Adaptive Sampling Quasi-Newton Methods for Zeroth-Order Stochastic Optimization

Raghu Bollapragada* Stefan M. Wild†

September 23, 2021

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 of a stochastic function using finite differences 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 the 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.

1 Introduction

We consider unconstrained stochastic optimization problems of the form

$$\min_{x\in\mathbb{R}^d} F(x) = \mathbb{E}_\zeta \left[f(x,\zeta) \right], \tag{1}$$

where one has access only to an oracle or a black-box procedure that outputs realizations of the stochastic function values $f(x,\zeta)$ and cannot access explicit estimates of the gradient $\nabla F(x)$. Such stochastic optimization problems arise in myriad science and engineering applications, from simulation optimization

*Operations Research and Industrial Engineering, The University of Texas at Austin, Austin, TX 78712. raghu.bollapragada@utexas.edu

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

[11, 26, 35, 47, 48] to reinforcement learning [9, 41, 52]. Several methods have been proposed to solve such derivative-free stochastic optimization problems, and we refer the reader to [3, 38] for surveys of these methods. A popular class of these methods estimate the gradients using function values and employ standard gradient-based optimization methods using these estimators.

Quasi-Newton methods 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 [15] 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 function values.

We propose finite-difference derivative-free stochastic quasi-Newton methods for solving (1) by exploiting *common random number (CRN)* evaluations of f. The CRN setting allows us to define subsampled gradient estimators

$$\left[\nabla^{\mathrm{FD}} F_{\zeta_i}(x)\right]_j := \frac{f(x+\nu e_j, \zeta_i) - f(x, \zeta_i)}{\nu}, \; j = 1, \ldots, d \tag{2}$$

$$\nabla^{\mathrm{FD}} F_{S_k}(x) := \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \nabla^{\mathrm{FD}} F_{\zeta_i}(x), \tag{3}$$

which employ forward differences for the independent and identically distributed (i.i.d.) samples of ζ 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.

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 [13, 15, 16] to the optimal solution. 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.

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 Section 3 establishes theoretical convergence results. Section 4 describes the algorithmic components for handling nonsmooth subsampled functions. Numerical experiments are provided in Section 5, and concluding remarks are provided in Section 6.

1.1 Literature Review

Finite-difference-based versions of the standard stochastic gradient method ("stochastic approximation") of Robbins and Monro [50] soon followed that work, in both univariate [34] and multivariate [12] settings. Stochastic approximation methods based on CRNs were analyzed in [36, 39].

Kelley [33] 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 [43] 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, 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, 22, 25, 27, 29, 40, 45, 51, 53, 56].

Related classes of methods for nonconvex stochastic optimization include zeroth-order extensions of both conditional gradient methods [4, 5, 28] and other proximal-point approaches [31, 32].

Model-based trust-region methods [11, 19, 23, 24, 37, 54, 55] and direct search methods [2, 18, 20, 21] 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.

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

$$F_{S_k}(x) := \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} f(x, \zeta_i). \tag{4}$$

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 ζ. That is, for all $x \in \mathbb{R}^d$ and $k \in \mathbb{Z}_+$,*

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

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

$$\begin{aligned} \mathbb{E}_{S_k}\left[\nabla^{\mathrm{FD}} F_{S_k}(x_k)\right] &= \mathbb{E}_{S_k}\left[\tfrac{1}{|S_k|}\textstyle\sum_{\zeta_i \in S_k}\left[\tfrac{f(x_k + \nu e_j, \zeta_i) - f(x_k, \zeta_i)}{\nu}\right]_{j=1}^{d}\right] \\ &= \nabla^{\mathrm{FD}} F(x_k), \end{aligned} \tag{5}$$

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

$$\nabla^{\mathrm{FD}} F(x) := \left[\frac{F(x + \nu e_j) - F(x)}{\nu}\right]_{j=1}^{d}. \tag{6}$$

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 $L_{\nabla F} > 0$.*

When combined with Assumption A, Assumption B implies that $\nabla^{\mathrm{FD}} F_{S_k}(x_k)$ is a biased estimator of the gradient $\nabla F(x_k)$ and that the bias can be deterministically quantified by

$$\begin{aligned} \left\|\nabla^{\mathrm{FD}} F(x_k) - \nabla F(x_k)\right\|^2 &= \sum_{j=1}^{d} \left(\frac{F(x_k + \nu e_j) - F(x_k)}{\nu} - \left[\nabla F(x_k)\right]_j\right)^2 \\ &\leq \sum_{j=1}^{d} \left(\frac{L_{\nabla F}\nu}{2}\right)^2 \\ &= \left(\frac{L_{\nabla F}\nu\sqrt{d}}{2}\right)^2, \end{aligned} \tag{7}$$

where the inequality follows from the following result, which holds for functions F with $L_{\nabla F}$-Lipschitz continuous gradients.

