) \|^2 \right) \right]$$

$$\leq \theta^2 \| H_k \nabla^\text{FD} F(x_k) \|^4,$$ \hspace{1cm} (17)
where $\mathbb{E}_{S_k} \left[ H_k \nabla^{FD} F_{S_k}(x_k) \right] = H_k \nabla^{FD} F(x_k)$ by Assumption A. The left-hand side of (17) can be bounded by the true variance as done above; the proof that the true variance is bounded is given in Online Appendix A of the Supplementary Material [18]. Therefore, for ensuring (17), it is sufficient for

$$
\frac{1}{|S_k|} \mathbb{E}_{\zeta_i} \left[ \left( \left( H_k \nabla^{FD} F_{\{\zeta_i\}}(x_k) \right)^T H_k \nabla^{FD} F_{S_k}(x_k) - \left\| H_k \nabla^{FD} F(x_k) \right\|^2 \right)^2 \right]
$$

to be bounded by the right-hand side of (17). Approximating the true expected gradient and variance with sample gradient and variance estimates results in the practical finite-difference inner product quasi-Newton test

$$
\text{Var}_{\zeta_i \in S_k^v} \left[ \left( H_k \nabla^{FD} F_{\{\zeta_i\}}(x_k) \right)^T H_k \nabla^{FD} F_{S_k}(x_k) \right] \leq \theta^2 \left\| H_k \nabla^{FD} F_{S_k}(x_k) \right\|^4, \quad \text{(IPQN)}
$$

where $S_k^v \subseteq S_k$ is a subset of the current sample with $|S_k^v| > 1$ and the variance term is defined as

$$
\text{Var}_{\zeta_i \in S_k^v} \left[ \left( H_k \nabla^{FD} F_{\{\zeta_i\}}(x_k) \right)^T H_k \nabla^{FD} F_{S_k}(x_k) \right] := \frac{1}{|S_k^v| - 1} \sum_{\zeta_i \in S_k^v} \left( \left( H_k \nabla^{FD} F_{S_k}(x_k) \right)^T H_k \nabla^{FD} F_{\{\zeta_i\}}(x_k) - \left\| H_k \nabla^{FD} F_{S_k} \right\|^2 \right)^2.
$$

This variance computation requires only one additional Hessian-vector product (i.e., the product of $H_k$ with $H_k \nabla^{FD} F_{S_k}(x_k)$). In our algorithm we test the condition (IPQN); whenever it is not satisfied, we increase $|S_k|$ until the condition is satisfied.

### 2.2 Finite-difference parameter selection

The finite-difference parameter $\nu > 0$ plays a significant role in the performance of optimization methods. Here we select the parameter by minimizing an upper bound on the gradient estimation error

$$
\nabla^{FD} F_{S_k}(x_k) - \nabla F(x_k) = \underbrace{\nabla^{FD} F_{S_k}(x_k) - \nabla F_{S_k}(x_k)}_{\text{Term 1}} + \underbrace{\nabla F_{S_k}(x_k) - \nabla F(x_k)}_{\text{Term 2}}. \quad (18)
$$

We observe that Term 2 in (18) is independent of the parameter $\nu$. Using Assumption C on the subsampled functions, we can bound Term 1 by

$$
\left\| \nabla^{FD} F_{S_k}(x_k) - \nabla F_{S_k}(x_k) \right\|^2 = \sum_{j=1}^{d} \left( \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \left( f(x_k + \nu e_j, \zeta_i) - f(x_k, \zeta_i) - [\nabla x f(x_k, \zeta_i)]_j \right)^2 \right)
$$
which decreases as $\nu$ decreases. In any practical implementation, however, one has to account also for the numerical errors associated with the numerical evaluation of the stochastic function values. We employ the following assumption on a uniform bound for these errors.

**Assumption E** The function values $f(x, \xi)$ in (1) are corrupted by numerical noise $\epsilon(x, \xi)$ (i.e., the corrupted values $f(x, \xi) + \epsilon(x, \xi)$ are provided to the algorithm rather than the true stochastic values $f(x, \xi)$) uniformly bounded by $\epsilon_m > 0$; that is,

$$|\epsilon(x, \xi)| \leq \epsilon_m \quad \text{for all } x, \xi.$$

Applying Assumption E, we get the corrupted gradient estimator

$$\nabla^{\text{FD}} \hat{F}_{S_k}(x_k) := \frac{1}{|S_k|} \sum_{\xi_i \in S_k} \left[ f(x + \nu e_j, \xi_i) + \epsilon(x + \nu e_j, \xi_i) - f(x, \xi_i) - \epsilon(x, \xi_i) \right]_{j=1}^d,$$

and hence

$$\|\nabla^{\text{FD}} \hat{F}_{S_k}(x_k) - \nabla^{\text{FD}} F_{S_k}(x_k)\| \leq \frac{2 \epsilon_m \sqrt{d}}{\nu}.$$  

Combining this with (18) and minimizing the resulting upper bound, we get the parameter value

$$\nu^* := 2 \frac{\epsilon_m}{L\nabla f}.$$

This optimal finite-difference parameter is analogous to the one derived in [50], which depends on the variance in stochastic models of the numerical noise. We note that because we assume that one can employ CRNs in the stochastic function estimations, this leads to lower variance in the gradient estimators and makes the parameter selection independent of the variance from the random variable $\xi$. In Sect. 5 we forgo estimation of Lipschitz constants and simply employ $\nu \approx \sqrt{\epsilon_m}$.

### 2.3 Step-length Selection

We employ a stochastic line search to choose the step length $\alpha_k$ in (8) by using a sufficient decrease condition on the subsampled function. In particular, we would like
α_k to satisfy
\[ F_{S_k} \left( x_k - \alpha_k H_k \nabla^{FD} F_{S_k}(x_k) \right) \]
\[ \leq F_{S_k}(x_k) - c_1 \alpha_k \left( \nabla^{FD} F_{S_k}(x_k) \right)^T H_k \nabla^{FD} F_{S_k}(x_k) + c_2, \tag{22} \]

where \( c_1 \in (0, 0.5) \) and \( c_2 > 0 \) are user-specified parameters. We employ a backtracking procedure wherein a trial step length \( \alpha_k \) that does not satisfy (22) is reduced by a fixed fraction \( \tau < 1 \) (i.e., \( \alpha_k \to \tau \alpha_k \)). In Theorem 2, we establish that there exists a nontrivial interval for \( \alpha_k \) such that the condition (22) is always satisfied.

**Theorem 2** If Assumption C is satisfied, \( c_1 \in (0, 0.5) \), \( c_2 > 0 \), and \( \lambda_{\min}(H_k) > 0 \), then there exists a nontrivial interval such that (22) is satisfied. Specifically, (22) holds for any

\[ \alpha_k \in \left( 0, \min \left\{ \frac{1 - 2c_1}{L \nabla f \lambda_{\max}(H_k)}, \frac{8c_2}{\lambda_{\max}(H_k) L \nabla f \nu^2 d} \right\} \right). \tag{23} \]

**Proof** We first note from (23) that

\[ \alpha_k \leq \frac{1 - 2c_1}{L \nabla f \lambda_{\max}(H_k)} \leq \frac{1}{L \nabla f \lambda_{\min}(H_k)}, \]

since \( c_1 > 0 \) and \( \lambda_{\max}(H_k) \geq \lambda_{\min}(H_k) > 0 \). By using this inequality and Lemma 1 applied to \( F_{S_k} \) (a consequence of Assumption C), we have that

\[ F_{S_k} \left( x_k - \alpha_k H_k \nabla^{FD} F_{S_k}(x_k) \right) \]
\[ \leq F_{S_k}(x_k) - \alpha_k \nabla F_{S_k}(x_k)^T H_k \nabla^{FD} F_{S_k}(x_k) + \frac{L \nabla f \alpha_k^2}{2} \| H_k \nabla^{FD} F_{S_k}(x_k) \|^2 \]
\[ = F_{S_k}(x_k) - \alpha_k \nabla F_{S_k}(x_k)^T H_k \nabla^{FD} F_{S_k}(x_k) \]
\[ + \alpha_k \left( \nabla^{FD} F_{S_k}(x_k) - \nabla F_{S_k}(x_k) \right)^T H_k \nabla^{FD} F_{S_k}(x_k) + \frac{L \nabla f \alpha_k^2}{2} \| H_k \nabla^{FD} F_{S_k}(x_k) \|^2 \]
\[ \leq F_{S_k}(x_k) - \alpha_k \nabla F_{S_k}(x_k)^T H_k \nabla^{FD} F_{S_k}(x_k) + \frac{\alpha_k}{2} \nabla^{FD} F_{S_k}(x_k)^T H_k (\nabla^{FD} F_{S_k}(x_k) - \nabla F_{S_k}(x_k)) \]
\[ + \frac{L \nabla f \alpha_k^2}{2} \| H_k \nabla^{FD} F_{S_k}(x_k) \|^2 \]
\[ = F_{S_k}(x_k) - \alpha_k \nabla^{FD} F_{S_k}(x_k)^T H_k \left( I - L \nabla f \alpha_k H_k \right) H_k^{1/2} \nabla^{FD} F_{S_k}(x_k) \]
\[ + \frac{\alpha_k}{2} (\nabla^{FD} F_{S_k}(x_k) - \nabla F_{S_k}(x_k))^T H_k (\nabla^{FD} F_{S_k}(x_k) - \nabla F_{S_k}(x_k)) \]
\[ \leq F_{S_k}(x_k) - \alpha_k \left( 1 - \alpha_k L \nabla f \lambda_{\max}(H_k) \right) \nabla^{FD} F_{S_k}(x_k)^T H_k \nabla^{FD} F_{S_k}(x_k) \]

\[ \geq \frac{1}{2} \]
\[ + \frac{\alpha_k \lambda_{\text{max}}(H_k)}{2} \| \nabla^{FD} F_{S_k}(x_k) - \nabla F_{S_k}(x_k) \|^2 \]
\[ \leq F_{S_k}(x_k) - \alpha_k \left( 1 - \alpha_k L \sqrt{f} \lambda_{\text{max}}(H_k) \right) \nabla^{FD} F_{S_k}(x_k)^T H_k \nabla^{FD} F_{S_k}(x_k) \]
\[ + \frac{\alpha_k \lambda_{\text{max}}(H_k) L \sqrt{f} v^2 d}{8} \]
\[ \leq F_{S_k}(x_k) - c_1 \alpha_k (\nabla^{FD} F_{S_k}(x_k))^T H_k \nabla^{FD} F_{S_k}(x_k) + c_2, \]

where the second inequality is because \( H_k \) is positive definite and because, for any positive-definite matrix \( A \), \( x^T A y \leq \frac{x^T A x + y^T A y}{2} \); the fourth inequality is due to (19) (Assumption C); and the last inequality is due to (23). \( \square \)

We also note that because of the stochasticity in the function values, it is not guaranteed that a decrease in stochastic function realizations \( f \) can ensure decrease in the expected function \( F \). A conservative strategy to address this issue is to choose the initial trial step length to be small enough to control the potential increase in \( F \) values when the stochastic estimations are not good. Bollapragada et al. [17] proposed a heuristic to choose the initial trial estimate for \( \alpha_k \) such that there is a decrease in the expected function value. Following a similar strategy, we derive a heuristic to choose the initial trial step length as

\[ \hat{\alpha}_k = \left( 1 + \frac{\text{Var}_{\zeta_i \in S_k} \left[ \nabla^{FD} F_{\{\zeta_i\}}(x_k) \right]}{|S_k| \| \nabla^{FD} F_{S_k}(x_k) \|^2} \right)^{-1}. \tag{24} \]

The formal reasoning for this choice is provided in Online Appendix A.2 which uses Assumption (G) introduced in Sect. 3 of the Supplementary Material [18].