Lemma 1 (Descent Lemma [10]). *If $F : \mathbb{R}^d \mapsto \mathbb{R}$ is continuously differentiable with a $L_{\nabla F}$-Lipschitz continuous gradient on $\mathbb{R}^d$, then*

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

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 Section 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 ζ, the stochastic function $f(\cdot, \zeta)$ in (1) is continuously differentiable and has Lipschitz continuous gradients with Lipschitz constant $L_{\nabla f} > 0$.*

Assumption C implies that any subsampled gradient

$$\nabla F_{S_k}(x) := \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \nabla_x f(x, \zeta_i)$$

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 ∇F. In Section 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*

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

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, the only remaining source of randomness is from the sample set S_k. For ease of exposition, we drop this conditional dependence on x_k and hence expectations are shown with respect to only the sampling until the analysis of Section 3.3.

2 A Zeroth-Order Stochastic Quasi-Newton Algorithm

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

$$x_{k+1} = x_k - \alpha_k H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k), \tag{8}$$

where $\alpha_k > 0$ is the step length, H_k is a positive-definite quasi-Newton matrix, and $\nabla^{\mathrm{FD}} F_{S_k}(x_k)$ 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 [38]) can be employed to estimate the gradient.

We now discuss the algorithmic components consisting of sample size selection (Section 2.1), finite-difference parameter and step-length selection (Sections 2.2 and 2.3, respectively), and quasi-Newton updates (Section 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 [16], Equation (15) in [17]) for gradient estimators g_k to satisfy is the *norm condition* given by

$$\|g_k - \nabla F(x_k)\|^2 \leq \theta^2 \|\nabla F(x_k)\|^2, \quad \theta > 0. \tag{9}$$

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,

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

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

$$\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F(x_k) = \underbrace{\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla^{\mathrm{FD}} F(x_k)}_{\text{sampling error}} + \underbrace{\nabla^{\mathrm{FD}} F(x_k) - \nabla F(x_k)}_{\text{bias}}, \tag{11}$$

where $\nabla^{\mathrm{FD}} F$ 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

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

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

$$\begin{aligned}
&\mathbb{E}_{S_k}\left[\left\|\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F(x_k)\right\|^2\right] \\
&\quad\leq \mathbb{E}_{S_k}\left[\left\|\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right] + \left\|\nabla^{\mathrm{FD}} F(x_k) - \nabla F(x_k)\right\|^2 \\
&\quad\leq \theta^2 \left\|\nabla^{\mathrm{FD}} F(x_k)\right\|^2 + \left\|\nabla^{\mathrm{FD}} F(x_k) - \nabla F(x_k)\right\|^2 \\
&\quad\leq 2\theta^2 \left\|\nabla F(x_k)\right\|^2 + (1+2\theta^2)\left\|\nabla^{\mathrm{FD}} F(x_k) - \nabla F(x_k)\right\|^2 \\
&\quad\leq 2\theta^2 \left\|\nabla F(x_k)\right\|^2 + \frac{(1+2\theta^2) L_{\nabla F}^2 \nu^2 d}{4},
\end{aligned}$$

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 ($\nabla^{\mathrm{FD}} F_{\zeta_i}$; recall (2)). That is,

$$\mathbb{E}_{S_k}\left[\left\|\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right] \leq \frac{\mathbb{E}_{\zeta_i}\left[\left\|\nabla^{\mathrm{FD}} F_{\zeta_i}(x_k) - \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right]}{|S_k|}. \tag{13}$$

To be meaningful, such a bound requires that the true variance be bounded, which is guaranteed by Assumption D; the proof is given in Appendix A.1. Consequently, the condition

$$\frac{\mathbb{E}_{\zeta_i}\left[\left\|\nabla^{\mathrm{FD}} F_{\zeta_i}(x_k) - \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right]}{|S_k|} \leq \theta^2 \|\nabla^{\mathrm{FD}} F(x_k)\|^2 \tag{14}$$

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*

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

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

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

In our algorithm, we test condition (Norm); and whenever it is not satisfied, we increase $|S_k|$ until (Norm) is satisfied.

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. [15] 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,

$$\left(H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right)^T H_k \nabla F(x_k) > 0 \tag{15}$$

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

$$(H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k))^T H_k \nabla^{\mathrm{FD}} F(x_k) + (H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k))^T (H_k \nabla F(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)), \tag{16}$$

and, taking an expectation, we obtain

$$\begin{aligned}
&\mathbb{E}_{S_k}\left[(H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k))^T H_k \nabla F(x_k)\right] \\
&\quad = \left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2 + (H_k \nabla^{\mathrm{FD}} F(x_k))^T \left(H_k \nabla F(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)\right) \\
&\quad \geq \left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2 - \left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\| \left\|H_k \nabla F(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)\right\| \\
&\quad \geq \left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\| \left(\|H_k \nabla F(x_k)\| - 2\left\|H_k \nabla F(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)\right\|\right) \\
&\quad \geq \left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\| \left(\|H_k \nabla F(x_k)\| - 2\|H_k\| \left\|\nabla F(x_k) - \nabla^{\mathrm{FD}} F(x_k)\right\|\right) \\
&\quad \geq \left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\| \left(\|H_k \nabla F(x_k)\| - \|H_k\| L_{\nabla F} \nu \sqrt{d}\right),
\end{aligned}$$

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 $\|\nabla F(x_k)\| < \frac{\lambda_{\max}(H_k) L_{\nabla F} \nu \sqrt{d}}{\lambda_{\min}(H_k)}$, where $\lambda_{\max}(H_k)$ and $\lambda_{\min}(H_k) > 0$ are the largest and smallest eigenvalues of H_k, respectively, it is not guaranteed that the inequality in (15) can be satisfied in expectation. 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

$$\mathbb{E}_{S_k}\left[\left(\left(H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right)^T H_k \nabla^{\mathrm{FD}} F(x_k) - \left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right)^2\right] \leq \theta^2 \|H_k \nabla^{\mathrm{FD}} F(x_k)\|^4, \quad (17)$$

where $\mathbb{E}_{S_k}\left[H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right] = H_k \nabla^{\mathrm{FD}} F(x_k)$ 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 Appendix A.1. Therefore, for ensuring (17), it is sufficient for

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

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*

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

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

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

This variance computation requires only one additional Hessian-vector product (i.e., the product of H_k with $H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)$). 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

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

We observe that Term 2 in (18) is independent of the parameter ν. Using Assumption C on the sample path functions, we can bound Term 1 by

$$\begin{aligned}
&\left\|\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F_{S_k}(x_k)\right\|^2 \\
&\qquad = \sum_{j=1}^{d} \left(\frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \left(\frac{f(x_k + \nu e_j, \zeta_i) - f(x_k, \zeta_i)}{\nu} - [\nabla_x f(x_k, \zeta_i)]_j \right) \right)^2 \\
&\qquad \leq \left(\frac{L_{\nabla f} \nu \sqrt{d}}{2} \right)^2,
\end{aligned} \tag{19}$$

which decreases as ν decreases. In any practical implementation, however, one has to account for the numerical errors associated with the numerical evaluation of the function values. We employ the following assumption on a uniform bound for these errors.

Assumption E. *The function values $f(x, \zeta)$ in (1) are corrupted by numerical noise $\epsilon(x, \zeta)$ uniformly bounded by $\epsilon_m > 0$; that is,*

$$|\epsilon(x, \zeta)| \leq \epsilon_m \qquad \textit{for all } x, \zeta.$$

Applying Assumption E, we get the corrupted gradient estimator

$$\nabla^{\mathrm{FD}} \hat{F}_{S_k}(x_k) := \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \left[\frac{f(x + \nu e_j, \zeta_i) + \epsilon(x + \nu e_j, \zeta_i) - f(x, \zeta_i) - \epsilon(x, \zeta_i)}{\nu} \right]_{j=1}^{d} \tag{20}$$

$$= \nabla^{\mathrm{FD}} F_{S_k}(x_k) + \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \left[\frac{\epsilon(x + \nu e_j, \zeta_i) - \epsilon(x, \zeta_i)}{\nu} \right]_{j=1}^{d},$$

and hence

$$\|\nabla^{\mathrm{FD}} \hat{F}_{S_k}(x_k) - \nabla^{\mathrm{FD}} F_{S_k}(x_k)\| \leq \frac{2\epsilon_m \sqrt{d}}{\nu}. \tag{21}$$

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

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

This optimal finite-difference parameter is analogous to the one derived in [44], 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 ζ.