### 2.4 Stable quasi-Newton update

In the BFGS and L-BFGS methods, the inverse Hessian approximation is updated by using the formulae

\[ H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T, \quad \rho_k = \left( y_k^T s_k \right)^{-1}, \quad V_k = I - \rho_k y_k s_k^T, \]

where \( s_k = x_{k+1} - x_k \) and \( y_k \) is the difference in the gradients at \( x_{k+1} \) and \( x_k \). In stochastic settings, \( y_k \) is typically defined as the difference in gradients measured on the same sample \( S_k \) to ensure stability in the quasi-Newton approximation [17]. We follow the same approach and define

\[ y_k := \nabla^{FD} F_{S_k}(x_{k+1}) - \nabla^{FD} F_{S_k}(x_k). \tag{25} \]

However, even though computing gradient differences on common sample sets can improve stability, the curvature pair \((y_k, s_k)\) still may not satisfy the condition \( y_k^T s_k > 0 \)
required to ensure positive definiteness of the quasi-Newton matrix $H_k$. In particular, for any $\mu$-strongly convex function $F_{S_k}$, we have that

$$y_k^T s_k = \left( \nabla^\text{FD} F_{S_k}(x_{k+1}) - \nabla^\text{FD} F_{S_k}(x_k) \right)^T s_k$$

$$= \left( \nabla F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_k) \right)^T s_k$$

$$+ \left( \nabla^\text{FD} F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_{k+1}) + \nabla F_{S_k}(x_k) - \nabla^\text{FD} F_{S_k}(x_k) \right)^T s_k$$

$$\geq \mu \| s_k \|^2$$

$$- \left( \| \nabla^\text{FD} F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_{k+1}) \| + \| \nabla F_{S_k}(x_k) - \nabla^\text{FD} F_{S_k}(x_k) \| \right) \| s_k \|$$

$$\geq \mu \| s_k \|^2 - L_{\nabla f} \sqrt{d} \| s_k \| = \| s_k \| \left( \mu \| s_k \| - L_{\nabla f} \sqrt{d} \right),$$

where the first inequality is due to strong convexity and the last inequality is due to (19) (by Assumption C). Therefore, the condition $y_k^T s_k > 0$ is guaranteed to be satisfied when $\| s_k \| > \frac{L_{\nabla f} \sqrt{d}}{\mu}$. Recently, Xie et al. [65] proposed modifying the curvature pair update whenever the step $s_k$ is too small so that $y_k^T s_k > 0$. However, this modification requires knowledge of some unknown problem parameters and may not provide guarantees in the case when $F_{S_k}$ is nonconvex. Therefore, we skip the quasi-Newton update if the following curvature condition is not satisfied:

$$y_k^T s_k > \beta_1 \| s_k \|^2,$$  \hspace{1cm} (26)

where $\beta_1 > 0$ is a predetermined constant.

Moreover, to ensure that the eigenvalues of the quasi-Newton matrix are bounded, we require the ratio $\frac{y_k^T y_k}{y_k^T s_k}$ to be bounded. We note, however, that this requirement may not always be possible to satisfy because of the presence of the bias term. That is,

$$\frac{y_k^T y_k}{y_k^T s_k} = \frac{\| \nabla^\text{FD} F_{S_k}(x_{k+1}) - \nabla^\text{FD} F_{S_k}(x_k) \|^2}{y_k^T s_k}$$

$$\leq 3 \frac{\| \nabla F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_k) \|^2}{\beta_1 \| s_k \|^2} + 3 \frac{\| \nabla^\text{FD} F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_{k+1}) \|^2}{\beta_1 \| s_k \|^2}$$

$$+ 3 \frac{\| \nabla F_{S_k}(x_k) - \nabla F_{S_k}(x_k) \|^2}{\beta_1 \| s_k \|^2}$$

$$\leq \frac{3L_{\nabla f}^2}{\beta_1} + \frac{3L_{\nabla f}^2 \nu^2 d}{2\beta_1 \| s_k \|^2},$$  \hspace{1cm} (27)

where the first inequality is due to the fact that $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ and (26) and the last inequality is due to Assumption C and (19). Therefore, for $\| s_k \|$ arbitrarily close to zero, this fraction may not be bounded. Thus, to ensure the eigenvalues are bounded, we skip the update whenever $\| s_k \|$ is too small. That is, we skip the update
whenever the following lengthening condition is not satisfied:

\[ \|s_k\| > \beta_2 > 0, \]  

(28)

where \( \beta_2 > 0 \) is a small predetermined constant.

### 2.5 The complete algorithm

We use L-BFGS as the method for incorporating quasi-Newton information and the search direction is computed using a two-loop recursion procedure [52], which has a computational cost of \( \mathcal{O}(md) \) where \( m \) represents the number of curvature pairs in the memory. Note that \( m \) is typically chosen to be less than \( d \) and so the overall computational cost is typically linear in terms of the dimension of the problem for large-scale problems. The pseudocode of the resulting finite-difference stochastic L-BFGS method is given in Algorithm 1. We summarize the assumptions on the algorithmic parameters in Assumption F. The initial Hessian matrix \( H_0^k \) in the L-BFGS recursion at each iteration is chosen as \( \kappa_k I \), where \( \kappa_k = \frac{y^T_k s_k}{y^T_k y_k} \).

**Assumption F** The algorithmic parameters satisfy \( \tau \in (0, 1) \), \( c_1 \in (0, 0.5) \), \( c_2 > 0 \), \( \theta_0 > 0 \), \( \gamma < 1 \), \( m \in \mathbb{Z}_{++} \), \( |S_0| \in \mathbb{Z}_{++} \), \( \beta_1 > 0 \), and \( \beta_2 > 0 \).

In the sampling tests, we employ sample approximations to compute the sample size. These sample estimates are sufficiently accurate except if the sample size is too small. To avoid the scenario of not increasing the sample sizes at all, we employ the following strategy. Instead of choosing the parameter \( \theta \) to be a fixed parameter, we make it iteration dependent and control it adaptively.

The parameter \( \theta \) controls the probability of satisfying the underlying deterministic condition. For example, in the inner product quasi-Newton test, \( \theta \) controls the probability of generating a quasi-Newton direction that makes an acute angle with the true quasi-Newton direction. Smaller \( \theta \) values increase the probability of satisfying the underlying conditions and promote large sample sizes. Motivated by this property, we propose to increase the probability of satisfying the deterministic conditions when the approximations are not reliable. Although it is hard to identify whether the approximations are accurate or not solely based on sample sizes, we can monitor the potential ill effects of such scenarios. In particular, whenever the sample sizes remain constant, it is either because the current sample size is large enough to satisfy the true condition or because the approximations are not accurate. Therefore, in this scenario we decrease the \( \theta \) value in the next iteration. If the sample size has increased in the next iteration, we reset the value to its default value \( \theta_0 \). Otherwise, we continue to decrease its value until the sample sizes are increased. More precisely, at each iteration \( k \) we set \( \theta_k = \theta_{k-1} \gamma \) if \( |S_k| = |S_{k-1}| \), where \( \gamma < 1 \); otherwise we reset its value to a default value \( \theta_0 \).
Algorithm 1 Finite-Difference Stochastic L-BFGS Method

**Input:** Initial iterate $x_0$, initial sample size $|S_0|$, L-BFGS memory $m$, finite-difference parameter $\nu$

line search parameters $(c_1, c_2, \tau)$, sample test parameters $(\theta_0, \gamma)$, quasi-Newton safeguarding parameters $(\beta_1, \beta_2)$, selection of a (Norm) or (IPQN) test

**Initialization:** Set $k \leftarrow 1; \theta = \theta_0$

**Repeat** until convergence:

1: Choose a set $S_k$ consisting of $|S_k|$ i.i.d. realizations of $\zeta$

2: if (Norm) test selected and (Norm) is not satisfied then

3: $|S_k| \leftarrow$ least $|S_k|$ such that the inequality in (Norm) is satisfied

4: else if (IPQN) test selected and (IPQN) is not satisfied then

5: $|S_k| \leftarrow$ least $|S_k|$ such that the inequality in (IPQN) is satisfied

6: end if

7: if $|S_k| = |S_{k-1}|$ then

8: Set $\theta \leftarrow \theta_0$

9: else

10: Set $\theta \leftarrow \theta_0$

11: end if

12: Compute $\nabla_{\text{FD}} F_{S_k}(x_k)$

13: Compute $p_k = -H_k \nabla_{\text{FD}} F_{S_k}(x_k)$ using L-BFGS two-loop recursion in [52]

14: Compute $\alpha_k$ using (24)

15: while Armijo condition (22) is not satisfied do

16: Set $\alpha_k \leftarrow \alpha_k \tau$

17: end while

18: Compute $x_{k+1} = x_k + \alpha_k p_k$

19: Compute $y_k$ using (25) and set $s_k = x_{k+1} - x_k$

20: if $y_k^T s_k > \beta_1 \|s_k\|^2$ and $\|s_k\| > \beta_2$ then

21: if number of stored $(y_j, s_j)$ exceeds $m$ then

22: Discard oldest curvature pair $(y_j, s_j)$

23: end if

24: Store new curvature pair $(y_k, s_k)$

25: end if

26: Set $k \leftarrow k + 1$

27: Set $|S_k| = |S_{k-1}|$

3 Analysis of Algorithm 1

We now establish convergence results for the finite-difference quasi-Newton methods with the norm test and inner product quasi-Newton test. We make use of the following additional assumption for the analysis.

**Assumption G** For all $k$, the eigenvalues of $H_k$ are contained in an interval in $\mathbb{R}_{++}$; that is, there exist constants $\Lambda_2 \geq \Lambda_1 > 0$ such that

$$\Lambda_1 I \preceq H_k \preceq \Lambda_2 I, \quad \forall k.$$  

Assumption G can be shown to hold for both convex and nonconvex twice-differentiable functions $F$ by updating $H_k$ only when $y_k^T s_k \geq \beta_1 \|s_k\|^2$, where $\beta_1 > 0$ is a predetermined constant [8]. We provide the proof for the sake of completeness in Online Appendix A.3 of the Supplementary Material [18]. We note that as a consequence of this assumption, the analysis provided here is more general and can be used for a method with any positive-definite matrix $H_k$.  

\[ \text{Springer} \]
We now establish technical lemmas for both the norm and the inner product quasi-Newton tests.

### 3.1 Norm test

We begin in Lemma 3 by establishing a descent result for cases where the sample size $|S_k|$ satisfies the norm test.

**Lemma 3** For any $x_0$, let $\{x_k : k \in \mathbb{Z}_{++}\}$ be generated by iteration (8) with $|S_k|$ chosen by the (exact variance) finite-difference norm test (13) for a given constant $\theta > 0$, and suppose that Assumptions A, B, and G hold. Then, for any $k$ where $\alpha_k$ satisfies

$$0 < \alpha_k \leq \frac{\Lambda_1}{4 \left(1 + \theta^2\right) L \nabla F \Lambda_2^2},$$

we have that

$$\mathbb{E}_{S_k} \left[F(x_{k+1})\right] \leq F(x_k) - \frac{\alpha_k \Lambda_1}{4} \| \nabla F(x_k) \|^2 + \frac{\alpha_k (\Lambda_1 + 2 \Lambda_2)}{4} \|\nabla F(x_k) - \nabla F(x_k)\|^2.$$  

**Proof** By Assumption B and Lemma 1, we have that

$$\mathbb{E}_{S_k} \left[F(x_{k+1})\right] \leq F(x_k) - \mathbb{E}_{S_k} \left[ \alpha_k \left( H_k \nabla^F F_{S_k}(x_k) \right)^T \nabla F(x_k) \right]$$

$$+ \mathbb{E}_{S_k} \left[ \frac{L \nabla F \alpha_k^2}{2} \| H_k \nabla^F F_{S_k}(x_k) \|^2 \right]$$

$$= F(x_k) - \alpha_k \nabla^F F(x_k)^T H_k \nabla F(x_k)$$

$$+ \frac{L \nabla F \alpha_k^2}{2} \mathbb{E}_{S_k} \left[ \| H_k \nabla^F F_{S_k}(x_k) \|^2 \right],$$

where the equality follows from Assumption A. Defining

$$\delta_k := \nabla^F F(x_k) - \nabla F(x_k)$$

$$T_k := \frac{L \nabla F \alpha_k^2}{2} \mathbb{E}_{S_k} \left[ \| H_k \nabla^F F_{S_k}(x_k) \|^2 \right],$$

we have that

$$\mathbb{E}_{S_k} \left[F(x_{k+1})\right] \leq F(x_k) - \alpha_k \left( \nabla F(x_k) + \delta_k \right)^T H_k \nabla F(x_k) + T_k$$

$$= F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) - \alpha_k \delta_k^T H_k \nabla F(x_k) + T_k$$
\[ F(x_k) - \frac{\alpha_k}{2} \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\alpha_k}{2} \delta_k^T H_k \delta_k + T_k \]

where the second inequality is obtained by using the fact that \(2|\mathbf{x}^T \mathbf{y}| \leq \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{y}^T \mathbf{A} \mathbf{y}\) for any positive-definite matrix \(\mathbf{A}\).

Now, using (12) and Assumption G, we have that

\[
\mathbb{E}_{S_k} \left[ \| H_k \nabla F_{S_k}(x_k) \|^2 \right] \\
= \mathbb{E}_{S_k} \left[ \| H_k \left( \nabla F_{S_k}(x_k) - \nabla F(x_k) \right) \|^2 \right] + \| H_k \nabla F(x_k) \|^2 \\
\leq \Lambda_2^2 \mathbb{E}_{S_k} \left[ \| \nabla F_{S_k}(x_k) - \nabla F(x_k) \|^2 \right] + \Lambda_2^2 \| \nabla F(x_k) \|^2 \\
\leq \Lambda_2^2 \| \nabla F(x_k) \|^2 \\
\leq 2 \Lambda_2^2 (1 + \theta^2) \left( \| \nabla F(x_k) - \nabla F(x_k) \|^2 + \| \nabla F(x_k) \|^2 \right) \\
= 2 \Lambda_2^2 (1 + \theta^2) \| \delta_k \|^2 + 2 \Lambda_2^2 (1 + \theta^2) \| \nabla F(x_k) \|^2.
\]

Substituting this into \(T_k\) in (32) and using (29) and Assumption G, we obtain

\[
\mathbb{E}_{S_k} \left[ F(x_{k+1}) \right] \leq F(x_k) - \frac{\alpha_k}{2} \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\alpha_k}{2} \delta_k^T H_k \delta_k \\
+ L \nabla F_{\alpha_k} (1 + \theta^2) \| \delta_k \|^2 + L \nabla F_{\alpha_k} \Lambda_2^2 (1 + \theta^2) \| \nabla F(x_k) \|^2 \\
\leq F(x_k) - \frac{\alpha_k \Lambda_1}{2} \| \nabla F(x_k) \|^2 + \frac{\alpha_k \Lambda_2}{2} \| \delta_k \|^2 \\
+ L \nabla F_{\alpha_k} \Lambda_2^2 (1 + \theta^2) \| \delta_k \|^2 + L \nabla F_{\alpha_k} \Lambda_2^2 (1 + \theta^2) \| \nabla F(x_k) \|^2 \\
\leq F(x_k) - \frac{\alpha_k \Lambda_1}{2} \| \nabla F(x_k) \|^2 + \frac{\alpha_k \Lambda_2}{2} \| \delta_k \|^2 \\
+ \frac{\alpha_k \Lambda_1}{4} \| \delta_k \|^2 + \frac{\alpha_k \Lambda_1}{4} \| \nabla F(x_k) \|^2 \\
= F(x_k) - \frac{\alpha_k \Lambda_1}{4} \| \nabla F(x_k) \|^2 + \frac{\alpha_k (\Lambda_1 + 2 \Lambda_2)}{4} \| \delta_k \|^2,
\]

which establishes (30).

\[ \square \]

### 3.2 Inner product quasi-Newton test

We now consider the case where the sample size \(|S_k|\) satisfies the inner product quasi-Newton test. Following the strategy provided in [17], we assume that the orthogonality condition is satisfied by the stochastic finite-difference quasi-Newton directions.
Assumption H For the norm of the component of the stochastic direction orthogonal to the true direction,

\[ U_{i,k} := \left\| H_k \nabla F_{(\xi_i)}(x_k) - \frac{(H_k \nabla F_{(\xi_i)}(x_k))^T (H_k \nabla F(x_k))}{\| H_k \nabla F(x_k) \|^2} H_k \nabla F(x_k) \right\|^2, \]

there exists \( \psi > 0 \) such that

\[ \mathbb{E}_{\xi_i} \left[ \frac{U_{i,k}}{|S_k|} \right] \leq \psi^2 \left\| H_k \nabla F(x_k) \right\|^2 \forall k. \]

Using the proof techniques in [17, Lemma 1], we thus have the following bound on the length of the search direction:

\[ \mathbb{E}_{S_k} \left[ \left\| H_k \nabla F_{S_k}(x_k) \right\|^2 \right] \leq (1 + \theta^2 + \psi^2) \left\| H_k \nabla F(x_k) \right\|^2. \] (33)

Using this bound, we first establish a technical lemma.

Lemma 4 For any \( x_0 \), let \( \{ x_k : k \in \mathbb{Z}_{++} \} \) be generated by iteration (8) with \( |S_k| \) chosen by the (exact variance) finite-difference inner product quasi-Newton test (17), and suppose that Assumptions A, B, G, and H hold. Then, for any \( k \) where \( \alpha_k \) satisfies

\[ 0 < \alpha_k < \frac{1}{(1 + \theta^2 + \psi^2) L \nabla F \Lambda_2}, \] (34)

we have that

\[ \mathbb{E}_{S_k} \left[ F(x_{k+1}) \right] \leq F(x_k) - \frac{\alpha_k \Lambda_1}{2} \left\| \nabla F(x_k) \right\|^2 + \frac{\alpha_k \Lambda_2}{2} \left\| \nabla F(x_k) - \nabla F(x_k) \right\|^2. \] (35)

Proof By Assumptions A, B and H, and Lemma 1, we have that

\[
\mathbb{E}_{S_k} \left[ F(x_{k+1}) \right] \\
\leq F(x_k) - \mathbb{E}_{S_k} \left[ \alpha_k \left( H_k \nabla F_{S_k}(x_k) \right)^T \nabla F(x_k) \right] \\
+ \mathbb{E}_{S_k} \left[ \frac{L \nabla F \alpha_k^2}{2} \left\| H_k \nabla F_{S_k}(x_k) \right\|^2 \right] \\
= F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{L \nabla F \alpha_k^2}{2} \mathbb{E}_{S_k} \left[ \left\| H_k \nabla F_{S_k}(x_k) \right\|^2 \right] \\
\leq F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{L \nabla F \alpha_k^2}{2} \mathbb{E}_{S_k} \left[ \left\| H_k \nabla F_{S_k}(x_k) \right\|^2 \right] \\
\leq F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{L \nabla F \alpha_k^2}{2} \left( 1 + \theta^2 + \psi^2 \right) \left\| H_k \nabla F(x_k) \right\|^2, \]
where the last inequality is due to Assumption H and (33).