2.3 Step-Length Selection

We employ a stochastic line search to choose the step length α_k in (8) by using a sufficient decrease condition on the subsampled function. In particular, we would like α_k to satisfy

$$F_{S_k}\left(x_k - \alpha_k H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right) \leq F_{S_k}(x_k) - c_1\alpha_k (\nabla^{\mathrm{FD}} F_{S_k}(x_k))^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) + c_2, \tag{22}$$

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 α_k that does not satisfy (22) is reduced by a fixed fraction $\tau < 1$ (i.e., $\alpha_k \leftarrow \tau\alpha_k$). In Theorem 2, we establish that there exists a nontrivial interval for α_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 (22) holds for any*

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

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

$$
\begin{aligned}
&F_{S_k}\left(x_k - \alpha_k H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right) \\
&\quad \leq F_{S_k}(x_k) - \alpha_k \nabla F_{S_k}(x_k)^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) + \frac{L_{\nabla f}\alpha_k^2}{2}\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\|^2 \\
&\quad = F_{S_k}(x_k) - \alpha_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) \\
&\qquad + \alpha_k (\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F_{S_k}(x_k))^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) + \frac{L_{\nabla f}\alpha_k^2}{2}\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\|^2 \\
&\quad \leq F_{S_k}(x_k) - \alpha_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) + \frac{\alpha_k}{2}\nabla^{\mathrm{FD}} F_{S_k}(x_k)^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) \\
&\qquad + \frac{\alpha_k}{2}(\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F_{S_k}(x_k))^T H_k (\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F_{S_k}(x_k)) \\
&\qquad + \frac{L_{\nabla f}\alpha_k^2}{2}\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\|^2 \\
&\quad = F_{S_k}(x_k) - \frac{\alpha_k}{2}\nabla^{\mathrm{FD}} F_{S_k}(x_k)^T H_k^{1/2}\left(I - L_{\nabla f}\alpha_k H_k\right) H_k^{1/2} \nabla^{\mathrm{FD}} F_{S_k}(x_k) \\
&\qquad + \frac{\alpha_k}{2}(\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F_{S_k}(x_k))^T H_k (\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F_{S_k}(x_k)) \\
&\quad \leq F_{S_k}(x_k) - \frac{\alpha_k\left(1 - \alpha_k L_{\nabla f}\lambda_{\max}(H_k)\right)}{2}\nabla^{\mathrm{FD}} F_{S_k}(x_k)^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) \\
&\qquad + \frac{\alpha_k \lambda_{\max}(H_k)}{2}\|\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F_{S_k}(x_k)\|^2 \\
&\quad \leq F_{S_k}(x_k) - \frac{\alpha_k\left(1 - \alpha_k L_{\nabla f}\lambda_{\max}(H_k)\right)}{2}\nabla^{\mathrm{FD}} F_{S_k}(x_k)^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) \\
&\qquad + \frac{\alpha_k \lambda_{\max}(H_k) L_{\nabla f}^2 \nu^2 d}{8} \\
&\quad \leq F_{S_k}(x_k) - c_1 \alpha_k (\nabla^{\mathrm{FD}} F_{S_k}(x_k))^T H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) + c_2,
\end{aligned}
$$

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). □

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. [15] proposed a heuristic to choose the initial trial estimate for α_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

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

The formal reasoning for this choice is provided in Appendix A.3.

2.4 Stable Quasi-Newton Update

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

$$H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T, \qquad \rho_k = (y_k^T s_k)^{-1}, \qquad V_k = I - \rho_k y_k s_k^T,$$

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 [15]. We follow the same approach and define

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

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 μ-strongly convex function F_{S_k}, we have that

$$\begin{aligned}
y_k^T s_k &= \left(\nabla^{\mathrm{FD}} F_{S_k}(x_{k+1}) - \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right)^T s_k \\
&= \left(\nabla F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_k)\right)^T s_k \\
&\quad + \left(\nabla^{\mathrm{FD}} F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_{k+1}) + \nabla F_{S_k}(x_k) - \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right)^T s_k \\
&\geq \mu \|s_k\|^2 \\
&\quad - \left(\|\nabla^{\mathrm{FD}} F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_{k+1})\| + \|\nabla F_{S_k}(x_k) - \nabla^{\mathrm{FD}} F_{S_k}(x_k)\|\right) \|s_k\| \\
&\geq \mu \|s_k\|^2 - L_{\nabla f} \nu \sqrt{d} \|s_k\| = \|s_k\| \left(\mu \|s_k\| - L_{\nabla f} \nu \sqrt{d}\right),
\end{aligned}$$