By using $\delta_k$ from (31), $\tilde{L}_F \alpha := L_F(1 + \theta^2 + \psi^2)$, and Assumption G, we have that

$$
\mathbb{E}_{S_k} \left[ F(x_{k+1}) \right] 
\leq F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\tilde{L}_F \alpha_k^2}{2} \| H_k (\nabla F(x_k) + \delta_k) \|^2
$$

$$
= F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\tilde{L}_F \alpha_k^2}{2} \left( \| H_k \nabla F(x_k) \|^2 + \| H_k \delta_k \|^2 \right)
$$

$$
- \alpha_k (H_k^{1/2} \delta_k)^T (I - \tilde{L}_F \alpha_k H_k) (H_k^{1/2} \nabla F(x_k))
$$

$$
\leq F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\tilde{L}_F \alpha_k^2}{2} \left( \| H_k \nabla F(x_k) \|^2 + \| H_k \delta_k \|^2 \right)
$$

$$
+ \frac{\alpha_k}{2} \left( H_k^{1/2} \nabla F(x_k) \right)^T (I - \tilde{L}_F \alpha_k H_k) \left( H_k^{1/2} \nabla F(x_k) \right)
$$

$$
+ \frac{\alpha_k}{2} \left( H_k^{1/2} \delta_k \right)^T (I - \tilde{L}_F \alpha_k H_k) \left( H_k^{1/2} \delta_k \right)
$$

$$
= F(x_k) - \frac{\alpha_k}{2} \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\alpha_k}{2} \delta_k^T H_k \delta_k
$$

$$
\leq F(x_k) - \frac{\alpha_k \Lambda_1}{2} \| \nabla F(x_k) \|^2 + \frac{\alpha_k \Lambda_2}{2} \| \delta_k \|^2,
$$

where the second inequality is obtained by using the fact that $I - \tilde{L}_F \alpha_k H_k$ is a positive-definite matrix due to (34) and Assumption G, and $2|x^T A y| \leq x^T A x + y^T A y$ for any positive-definite matrix $A$, and the last inequality is due to Assumption G. Substituting $\delta_k$ with its definition in (31) completes the proof. \hfill \Box

### 3.3 Convergence results

We now show that the finite-difference stochastic quasi-Newton iteration (8) with a fixed step length $\alpha_k = \alpha$ is convergent to a neighborhood of a stationary point $x^*$ when the sample sizes $|S_k|$ satisfy either the norm test or the inner product quasi-Newton test.

Throughout this section we let $\mathbb{E} \cdot$ denote the total expectation with respect to all the random sets $\{S_k\}_k$. That is, $\mathbb{E} \cdot = \mathbb{E}_{T_k} [\mathbb{E} \cdot | T_k]$ where $T_k = \sigma(x_0, S_1, S_2, \ldots, S_{k-1})$ is the filtration defined in Sect. 1.2. More precisely, the total expectation of $F(x_k)$ and $\| \nabla F(x_k) \|^2$ are given as,

$$
\mathbb{E} [F(x_k)] = \mathbb{E}_{S_1} \left[ \mathbb{E}_{S_2} \left[ \cdots \mathbb{E}_{S_{k-1}} [F(x_k)] \right] \right]
$$

$$
\mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] = \mathbb{E}_{S_1} \left[ \mathbb{E}_{S_2} \left[ \cdots \mathbb{E}_{S_{k-1}} \left[ \| \nabla F(x_k) \|^2 \right] \right] \right]
$$
3.3.1 Strongly convex functions

We first consider strongly convex functions \( F \) with \( x^* \) denoting the unique minimizer of \( F \). This is formalized in the following assumption, which supposes that \( \nabla F \) exists (as it is the case under either Assumption B or Assumption C).

**Assumption I** There exists a parameter \( \mu > 0 \) such that

\[
\| \nabla F(x) \|^2 \geq 2\mu \left( F(x) - F(x^*) \right) \quad \forall x \in \mathbb{R}^d.
\]

We first establish a general lemma whose result can be used in proving convergence results for both the tests.

**Lemma 5** Suppose Assumption I is satisfied. For any \( x_0 \), let \( \{x_k : k \in \mathbb{Z}_{++}\} \) be generated by iteration (8), with \( |S_k| \) chosen such that

\[
\mathbb{E}_{S_k} \left[ F(x_{k+1}) \right] \leq F(x_k) - \frac{a_1}{2} \| \nabla F(x_k) \|^2 + a_2
\]

for some constants \( a_1 > 0 \) and \( a_2 > 0 \). Then,

\[
\mathbb{E} \left[ F(x_k) - F(x^*) \right] \leq (1 - \mu a_1)^k \left( F(x_0) - F(x^*) - \frac{a_2}{\mu a_1} \right) + \frac{a_2}{\mu a_1} \quad \forall k \in \mathbb{Z}_+.
\]

**Proof** Employing Assumption I at iteration \( k \), substituting into (36), and subtracting \( F(x^*) \) from both sides, we obtain

\[
\mathbb{E}_{S_k} \left[ F(x_{k+1}) - F(x^*) \right] \leq F(x_k) - F(x^*) - \mu a_1 (F(x_k) - F(x^*)) + a_2.
\]

Subtracting the constant \( \frac{a_2}{\mu a_1} \) from both sides and taking total expectation, we obtain

\[
\mathbb{E} \left[ F(x_{k+1}) - F(x^*) \right] - \frac{a_2}{\mu a_1} \leq (1 - \mu a_1) \mathbb{E} \left[ F(x_k) - F(x^*) \right] + a_2 - \frac{a_2}{\mu a_1} \]

\[= (1 - \mu a_1) \left( \mathbb{E} \left[ F(x_k) - F(x^*) \right] - \frac{a_2}{\mu a_1} \right). \tag{37} \]

The lemma follows by applying (37) repeatedly through iteration \( k \in \mathbb{Z}_+ \).

We can now apply this general lemma to show results for sample sizes \( |S_k| \) satisfying either the norm test (Theorem 6) or the inner product quasi-Newton test (Theorem 7). We note that in the remainder of this section we assume a constant step length, but this can readily be generalized as established in Online Appendix A.4 of the Supplementary Material [18].

**Theorem 6** (Norm Test) For any \( x_0 \), let \( \{x_k : k \in \mathbb{Z}_{++}\} \) be generated by iteration (8) with \( |S_k| \) chosen by the (exact variance) finite-difference norm test (12), and suppose
that Assumptions $A$, $B$, $G$, and $I$ hold. Then, if $\alpha_k = \alpha$ satisfies (29), we have that

$$
\mathbb{E}\left[F(x_k) - F(x^*)\right] \leq \left(1 - \frac{\mu \Lambda_1 \alpha}{2}\right)^k \left(F(x_0) - F(x^*)\right) + \frac{(\Lambda_1 + 2\Lambda_2)L_{\nabla F}^2 v^2 d}{8\mu \Lambda_1}.
$$

(38)

**Proof** Applying Lemma 3 and substituting (7) into (30), we obtain

$$
\mathbb{E}_{S_k}\left[F(x_{k+1})\right] \leq F(x_k) - \frac{\alpha \Lambda_1}{4} \|\nabla F(x_k)\|^2 + \frac{\alpha (\Lambda_1 + 2\Lambda_2)L_{\nabla F}^2 v^2 d}{16}.
$$

(39)

Applying Lemma 5 with constants $a_1 = \frac{\alpha \Lambda_1}{4}$ and $a_2 = \frac{\alpha (\Lambda_1 + 2\Lambda_2)L_{\nabla F}^2 v^2 d}{16}$ yields (38). $\square$

**Theorem 7** (Inner Product Quasi-Newton Test) For any $x_0$, let $\{x_k : k \in \mathbb{Z}^{++}\}$ be generated by iteration (8) with $|S_k|$ chosen by the (exact variance) finite-difference inner product quasi-Newton test (17), and suppose that the Assumptions $A$, $B$, $G$, $H$, and $I$ hold. Then, if $\alpha_k = \alpha$ satisfies (34) we have that

$$
\mathbb{E}\left[F(x_k) - F(x^*)\right] \leq (1 - \mu \Lambda_1 \alpha)^k \left(F(x_0) - F(x^*)\right) + \frac{\Lambda_2 L_{\nabla F}^2 v^2 d}{8\mu \Lambda_1}.
$$

**Proof** Applying Lemma 4 and substituting (7) into (35), we obtain

$$
\mathbb{E}_{S_k}\left[F(x_{k+1})\right] \leq F(x_k) - \frac{\alpha \Lambda_1}{2} \|\nabla F(x_k)\|^2 + \frac{\alpha \Lambda_2 L_{\nabla F}^2 v^2 d}{8}.
$$

(40)

Applying Lemma 5 with $a_1 = \alpha \Lambda_1$ and $a_2 = \frac{\alpha \Lambda_2 L_{\nabla F}^2 v^2 d}{8}$ completes the proof. $\square$

The above results are similar to the ones established for stochastic first-order optimization methods [20, Theorem 4.6]. However, the size of the neighborhood in [20, Theorem 4.6] depends on the stochasticity in the problem, whereas here it is independent of the stochasticity in the problem. This independence is due to the sample size selection procedure and is equivalent to the settings where the sample sizes are chosen to be sufficiently large (depending on the variance in the stochastic functions) to decrease the size of the neighborhood. That is, we achieve the above results by adaptively increasing the sample sizes instead of choosing the large sample size from the beginning, thereby reducing the computational efforts. Moreover, the neighborhood obtained here is similar to the one that is achieved for deterministic finite-difference gradient methods [9, Theorem 4.1] where $\bar{\epsilon}_g^2$ in those settings is equal to $\frac{L_{\nabla f}^2 v^2 d}{4}$ due to (7).

The above results also show the dependence of the convergence rate and size of the neighborhood on the condition number of the quasi-Newton matrix (i.e., $\frac{\Lambda_2}{\Lambda_1}$). This dependence on the condition number of the quasi-Newton matrix is because our
analysis is for a general positive-definite matrix and such dependencies exist in global convergence results of Newton and quasi-Newton methods; see, e.g., [8, 15, 56, 66]. Moreover, we only consider limited-memory quasi-Newton variants in our algorithm for which the best known worst-case theoretical convergence results are worse than steepest descent methods [52]. However, the practical performance of such variants is often superior to steepest descent [8].

3.3.2 Nonconvex functions

We now consider the case when $F$ is bounded below but not necessarily convex. In this setting, we replace Assumption I and Lemma 5 as follows.

**Assumption J** There exists a constant $F_{\text{min}}$ with $-\infty < F_{\text{min}} \leq F(x)$ $\forall x \in \mathbb{R}^d$.