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 $\|s_k\| > \frac{L_{\nabla f} \nu \sqrt{d}}{\mu}$. Recently, Xie et al. [57] 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:

$$y_k^T s_k > \beta_1 \|s_k\|^2, \tag{26}$$

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,

$$\begin{aligned}\frac{y_k^T y_k}{y_k^T s_k} &= \frac{\|\nabla^{\mathrm{FD}} F_{S_k}(x_{k+1}) - \nabla^{\mathrm{FD}} F_{S_k}(x_k)\|^2}{y_k^T s_k} \\ &\leq 3\frac{\|\nabla F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_k)\|^2}{\beta_1\|s_k\|^2} + 3\frac{\|\nabla^{\mathrm{FD}} F_{S_k}(x_{k+1}) - \nabla F_{S_k}(x_{k+1})\|^2}{\beta_1\|s_k\|^2} \\ &\quad + 3\frac{\|\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F_{S_k}(x_k)\|^2}{\beta_1\|s_k\|^2} \\ &\leq \frac{3L_{\nabla f}^2}{\beta_1} + \frac{3L_{\nabla f}^2\nu^2 d}{2\beta_1\|s_k\|^2}, \end{aligned} \tag{27}$$

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:

$$\|s_k\| > \beta_2 > 0, \tag{28}$$

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. 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 $\kappa_k = \frac{y_k^T s_k}{y_k^T y_k}$.

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 θ to be a fixed parameter, we make it iteration dependent and control it adaptively.

The parameter θ controls the probability of satisfying the underlying deterministic condition. For example, in the inner product quasi-Newton test, θ controls the probability of generating a quasi-Newton direction that makes an acute angle with the true quasi-Newton direction. Smaller θ values increase the probability of

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 ν
line search parameters (c_1, c_2, τ), sample test parameters θ_0, γ.
Initialization: Set $k \leftarrow 0$; $\theta = \theta_0$
Repeat until convergence:

Choose a set S_k consisting of $|S_k|$ i.i.d. realizations of ζ
switch (Sample Selection:)
case Finite-Difference Norm Test**:**
  if (Norm) is not satisfied **then**
    Choose least $|S_k|$ such that the inequality in (Norm) is satisfied
  end if
case Finite-Difference Inner Product Quasi-Newton Test**:**
  if (IPQN) is not satisfied **then**
    Choose least $|S_k|$ such that the inequality in (IPQN) is satisfied
  end if
end switch
if $|S_k| = |S_{k-1}|$ **then**
  Set $\theta \leftarrow \theta\gamma$
else
  Set $\theta \leftarrow \theta_0$
end if
Compute $\nabla^{\mathrm{FD}} F_{S_k}(x_k)$
Compute $p_k = -H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)$ using L-BFGS two-loop recursion in [46]
Compute α_k using (24)
while Armijo condition (22) is not satisfied **do**
  Set $\alpha_k \leftarrow \alpha_k \tau$
end while
Compute $x_{k+1} = x_k + \alpha_k p_k$
Compute y_k using (25) and set $s_k = x_{k+1} - x_k$
if $y_k^T s_k > \beta_1 \|s_k\|^2$ and $\|s_k\| > \beta_2$ **then**
  if number of stored (y_j, s_j) exceeds m **then**
    Discard oldest curvature pair (y_j, s_j)
  end if
  Store new curvature pair (y_k, s_k)
end if
Set $k \leftarrow k + 1$
Set $|S_k| = |S_{k-1}|$

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 θ value in the next iteration. If the sample size has increased in the next iteration, we reset the value to its default value θ_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 θ_0.

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*

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

Assumption G can be shown to hold for both convex and nonconvex twice-differentiable functions by updating H_k only when $y_k^T s_k \geq \beta_1 \|s_k\|_2^2$, where $\beta_1 > 0$ is a predetermined constant [8]. We provide the proof for the sake of completeness in Appendix A.4. 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.

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 α_k satisfies*

$$0 < \alpha_k \leq \frac{\Lambda_1}{4(1+\theta^2)L_{\nabla F}\Lambda_2^2}, \tag{29}$$

we have that

$$\mathbb{E}_{S_k}\left[F(x_{k+1})\right] \leq F(x_k) - \frac{\alpha_k \Lambda_1}{4}\|\nabla F(x_k)\|^2 + \frac{\alpha_k(\Lambda_1 + 2\Lambda_2)}{4}\|\nabla^{FD} F(x_k) - \nabla F(x_k)\|^2. \tag{30}$$