**Lemma 8** Suppose Assumption J is satisfied. For any $x_0$, let $\{x_k : k \in \mathbb{Z}_{++}\}$ be generated by iteration (8) with $|S_k|$ chosen such that inequality (36) is satisfied with some constants $a_1, a_2 > 0$. Then, for any $T \in \mathbb{Z}_{++}$, we have that

$$
\min_{0 \leq k \leq T-1} \mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] \leq \frac{2}{Ta_1} (F(x_0) - F_{\text{min}}) + \frac{2a_2}{a_1}.
$$

**Proof** Taking total expectation in (36), we obtain

$$
\mathbb{E} \left[ F(x_{k+1}) \right] \leq \mathbb{E} \left[ F(x_k) \right] - \frac{a_1}{2} \mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] + a_2,
$$

and hence

$$
\mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] \leq \frac{2}{a_1} \mathbb{E} \left[ F(x_k) - F(x_{k+1}) \right] + \frac{2a_2}{a_1}.
$$

Summing both sides of this inequality from $k = 0$ to $T - 1$, and since $F$ is bounded below by $F_{\text{min}}$, we get

$$
\sum_{k=0}^{T-1} \mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] \leq \frac{2}{a_1} \mathbb{E} \left[ F(x_0) - F(x_T) \right] + T \frac{2a_2}{a_1} \leq \frac{2}{a_1} (F(x_0) - F_{\text{min}} + Ta_2).
$$

Therefore, we can conclude that

$$
\min_{0 \leq k \leq T-1} \mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] \leq \frac{1}{T} \sum_{k=0}^{T} \mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] \leq \frac{2}{Ta_1} (F(x_0) - F_{\text{min}}) + \frac{2a_2}{a_1}.
$$

We can now apply this general lemma to show results for sample sizes $|S_k|$ satisfying either the norm test (Theorem 9) or the inner product quasi-Newton test (Theorem 10).
Theorem 9 (Norm Test) For any $x_0$, let $\{x_k : k \in \mathbb{Z}_{++}\}$ be generated by iteration (8) with $|S_k|$ chosen by the (exact variance) finite-difference norm test (12), and suppose that Assumptions A, B, G, and J hold. Then, if $\alpha_k = \alpha$ satisfies (29), for any $T \in \mathbb{Z}_{++}$ we have that

$$
\min_{0 \leq k \leq T-1} \mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] \leq \frac{4}{\alpha T \Lambda_1} (F(x_0) - F_{\text{min}}) + \frac{(\Lambda_1 + 2 \Lambda_2) L_{\nabla F}^2 \nu^2 d}{4 \Lambda_1}.
$$

Proof Applying Lemma 3, from inequality (39) we have that

$$
\mathbb{E}_{S_k} \left[ F(x_{k+1}) \right] \leq F(x_k) - \frac{\alpha \Lambda_1}{4} \| \nabla F(x_k) \|^2 + \frac{\alpha (\Lambda_1 + 2 \Lambda_2) L_{\nabla F}^2 \nu^2 d}{16}.
$$

Applying Lemma 8 with constants $a_1 = \frac{\alpha \Lambda_1}{2}$ and $a_2 = \frac{\alpha (\Lambda_1 + 2 \Lambda_2) L_{\nabla F}^2 \nu^2 d}{16}$ completes the proof. $\square$

Theorem 10 (Inner Product Quasi-Newton Test) For any $x_0$, let $\{x_k : k \in \mathbb{Z}_{++}\}$ be generated by iteration (8) with $|S_k|$ chosen by the (exact variance) finite-difference inner product quasi-Newton test (17), and suppose that Assumptions A, B, G, H, and J hold. Then, if $\alpha_k = \alpha$ satisfies (34), for any $T \in \mathbb{Z}_{++}$, we have that

$$
\min_{0 \leq k \leq T-1} \mathbb{E} \left[ \| \nabla F(x_k) \|^2 \right] \leq \frac{2}{\alpha T \Lambda_1} (F(x_0) - F_{\text{min}}) + \frac{\Lambda_2 L_{\nabla F}^2 \nu^2 d}{4 \Lambda_1}.
$$

Proof Applying Lemma 4, from inequality (40) we have that

$$
\mathbb{E}_{S_k} \left[ F(x_{k+1}) \right] \leq F(x_k) - \frac{\alpha \Lambda_1}{2} \| \nabla F(x_k) \|^2 + \frac{\alpha \Lambda_2 L_{\nabla F}^2 \nu^2 d}{8}.
$$

Applying Lemma 8 with $a_1 = \alpha \Lambda_1$ and $a_2 = \frac{\alpha \Lambda_2 L_{\nabla F}^2 \nu^2 d}{8}$ completes the proof. $\square$

We note that the above results are an improvement over the convergence results provided in [33, Corollary 3.3] for a randomized stochastic gradient-free method (RSGF), for which the rate of convergence is $O \left( \frac{d}{T} + \sqrt{d/T} \right)$ as opposed to the $O \left( \frac{1}{T} \right)$ provided here. This improvement requires an increasing number of stochastic function evaluations per iteration (namely, $(d+1)|S_k|$) as opposed to the two stochastic function evaluations performed in each iteration of RSGF. However, the sample sizes employed in our settings only depend on the stochasticity in the problem (variance in the stochastic gradients) and so the computational cost per iteration is still linear in terms of the dimension $d$ of the problem. As mentioned earlier, the complexity results depend on the condition number of the quasi-Newton matrix, $\frac{\Lambda_2}{\Lambda_1}$, which is commonly seen in global analysis of quasi-Newton and Newton methods [8, 15, 56, 66].

We conclude this section by noting that the conditions in Theorems 6, 7, 9, and 10 can be met and are well defined. In particular, we recall that Assumption D on the variance of the stochastic functions additionally ensures that a sample $S_k$ can be selected to satisfy (12) and (17).
4 Nonsmooth subsampled functions

In this section we consider the scenario where the subsampled functions are nonsmooth (i.e., Assumption C is not satisfied), but the expected function $F$ is differentiable. This scenario arises in many applications; see [60, Figure 4] for an illustration. Section 5.2 provides another such example. We note that the sample selection procedure and the convergence analysis are still valid in this case. Algorithm 1 still works after some modifications.

4.1 Finite-difference parameter selection

We choose the finite-difference parameter by minimizing an upper bound on the error in the gradient approximation. The subsampled gradients do not exist, however, and we need to consider a different gradient approximation error. Here, we consider the scaled gradient approximation error in terms of the true finite-difference gradient. That is,

$$ r_k := H_k \left( \nabla^{FD} F_{S_k}(x_k) - \nabla F(x_k) \right) $$

$$ = H_k \left( \nabla^{FD} F_{S_k}(x_k) - \nabla^{FD} F(x_k) \right) + H_k \left( \nabla^{FD} F(x_k) - \nabla F(x_k) \right), $$

where we assume that $H_k$ satisfies Assumption G.

If samples satisfy the norm test (14), we have

$$ \mathbb{E}_{S_k} \left[ \left\| H_k \left( \nabla^{FD} F_{S_k}(x_k) - \nabla^{FD} F(x_k) \right) \right\| \right] \leq \Lambda_2 \theta \left\| \nabla^{FD} F(x_k) \right\|. $$

If samples satisfy the inner product quasi-Newton test (17) along with Assumption H, then from (33) we have

$$ \mathbb{E}_{S_k} \left[ \left\| H_k \left( \nabla^{FD} F_{S_k}(x_k) - \nabla^{FD} F(x_k) \right) \right\| \right] \leq \Lambda_2 \sqrt{\theta^2 + \psi^2} \left\| \nabla^{FD} F(x_k) \right\|. $$

Therefore, in both these cases we have

$$ \mathbb{E}_{S_k} \left[ \left\| H_k \left( \nabla^{FD} F_{S_k}(x_k) - \nabla^{FD} F(x_k) \right) \right\| \right] \leq \kappa \Lambda_2 \left\| \nabla^{FD} F(x_k) \right\|, $$

where $\kappa = \theta$ for the norm test and $\kappa = \sqrt{\theta^2 + \psi^2}$ for the inner product quasi-Newton test. Now, consider

$$ \mathbb{E}_{S_k} \left[ \left\| r_k \right\| \right] \leq \kappa \Lambda_2 \left\| \nabla^{FD} F(x_k) \right\| + \left\| H_k \left( \nabla^{FD} F(x_k) - \nabla F(x_k) \right) \right\| $nonsmooth subsampled functions

$$ \leq \kappa \Lambda_2 \left\| \nabla^{FD} F(x_k) \right\| + \Lambda_2 \left\| \nabla^{FD} F(x_k) - \nabla F(x_k) \right\| $nonsmooth subsampled functions

$$ \leq \kappa \Lambda_2 \left\| \nabla F(x_k) \right\| + \Lambda_2 (1 + \kappa) \left\| \nabla^{FD} F(x_k) - \nabla F(x_k) \right\|. $$
\[
\leq \kappa \Lambda_2 \| \nabla F(x_k) \| + \frac{\Lambda_2(1 + \kappa)L_{\nabla F} \nu \sqrt{d}}{2},
\]

where the third inequality is due to the fact that \( \|a\| \leq \|a - b\| + \|b\| \) and the last inequality is due to (7). We observe that the first term in the right-hand side of (41) is independent of the parameter \( \nu \). As discussed in Sect. 2.2, in any practical implementation one has to account for the numerical errors associated with numerical evaluations of the stochastic function values. Therefore, from (20) and (21), we have

\[
\| H_k \left( \nabla^{FD} \hat{F}_{S_k}(x_k) - \nabla^{FD} F_{S_k}(x_k) \right) \| \leq \frac{2\Lambda_2 \epsilon_m \sqrt{d}}{\nu}.
\]

Combining this with (41) and minimizing the resulting upper bound yields the optimal parameter as

\[
\nu^* := 2\sqrt{\frac{\epsilon_m}{L_{\nabla F}(1 + \kappa)}},
\]

where \( \kappa = \theta \) for the norm test and \( \kappa = \sqrt{\theta^2 + \psi^2} \) for the inner product quasi-Newton test. We note that the only difference between the optimal parameters in the smooth and nonsmooth cases is the presence of \( \kappa \) in the denominator and the use of the Lipschitz constant of the gradient of the expected function \( (L_{\nabla F}) \) instead of the Lipschitz constant of the subsampled gradient \( (L_{\nabla f}) \).

4.2 Step-length selection

In the smooth case we employed a stochastic line search to choose the step length \( \alpha_k \) by using a sufficient decrease condition (22) based on the subsampled function. In the nonsmooth case, it is not guaranteed that such a step length always exists. Intuitively, however, if the sample approximations are reasonably good, such a step length may exist since the expected function’s gradient is Lipschitz continuous. Therefore, in the algorithm we can still employ the sufficient decrease condition with a safeguarding mechanism. That is, if the step length \( \alpha_k \) falls below some threshold \( \alpha_{\text{min}} > 0 \), then we ignore the sufficient decrease condition and choose \( \alpha_k = \alpha_{\text{min}} \). The initial trial step length (24) is still valid here, and the reasoning behind this choice remains the same.

As a result, we modify line 21 of Algorithm 1 to break from the line search with \( \alpha_k = \alpha_{\text{min}} \) if \( \alpha_k \) is attempted to be reduced below \( \alpha_{\text{min}} \).

4.3 Quasi-newton update

In the smooth case we skip the update of quasi-Newton matrix whenever (28) is not satisfied, to ensure that \( \frac{y_k^T y_k}{y_k^T s_k} \) is bounded; doing so results in bounded eigenvalues.

In the nonsmooth case condition (28) does not guarantee that the \( \frac{y_k^T y_k}{y_k^T s_k} \) is bounded.
Instead, we impose the condition
\[ \|y_k\| \leq M\|s_k\|. \] (42)

The condition (42), along with (26), implies that
\[ \frac{y_k^T y_k}{y_k^T s_k} \leq \frac{\|y_k\|^2}{\beta_1\|s_k\|^2} \leq \frac{M^2}{\beta_1}, \]
in which case Assumption G still holds.

As a result, we modify line 25 of Algorithm 1 to replace the condition \( \|s_k\| > \beta_2 \) with the condition (42).

5 Numerical experiments

We now examine empirical characteristics of our proposed algorithm in both smooth (Sect. 5.1) and nonsmooth (Sect. 5.2) settings.

We implemented two variants, “FD-Norm” and “FD-IPQN,” of the proposed algorithm with the sample size \( |S_k| \) update chosen based on the finite-difference norm test in (Norm) and the inner product quasi-Newton test in (IPQN), respectively. We used \( \theta_0 = 0.9, |S_0| = 2, |S_k^x| = \min\{S_k, 1000\} \), finite-difference parameter \( \nu = 10^{-8} \) (i.e., we avoid any estimation of associated Lipschitz constants required for optimal \( \nu^* \) determination), L-BFGS memory parameter \( m = 10 \), and line search parameters \( c_1 = 10^{-4}, c_2 = 10^{-14}, \) and \( \tau = 0.5 \). We used \( \beta_1 = 10^{-3} \) and did not use the condition with \( \beta_2 \) (effectively setting it to a smaller value than would ever been encountered). For the nonsmooth problems we used \( \alpha_{\min} = 10^{-8} \). None of these parameters have been tuned to the problems being considered. We chose \( \gamma = 0.99 \) for smaller variance problems and \( \gamma = 0.9 \) for larger variance problems (for the variances given in Sect. 5.1). Source code for this implementation may be found in [19].

We also implemented two stochastic methods of the form
\[ x_{k+1} = x_k - \alpha_0 g_k, \]
where \( g_k \) is an estimation of the gradient. The first method is based on a classical stochastic gradient algorithm where the gradients are estimated by using finite differences. This method is also referred as the Kiefer–Wolfowitz algorithm [39]. We call the method here the \textit{finite-difference stochastic gradient method}, “FD-SG,” and \( g_k \) is chosen as \( \nabla^{FD} F_{S_k}(x_k) \) defined in (3). The second method also estimates the stochastic gradient; however, instead of employing finite differences in all the coordinate directions, it estimates the gradients using a small number of random directions chosen within a unit sphere. We call this method the \textit{sphere smoothing stochastic gradient method}, “SS-SG,” and refer the reader to [7] for further details. The gradient estimate
at each iteration is given by
\[ g_k = \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \sum_{j=1}^{T} \frac{d}{T} \frac{f(x + \nu u_j, \zeta_i) - f(x, \zeta_i)}{\nu} u_j, \]
where \( \{u_j \in \mathbb{R}^d\}_{j=1}^T \) are i.i.d. random vectors following a uniform distribution on the unit sphere centered at 0 of radius 1 and \( \nu \) is the standard difference parameter. We chose \( T = 5 \) for all the problems.

We report results for the best versions of FD-SG and SS-SG based on tuning the constant step length for each problem (i.e., by considering \( \alpha_0 = 2^j \), for \( j \in \{-25, -24, \ldots, 9, 10\} \)). We chose \(|S_k| = |S_0| = 2 \) for both these methods and again use the finite-difference parameter \( \nu = 10^{-8} \). For all the experiments we report the minimum, maximum, and mean results across 5 different random runs.

We implemented all the algorithms and ran the experiments in MATLAB R2019a on a 64-bit machine (machine precision \( \epsilon_m = 10^{-16} \)) with Intel Core i5@2.4 GHz and 8 GB of RAM.

### 5.1 Smooth problems

We conducted numerical experiments on stochastic nonlinear least squares problems based on a mapping \( \phi : \mathbb{R}^d \to \mathbb{R}^p \) affected by two forms of stochastic noise. Our functions affected by relative noise are of the form
\[ f_{\text{rel}}(x, \zeta) := \frac{1}{1 + \sigma^2} \sum_{j=1}^{p} \phi_j^2(x) (1 + \zeta_j)^2, \]
and our functions affected by absolute noise are of the form
\[ f_{\text{abs}}(x, \zeta) := \sum_{j=1}^{p} \left( (\phi_j(x) + \zeta_j)^2 - \sigma^2 \right), \]
where \( \sigma^2 > 0 \) is a variance parameter and \( \zeta \sim \mathcal{N}(0, \sigma^2 I_p) \). We note that this form of noise results in both random functions satisfying \( \mathbb{E}_\zeta[f(x, \zeta)] = \sum_{j=1}^{p} \phi_j^2(x) \). Furthermore, since the random variable \( \zeta \) has unbounded support, both stochastic functions have unbounded support at all \( x \in \mathbb{R}^d \), except when \( f = f_{\text{rel}} \) and \( \sum_{j=1}^{p} \phi_j^2(x) = 0 \). In both cases, the function \( f(\cdot, \zeta) \) and the expected function \( \mathbb{E}_\zeta[f(\cdot, \zeta)] \) are twice continuously differentiable.

We considered five different problems for \( \phi \) from the CUTEr [35] collection of optimization problems and used two values for the standard deviation parameter (\( \sigma = 10^{-3} \) and \( \sigma = 10^{-5} \)). The details of these problems are given in Table 1.

In all the experiments, we chose the initial starting point as \( x_0 = 10x_s \), where \( x_s \) is the standard starting point for these problems given in [48]. We computed the minimum function values \( F^* \) by running the L-BFGS method on the noise-free (i.e.,
Adaptive Sampling Quasi-Newton Methods… 355

Table 1 Characteristics of the nonlinear least squares problems used in our experiments

| Function  | \( p \) | \( d \) |
|-----------|--------|------|
| Chebyquad | 45     | 30   |
| Osborne   | 65     | 11   |
| Bdqrtic  | 92     | 50   |
| Cube      | 30     | 20   |
| Heart8ls  | 8      | 8    |
| BRATU3D   | 125    | 125  |
| EIGENC    | 110    | 110  |
| ConnBand  | 100    | 100  |
| ROSENBR   | 198    | 100  |
| PENALTY2  | 200    | 100  |
| PENLT1NE  | 101    | 100  |

\( \sigma = 0 \) problems until \( \| \nabla F(x) \|_\infty \leq 10^{-10} \) or the maximum number of 2,000 function evaluations is reached.

Figure 1 reports results on the chebyquad function with abs-normal noise and rel-normal noise for \( \sigma \) values of \( 10^{-3} \) and \( 10^{-5} \). The vertical axis measures the error in the function \( F(x) - F^* \), and the horizontal axis measures in terms of the total (i.e., including those in the gradient estimates, curvature pair updates, and line search) number of evaluations of \( f(x, \zeta) \). The results show that both variants of our finite-difference quasi-Newton method are more efficient than the tuned finite-difference stochastic gradient method and the tuned sphere-smoothing stochastic gradient method. Furthermore, on three of the four problems, the stochastic gradient methods converged to a significantly larger neighborhood of the solution as compared with the quasi-Newton variants in the high-variance problems (\( \sigma = 10^{-3} \)).

Of the two stochastic gradient methods, we observe that FD-SG is more efficient than SS-SG. We suspect that this performance might be attributed to the fact that these are low-dimensional problems and the computational savings obtained by sampling only few random directions (recall from Table 1 that \( d \) ranges from 8/5 to 10) for estimating the stochastic gradient do not overweigh the benefits associated with estimating the stochastic gradient accurately.

We also observe that both the variants of our algorithm have similar performance in terms of total function evaluations while FD-Norm has less variability compared to FD-IPQN. This behavior is explained by the fact that both these variants increase the sample sizes in a similar manner for this problem, as seen in Fig. 2.

We also report the step lengths chosen at each iteration by the two variants of our algorithm in Fig. 3 to illustrate the performance of the line search mechanism. We note that initially the step lengths are chosen to be small but they quickly go to a larger step length and stay around 1 until they converge to the neighborhood of the solution.

Results for the other problems listed in Table 1 are given in Online Appendix B of the Supplementary Material [18]. A summary of the best average \( F(x) - F^* \) values attained in all these problems are reported in Table 2. We note that the proposed
methods perform best except on BRATU3D with relative noise, for which SS-SG showed the best results.

5.2 Nonsmooth problems

We also conducted an experiment on a synthetic nonsmooth problem to illustrate the robustness of the proposed algorithm with respect to nonsmoothness of the stochastic functions. We considered the stochastic nonsmooth function

$$f(x, \zeta) = \|Ax - b - \zeta\|_1 = \sum_{i=1}^{p} |a_i^T x - b_i - \zeta_i|$$

(43)

where $\zeta \in \mathbb{R}^p$ is a random vector drawn uniformly from $[-1, 1]^p$. We note that the expected function $\mathbb{E}_{\zeta}[f(\cdot, \zeta)]$ is continuously differentiable and strongly convex; for details, see Online Appendix C of the Supplementary Material [18]. We set $A \in \mathbb{R}^{50 \times 50}$ as a symmetric normal random matrix and $b = Ax^*$, where $x^* \in \mathbb{R}^{50}$ is a normal random vector. For this problem, the optimal function value is $F^* = 25$. 