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

$$\begin{aligned}
\mathbb{E}_{S_k}\left[F(x_{k+1})\right] &\leq F(x_k) - \mathbb{E}_{S_k}\left[\alpha_k \left(H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right)^T \nabla F(x_k)\right] \\
&\quad + \mathbb{E}_{S_k}\left[\frac{L_{\nabla F}\alpha_k^2}{2}\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\|^2\right] \\
&= F(x_k) - \alpha_k \nabla^{\mathrm{FD}} F(x_k)^T H_k \nabla F(x_k) \\
&\quad + \frac{L_{\nabla F}\alpha_k^2}{2}\mathbb{E}_{S_k}\left[\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\|^2\right],
\end{aligned}$$

where the equality follows from Assumption A. Defining

$$\begin{aligned}
\delta_k &:= \nabla^{\mathrm{FD}} F(x_k) - \nabla F(x_k) \\
T_k &:= \frac{L_{\nabla F}\alpha_k^2}{2}\mathbb{E}_{S_k}\left[\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\|^2\right],
\end{aligned} \tag{31}$$

we have that

$$\begin{aligned}
\mathbb{E}_{S_k}\left[F(x_{k+1})\right] &\leq F(x_k) - \alpha_k \left(\nabla F(x_k) + \delta_k\right)^T H_k \nabla F(x_k) + T_k \\
&= F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) - \alpha_k \delta_k^T H_k \nabla F(x_k) + T_k \\
&\leq F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) \\
&\quad + \frac{\alpha_k}{2}\left(\nabla F(x_k)^T H_k \nabla F(x_k) + \delta_k^T H_k \delta_k\right) + T_k \\
&= F(x_k) - \frac{\alpha_k}{2}\nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\alpha_k}{2}\delta_k^T H_k \delta_k + T_k,
\end{aligned} \tag{32}$$

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

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

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

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

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

which establishes (30). □

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 [15], we assume that the orthogonality condition is satisfied by the stochastic finite-difference quasi-Newton directions.

Assumption H. *For*

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

there exists $\psi > 0$ *such that*

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

Using the proof techniques in [15, Lemma 1], we thus have the following bound on the length of the search direction:

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

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 α_k satisfies*

$$0 < \alpha_k < \frac{1}{(1+\theta^2+\psi^2)\, L_{\nabla F} \Lambda_2}, \tag{34}$$

we have that

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

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

$$\begin{aligned}
&\mathbb{E}_{S_k}\left[F(x_{k+1})\right] \\
&\quad \leq F(x_k) - \mathbb{E}_{S_k}\left[\alpha_k \left(H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right)^T \nabla F(x_k)\right] \\
&\qquad + \mathbb{E}_{S_k}\left[\frac{L_{\nabla F}\alpha_k^2}{2}\left\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right\|^2\right] \\
&\quad = F(x_k) - \alpha_k \nabla^{\mathrm{FD}} F(x_k)^T H_k \nabla F(x_k) + \frac{L_{\nabla F}\alpha_k^2}{2}\,\mathbb{E}_{S_k}\left[\left\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right\|^2\right] \\
&\quad \leq F(x_k) - \alpha_k \nabla^{\mathrm{FD}} F(x_k)^T H_k \nabla F(x_k) \\
&\qquad + \frac{L_{\nabla F}\alpha_k^2}{2}\left(1+\theta^2+\psi^2\right)\left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2,
\end{aligned}$$

where the last inequality is due to Assumption H and (33).

By using δ_k from (31), $\tilde{L}_{\nabla F} := L_{\nabla F}(1+\theta^2+\psi^2)$, and Assumption G, we have that

$$\begin{aligned}
&\mathbb{E}_{S_k}\left[F(x_{k+1})\right] \\
&\quad \leq F(x_k) - \alpha_k(\nabla F(x_k)+\delta_k)^T H_k \nabla F(x_k) + \frac{\tilde{L}_{\nabla F}\alpha_k^2}{2}\|H_k(\nabla F(x_k)+\delta_k)\|^2 \\
&\quad = F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\tilde{L}_{\nabla F}\alpha_k^2}{2}(\|H_k \nabla F(x_k)\|^2 + \|H_k\delta_k\|^2) \\
&\qquad - \alpha_k (H_k^{1/2}\delta_k)^T (I - \tilde{L}_{\nabla F}\alpha_k H_k)(H_k^{1/2}\nabla F(x_k)) \\
&\quad \leq F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\tilde{L}_{\nabla F}\alpha_k^2}{2}(\|H_k \nabla F(x_k)\|^2 + \|H_k\delta_k\|^2) \\
&\qquad + \frac{\alpha_k}{2}\left((H_k^{1/2}\nabla F(x_k))^T (I - \tilde{L}_{\nabla F}\alpha_k H_k)(H_k^{1/2}\nabla F(x_k))\right) \\
&\qquad + \frac{\alpha_k}{2}\left((H_k^{1/2}\delta_k)^T (I - \tilde{L}_{\nabla F}\alpha_k H_k)(H_k^{1/2}\delta_k)\right) \\
&\quad = F(x_k) - \frac{\alpha_k}{2}\nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\alpha_k}{2}\delta_k^T H_k \delta_k \\
&\quad \leq F(x_k) - \frac{\alpha_k \Lambda_1}{2}\|\nabla F(x_k)\|^2 + \frac{\alpha_k \Lambda_2}{2}\|\delta_k\|^2,
\end{aligned}$$

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

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, which can be obtained by integrating all random variables $x_k, \ldots, x_1$ obtained through k iterations of the form (8).

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 ∇F exists (as is the case under either Assumption B or Assumption C).

Assumption I. *There exists a parameter $\mu > 0$ such that*

$$\|\nabla F(x)\|^2 \geq 2\mu\,(F(x) - F(x^*)) \qquad \forall x \in \mathbb{R}^d.$$

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*

$$\mathbb{E}_{S_k}[F(x_{k+1})] \leq F(x_k) - \frac{a_1}{2}\|\nabla F(x_k)\|^2 + a_2 \tag{36}$$

for some constants $a_1 > 0$ and $a_2 > 0$. Then,

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

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

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

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

$$\begin{aligned}\mathbb{E}\left[F(x_{k+1}) - F(x^*)\right] - \frac{a_2}{\mu a_1} &\leq (1 - \mu a_1)\,\mathbb{E}\left[F(x_k) - F(x^*)\right] + a_2 - \frac{a_2}{\mu a_1} \\ &= (1 - \mu a_1)\left(\mathbb{E}\left[F(x_k) - F(x^*)\right] - \frac{a_2}{\mu a_1}\right). \end{aligned} \tag{37}$$

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 Appendix A.2.

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*

$$\mathbb{E}\left[F(x_k) - F(x^*)\right] \leq \left(1 - \frac{\mu\Lambda_1\alpha}{2}\right)^k (F(x_0) - F(x^*)) + \frac{(\Lambda_1 + 2\Lambda_2)L_{\nabla F}^2\nu^2 d}{8\mu\Lambda_1}. \tag{38}$$

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

$$\mathbb{E}_{S_k}\left[F(x_{k+1})\right] \leq F(x_k) - \frac{\alpha\Lambda_1}{4}\|\nabla F(x_k)\|^2 + \frac{\alpha(\Lambda_1 + 2\Lambda_2)L_{\nabla F}^2\nu^2 d}{16}. \tag{39}$$

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

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*

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

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

$$\mathbb{E}_{S_k}\left[F(x_{k+1})\right] \leq F(x_k) - \frac{\alpha\Lambda_1}{2}\|\nabla F(x_k)\|^2 + \frac{\alpha\Lambda_2 L_{\nabla F}^2\nu^2 d}{8}. \tag{40}$$

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

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 $F_{\min}$ with $-\infty < F_{\min} \leq F(x) \quad \forall x \in \mathbb{R}^d$.*

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*

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

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 $F_{\min}$, we get

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

Therefore, we can conclude that

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

□

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*

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

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. □

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*

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

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. □

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; that is, Assumption C is not satisfied. We note that the sample selection procedure and the convergence analysis are still valid in this case. Algorithm 1 still works but requires some modifications tailored to this setting.

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,

$$\begin{aligned} r_k :=& H_k\left(\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla F(x_k)\right) \\ =& H_k\left(\nabla^{\mathrm{FD}} F_{S_k}(x_k) - \nabla^{\mathrm{FD}} F(x_k)\right) + H_k\left(\nabla^{\mathrm{FD}} F(x_k) - \nabla F(x_k)\right), \end{aligned}$$

where we assume that H_k satisfies Assumption G.

If samples satisfy the norm test, we have

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

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

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

Therefore, in both these cases we have

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

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

$$\begin{aligned}\mathbb{E}_{S_k}\left[\|r_k\|\right]&\leq\kappa\Lambda_2\left\|\nabla^{\mathrm{FD}}F(x_k)\right\|+\left\|H_k(\nabla^{\mathrm{FD}}F(x_k)-\nabla F(x_k))\right\|\\&\leq\kappa\Lambda_2\left\|\nabla^{\mathrm{FD}}F(x_k)\right\|+\Lambda_2\left\|\nabla^{\mathrm{FD}}F(x_k)-\nabla F(x_k)\right\|\\&\leq\kappa\Lambda_2\left\|\nabla F(x_k)\right\|+\Lambda_2(1+\kappa)\left\|\nabla^{\mathrm{FD}}F(x_k)-\nabla F(x_k)\right\|\\&\leq\kappa\Lambda_2\left\|\nabla F(x_k)\right\|+\frac{\Lambda_2(1+\kappa)L_{\nabla F}\nu\sqrt{d}}{2},\end{aligned}\tag{41}$$

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 ν. As discussed in Section 2.2, in any practical implementation one has to account for the numerical errors associated with numerical evaluations of the function values. Therefore, from (20) and (21), we have

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

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 κ 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 α_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 α_k falls below some threshold $\alpha_{\min} > 0$, then we ignore the sufficient decrease condition and choose $\alpha_k = \alpha_{\min}$. 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 $\alpha_k = \alpha_{\min}$ if α_k is attempted to be reduced below $\alpha_{\min}$.