 Springer
Table 2 Best average $F(x) - F^*$ values across different random runs for the remaining problems given in Table 1 with different forms of stochastic noise and different variance values are reported here.

| Function  | Noise form | $\sigma$ | FD-SG      | SS-SG      | FD-Norm    | FD-IPQN    |
|-----------|------------|----------|------------|------------|------------|------------|
| Osborne   | abs        | $10^{-3}$| $6.6753e-07$ | $7.4344e-04$ | $2.3282e-09$ | $1.8824e-09$ |
|           |            | $10^{-5}$| $1.0544e-10$ | $7.4275e-04$ | $1.6624e-13$ | $1.2509e-13$ |
|           | rel        | $10^{-3}$| $2.0054e-09$ | $7.4250e-04$ | $4.6549e-12$ | $2.6651e-12$ |
|           |            | $10^{-5}$| $6.8846e-12$ | $7.4275e-04$ | $1.6764e-14$ | $2.1837e-14$ |
| Bdqrtic   | abs        | $10^{-3}$| $5.6897e-08$ | $3.8773e-07$ | $2.4910e-09$ | $5.6414e-09$ |
|           |            | $10^{-5}$| $6.2528e-12$ | $3.0514e-08$ | $1.7053e-12$ | $2.3306e-12$ |
|           | rel        | $10^{-3}$| $4.9662e-07$ | $2.9049e-06$ | $1.5850e-08$ | $7.6843e-08$ |
|           |            | $10^{-5}$| $5.0050e-11$ | $3.0060e-08$ | $7.5856e-12$ | $9.4644e-12$ |
| Cube      | abs        | $10^{-3}$| $3.8707e-03$ | $3.9712e-03$ | $5.7521e-05$ | $1.2977e-04$ |
|           |            | $10^{-5}$| $3.8704e-03$ | $3.9565e-03$ | $4.3196e-06$ | $1.8119e-05$ |
|           | rel        | $10^{-3}$| $1.5729e-03$ | $3.9556e-03$ | $6.9749e-07$ | $1.0068e-05$ |
|           |            | $10^{-5}$| $2.1794e-03$ | $3.9564e-03$ | $2.9567e-07$ | $2.8714e-06$ |
| Heart8ls  | abs        | $10^{-3}$| $1.3466e-07$ | $4.3781e-07$ | $3.4123e-08$ | $1.4136e-07$ |
|           |            | $10^{-5}$| $1.3354e-11$ | $4.2001e-08$ | $4.0514e-12$ | $5.0554e-12$ |
|           | rel        | $10^{-3}$| $6.7971e-12$ | $3.3372e-12$ | $7.4510e-12$ | $7.4543e-12$ |
|           |            | $10^{-5}$| $6.8086e-12$ | $3.3074e-12$ | $7.1041e-12$ | $7.1196e-12$ |
| BRATU3D   | abs        | $10^{-3}$| $1.0143e-06$ | $2.1045e-05$ | $6.2317e-08$ | $7.3464e-08$ |
|           |            | $10^{-5}$| $4.3117e-10$ | $4.8757e-09$ | $8.2560e-12$ | $9.3605e-12$ |
|           | rel        | $10^{-3}$| $1.0008e-11$ | $2.7165e-13$ | $8.2044e-12$ | $8.2034e-12$ |
| Function   | Noise form | $\sigma$ | FD-SG       | SS-SG      | FD-Norm      | FD-IPQN      |
|------------|------------|----------|-------------|------------|--------------|--------------|
|            |            | $10^{-5}$ | $1.0011e-11$ | $2.7304e-13$ | $9.7587e-12$ | $9.7587e-12$ |
| EIGENC     | abs        | $10^{-3}$ | $5.1682e-03$ | $1.3140e-01$ | $3.1919e-03$ | $1.2952e-02$ |
|            | rel        | $10^{-5}$ | $5.1595e-03$ | $1.3138e-01$ | $1.0899e-04$ | $4.5401e-04$ |
| ConnBand   | abs        | $10^{-3}$ | $1.4857e-03$ | $2.7463e-03$ | $6.6425e-05$ | $8.6210e-05$ |
|            | rel        | $10^{-5}$ | $1.4857e-03$ | $2.7459e-03$ | $1.3944e-05$ | $2.4759e-05$ |
| ROSENBR    | abs        | $10^{-3}$ | $7.5380e-05$ | $9.0566e+01$ | $1.4431e-08$ | $1.2905e-08$ |
|            | rel        | $10^{-5}$ | $7.5707e-09$ | $9.0565e+01$ | $6.6840e-12$ | $1.3016e-11$ |
| PENALTY2   | abs        | $10^{-3}$ | $6.9520e-06$ | $7.2453e-06$ | $5.3978e-07$ | $1.6472e-06$ |
|            | rel        | $10^{-5}$ | $6.9520e-06$ | $7.2453e-06$ | $3.7247e-06$ | $4.3828e-06$ |
| PENLTINE   | abs        | $10^{-3}$ | $6.2638e+00$ | $1.7615e+03$ | $5.9617e-05$ | $4.0628e-05$ |
|            | rel        | $10^{-3}$ | $6.2638e+00$ | $1.7615e+03$ | $2.0923e-02$ | $2.2798e-02$ |

Table 2 continued
Figure 2 reports results for a random instance of this problem. We observe that both variants of our finite-difference quasi-Newton method are more efficient than the tuned finite-difference stochastic gradient method and the tuned sphere-smoothing stochastic gradient method. We further note that because of the high variance arising due to the nonsmoothness, the methods converge at a slower rate.

6 Final remarks

We presented finite-difference quasi-Newton methods for solving derivative-free stochastic optimization problems where the sample sizes used in finite-difference gradient estimators are controlled by a modified norm test or an inner product quasi-Newton test. The numerical results show that the modified tests have potential for stochastic problems where the CRN approach is feasible. Early results on a challenging class of simulation-based finite-sum problems illustrate that such methods can be competitive even in settings where the batch size adaptivity is severely limited [16].

In this work, we considered forward finite differences in all the coordinate directions to estimate the gradients. It is interesting to consider other derivative-free techniques that estimate the gradients in smaller subspaces ($< d$) that might result in lower
computational effort. However, these approaches are challenging and require special attention to the curvature information used in quasi-Newton updates.

**Supplementary Information**  The online version contains supplementary material available at https://doi.org/10.1007/s12532-023-00233-9.

**Acknowledgements**  This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, applied mathematics and SciDAC
programs under Contract Nos. DE-AC02-05CH11231 and DE-AC02-06CH11357. The authors are grateful to Matt Menickelly and anonymous referees for thoughtful comments on early drafts.

**Funding** This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, applied mathematics and SciDAC programs under Contract Nos. DE-AC02-05CH11231 and DE-AC02-06CH11357.

**Data Availability** All data analyzed during this study can be reproduced with the ZOAdaQN package publicly available at [https://github.com/POptUS/ZOAdaQN](https://github.com/POptUS/ZOAdaQN) and [19].

**Code Availability** The full code was made available for review. The code may be obtained via the above references.

**Declarations**

**Conflict of Interest** The authors declare that they have no conflict of interest.

**References**

1. Agarwal, A., Dekel, O., Xiao, L.: Optimal algorithms for online convex optimization with multi-point bandit feedback. In: 23rd Conference on Learning Theory, pp. 28–40. (2010). [http://www.learningtheory.org/colt2010/conference-website/papers/037agarwal.pdf](http://www.learningtheory.org/colt2010/conference-website/papers/037agarwal.pdf)
2. Audet, C., Dzahini, K.J., Kokkolaras, M., Le Digabel, S.: Stochastic mesh adaptive direct search for blackbox optimization using probabilistic estimates. Comput. Optim. Appl. 79(1), 1–34 (2021). [https://doi.org/10.1007/s10589-020-00249-0](https://doi.org/10.1007/s10589-020-00249-0)
3. Audet, C., Hare, W.L.: Derivative-Free and Blackbox Optimization. Springer, Heidelberg (2017)
4. Balasubramanian, K., Ghadimi, S.: Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. Curran Associates Inc, New York (2018)
5. Balasubramanian, K., Ghadimi, S.: Zeroth-order nonconvex stochastic optimization: handling constraints, high dimensionality, and saddle points. Found. Comput. Math. 22, 35–76 (2022). [https://doi.org/10.1007/s10208-021-09499-8](https://doi.org/10.1007/s10208-021-09499-8)
6. Berahas, A.S., Byrd, R.H., Nocedal, J.: Derivative-free optimization of noisy functions via quasi-Newton methods. SIAM J. Optim. 29(2), 965–993 (2019). [https://doi.org/10.1137/18m1177718](https://doi.org/10.1137/18m1177718)
7. Berahas, A.S., Cao, L., Choromanski, K., Scheinberg, K.: A theoretical and empirical comparison of gradient approximations in derivative-free optimization. Found. Comput. Math. 22, 507–560 (2022). [https://doi.org/10.1007/s10208-021-09513-z](https://doi.org/10.1007/s10208-021-09513-z)
8. Berahas, A.S., Nocedal, J., Takáč, M.: A multi-batch L-BFGS method for machine learning. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 1055–1063. Curran Associates Inc., New York (2016)
9. Berahas, A.S., Takáč, M.: A robust multi-batch l-BFGS method for machine learning. Optim. Methods Softw. 35(1), 191–219 (2020). [https://doi.org/10.1080/10556788.2019.1658107](https://doi.org/10.1080/10556788.2019.1658107)
10. Bertsekas, D.P.: Reinforcement Learning and Optimal Control. Athena Scientific, Nashua (2019)
11. Bertsekas, D.P., Nedić, A., Ozdaglar, A.E.: Convex Analysis and Optimization. Athena Scientific, Nashua (2003)
12. Blanchet, J., Cartis, C., Menickelly, M., Scheinberg, K.: Convergence rate analysis of a stochastic trust-region method via supermartingales. INFORMS J. Optim. 1(2), 92–119 (2019). [https://doi.org/10.1287/ijoo.2019.0016](https://doi.org/10.1287/ijoo.2019.0016)
13. Blum, J.R.: Multidimensional stochastic approximation methods. Ann. Math. Stat. 25(4), 737–744 (1954). [https://doi.org/10.1214/aoms/1177728659](https://doi.org/10.1214/aoms/1177728659)
14. Bollapragada, R., Byrd, R., Nocedal, J.: Adaptive sampling strategies for stochastic optimization. SIAM J. Optim. 28(4), 3312–3343 (2018). [https://doi.org/10.1137/17m1154679](https://doi.org/10.1137/17m1154679)
15. Bollapragada, R., Byrd, R.H., Nocedal, J.: Exact and inexact subsampled Newton methods for optimization. IMA J. Numer. Anal. 39(2), 545–578 (2018). https://doi.org/10.1093/imanum/dry009

16. Bollapragada, R., Menickelly, M., Nazarewicz, W., O’Neal, J., Reinhard, P.G., Wild, S.M.: Optimization and supervised machine learning methods for fitting numerical physics models without derivatives. J. Phys. G Nucl. Part. Phys. 48(2), 024001 (2021). https://doi.org/10.1088/1361-6471/abd009

17. Bollapragada, R., Nocedal, J., Mudigere, D., Shi, H.J., Tang, P.T.P.: A progressive batching L-BFGS method for machine learning. In: J. Dy, A. Krause (eds.) Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 620–629. PMLR (2018). http://proceedings.mlr.press/v80/bollapragada18a.html

18. Bollapragada, R., Wild, S.M.: Online supplement to “Adaptive Sampling Quasi-Newton Methods for Zeroth-Order Stochastic Optimization” (2023)

19. Bollapragada, R., Wild, S.M.: ZOAdaQN: An adaptive quasi-Newton algorithm for zeroth-order stochastic optimization. https://github.com/POptUS/ZOAdaQN (2023). https://doi.org/10.5281/zenodo.7579239

20. Bollapragada, R., Byrd, R.H., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Math. Program. 134(1), 127–155 (2012). https://doi.org/10.1007/s10107-012-0572-5

21. Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Math. Program. 134(1), 127–155 (2012). https://doi.org/10.1007/s10107-012-0572-5

22. Cartis, C., Scheinberg, K.: Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. Math. Program. 169(2), 337–375 (2018)

23. Chang, K.H.: Stochastic Nelder-Mead simplex method - A new globally convergent direct search method for simulation optimization. Eur. J. Oper. Res. 220(3), 684–694 (2012). https://doi.org/10.1016/j.ejor.2012.02.028

24. Chen, R., Menickelly, M., Scheinberg, K.: Stochastic optimization using a trust-region method and random models. Math. Program. 169(2), 447–487 (2018). https://doi.org/10.1007/s10107-017-1141-8

25. Chen, X., Kelley, C.T.: Optimization with hidden constraints and embedded Monte Carlo computations. Optim. Eng. 17(1), 157–175 (2016). https://doi.org/10.1007/s11081-015-9302-1

26. Chen, X., Kelley, C.T., Xu, F., Zhang, Z.: A smoothing direct search method for Monte Carlo-based bound constrained composite nonsmooth optimization. SIAM J. Sci. Comput. 40(4), A2174–A2199 (2018). https://doi.org/10.1137/17m1116714

27. Chen, X., Liu, S., Xu, K., Li, X., Lin, X., Hong, M., Cox, D.: ZO-AdaMM: zeroth-order adaptive momentum method for black-box optimization. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. Curran Associates Inc, New York (2019)

28. Deng, G., Ferris, M.C.: Adaptation of the UOBYQA algorithm for noisy functions. In: Proceedings of the Winter Simulation Conference, pp. 312–319 (2006). https://doi.org/10.1109/wsc.2006.323088

29. Deng, G., Ferris, M.C.: Variable-number sample-path optimization. Math. Program. 117, 81–109 (2009). https://doi.org/10.1007/s10107-007-0164-y

30. Duchi, J.C., Jordan, M.I., Wainwright, M.J., Wibisono, A.: Optimal rates for zero-order convex optimization: the power of two function evaluations. IEEE Trans. Inf. Theory 61(5), 2788–2806 (2015). https://doi.org/10.1109/TIT.2015.2409256

31. Fu, M.C., Glover, F.W., April, J.: Simulation optimization: a review, new developments, and applications. In: Proceedings of the Winter Simulation Conference. IEEE (2005). https://doi.org/10.1109/wsc.2005.1574242

32. Gasnikov, A.V., Krymova, E.A., Lagunovskaya, A.A., Usmanova, I.N., Fedorenko, F.A.: Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case. Automat. Remote Control 78(2), 224–234 (2017)

33. Ghadimi, S.: Conditional gradient type methods for composite nonlinear and stochastic optimization. Math. Program. 173(1–2), 431–464 (2019). https://doi.org/10.1007/s10107-017-1225-5

34. Ghadimi, S., Lan, G.: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. 23(4), 2341–2368 (2013). https://doi.org/10.1137/120880811

35. Gould, N.I.M., Orban, D., Toint, P.L.: CUTEr and SifDec: a constrained and unconstrained testing environment, revisited. ACM Trans. Math. Softw. 29(4), 373–394 (2003). https://doi.org/10.1145/962437.962439
36. Huang, F., Gu, B., Huo, Z., Chen, S., Huang, H.: Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 1503–1510. (2019). https://doi.org/10.1609/aaai.v33i01.33011503
37. Huang, F., Tao, L., Chen, S.: Accelerated stochastic gradient-free and projection-free methods. In: H.D. III, A. Singh (eds.) Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 119, pp. 4519–4530. PMLR (2020). http://proceedings.mlr.press/v119/huang20j.html
38. Kelley, C.T.: Users Guide for imfil version 0.5. Available at www4.ncsu.edu/~ctk/imfil.html
39. Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. 22(3), 462–466 (1952). https://doi.org/10.1214/aoms/1177729392
40. Kim, S., Pasupathy, R., Henderson, S.G.: A guide to sample average approximation. In: Fu, M. (ed.) Handbook of Simulation Optimization, International Series in Operations Research & Management Science, pp. 207–243. Springer, New York (2015)
41. Kleinman, N.L., Spall, J.C., Naiman, D.Q.: Simulation-based optimization with stochastic approximation using common random numbers. Manag. Sci. 45(11), 1570–1578 (1999). https://doi.org/10.1287/mnsc.45.11.1570
42. Kungurtsev, V., Rinaldi, F.: A zeroth order method for stochastic weakly convex optimization. Comput. Optim. Appl. 80(3), 731–753 (2021). https://doi.org/10.1007/s10589-021-00313-3
43. Larson, J., Billups, S.C.: Stochastic derivative-free optimization using a trust region framework. Comput. Optim. Appl. 64(3), 619–645 (2016). https://doi.org/10.1007/s10589-016-9827-z
44. Larson, J., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. Acta Numer. 28, 287–404 (2019). https://doi.org/10.1017/s0962492919000060
45. L’Ecuyer, P., Yin, G.: Budget-dependent convergence rate of stochastic approximation. SIAM J. Optim. 8(1), 217–247 (1998). https://doi.org/10.1137/S1056264295270723
46. Liu, S., Kailkhura, B., Chen, P.Y., Ting, P., Chang, S., Amini, L.: Zeroth-order stochastic variance reduction for nonconvex optimization. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. Curran Associates Inc, New York (2018)
47. Mania, H., Guy, A., Recht, B.: Simple random search of static linear policies is competitive for reinforcement learning. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. Curran Associates Inc, New York (2018)
48. Moré, J.J., Wild, S.M.: Benchmarking derivative-free optimization algorithms. SIAM J. Optim. 20(1), 172–191 (2009). https://doi.org/10.1137/080724083
49. Moré, J.J., Wild, S.M.: Estimating computational noise. SIAM J. Sci. Comput. 33(3), 1292–1314 (2011). https://doi.org/10.1137/100786125
50. Moré, J.J., Wild, S.M.: Estimating derivatives of noisy simulations. ACM Trans. Math. Softw. 38(3), 19:1-19:21 (2012). https://doi.org/10.1145/2168773.2168777
51. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. 17(2), 527–566 (2017). https://doi.org/10.1007/s10208-015-9296-2
52. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2006)
53. Pasupathy, R., Ghosh, S.: Simulation optimization: a concise overview and implementation guide. In: Theory Driven by Influential Applications, pp. 122–150. INFORMS (2013). https://doi.org/10.1287/educ.2013.0118
54. Pasupathy, R., Glynn, P., Ghosh, S., Hashemi, F.S.: On sampling rates in simulation-based recursions. SIAM J. Optim. 28(1), 45–73 (2018). https://doi.org/10.1137/140951679
55. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. 22(3), 400–407 (1951). https://doi.org/10.1214/aoms/1177729586
56. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods. Math. Program. 174(1–2), 293–326 (2019). https://doi.org/10.1007/s10107-018-1346-5
57. Ross, S.M.: Simulation, 4th edn. Elsevier, Amsterdam (2006)
58. Sahu, A.K., Zaheer, M., Kar, S.: Towards gradient free and projection free stochastic optimization. In: K. Chaudhuri, M. Sugiyama (eds.) Proceedings of Machine Learning Research, Proceedings of Machine Learning Research, vol. 89, pp. 3468–3477. PMLR (2019)
59. Salimans, T., Ho, J., Chen, X., Sidor, S., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. Tech. Rep. arXiv:1703.03864, (2017)
60. Schaul, T., LeCun, Y.: Adaptive learning rates and parallelization for stochastic, sparse, non-smooth gradients. In: 1st International Conference on Learning Representations, ICLR 2013 (2013)

61. Shamir, O.: An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. J. Mach. Learn. Res. 18(52), 1–11 (2017)

62. Shashaani, S., Hashemi, F.S., Pasupathy, R.: ASTRO-DF: a class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. SIAM J. Optim. 28(4), 3145–3176 (2018). https://doi.org/10.1137/15m1042425

63. Shashaani, S., Hunter, S.R., Pasupathy, R.: ASTRO-DF: adaptive sampling trust-region optimization algorithms, heuristics, and numerical experience. In: 2016 Winter Simulation Conference (WSC). IEEE (2016). https://doi.org/10.1109/wsc.2016.7822121

64. Wibisono, A., Wainwright, M.J., Jordan, M.I., Duchi, J.C.: Finite sample convergence rates of zero-order stochastic optimization methods. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. Curran Associates Inc, New York (2012)

65. Xie, Y., Byrd, R.H., Nocedal, J.: Analysis of the BFGS method with errors. SIAM J. Optim. 30(1), 182–209 (2020). https://doi.org/10.1137/19m1240794

66. Xu, P., Roosta, F., Mahoney, M.W.: Newton-type methods for non-convex optimization under inexact Hessian information. Math. Program. 184(1–2), 35–70 (2020). https://doi.org/10.1007/s10107-019-01405-z

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.