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

$$\|y_k\| \leq M\|s_k\|. \tag{42}$$

The condition (42), along with (26), implies that

$$\frac{y_k^T y_k}{y_k^T s_k} \leq \frac{\|y_k\|^2}{\beta_1 \|s_k\|^2} \leq \frac{M}{\beta_1},$$

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 (Section 5.1) and nonsmooth (Section 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 $\theta_0 = 0.9$, $|S_0| = 2$, finite-difference parameter $\nu = 10^{-8}$, 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 β_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.

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 [34]. We call the method here the *finite-difference stochastic gradient method*, "FD-SG," and g_k is chosen as $\nabla^{\mathrm{FD}} F_{S_k}(x_k)$ 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 *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

$$g_k = \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \frac{d}{T} \sum_{j=1}^{T} \frac{f\left(x + \nu u_j, \zeta_i\right) - f(x, \zeta_i)}{\nu} u_j,$$

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 ν 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 $j \in \{-20, -9, \ldots, 9, 10\}$). 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

$$f_{\mathrm{rel}}(x, \zeta) := \frac{1}{1 + \sigma^2} \sum_{j=1}^{p} \phi_j^2(x) \left(1 + \zeta_j\right)^2,$$

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, both functions are of unbounded support 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 ϕ from the CUTEr [30] collection of optimization problems and used two different σ values $\{10^{-3}, 10^{-5}\}$. The details of these problems are given in Table 1.

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

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 [42]. We computed the minimum function values F^* by running the L-BFGS method on the noise-free (i.e., $\sigma = 0$) 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 σ 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

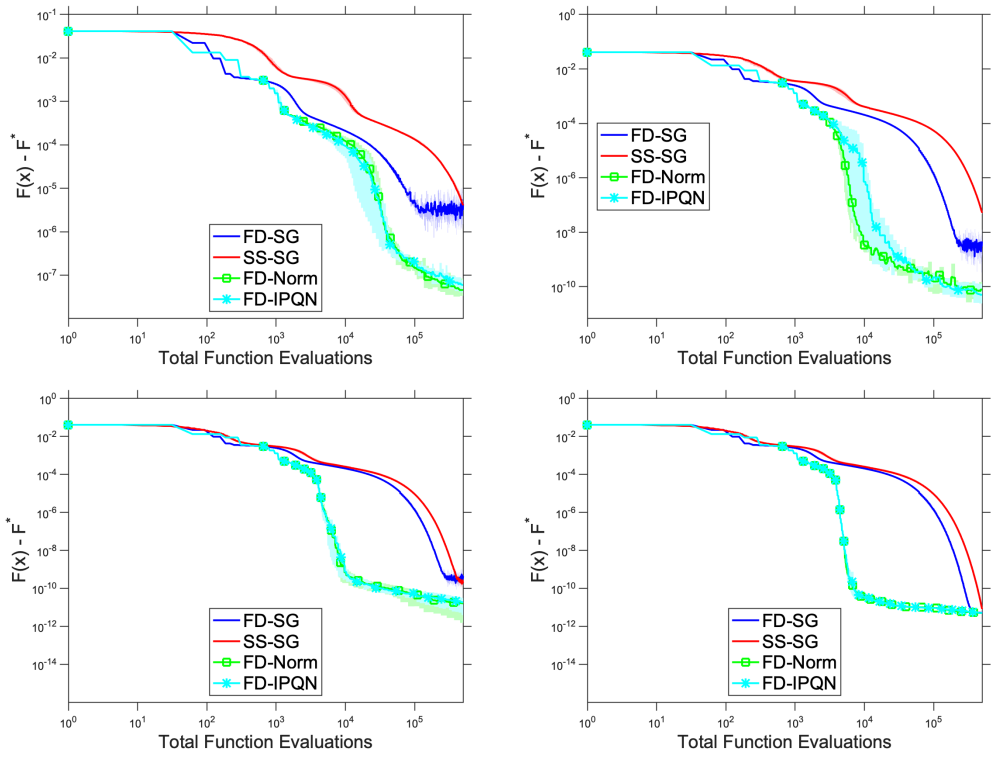

Figure 1: Chebyquad function results based on the total number of f evaluations: Using f_{abs} (left column) and f_{rel} (right column) with $\sigma = 10^{-3}$ (top row) and $\sigma = 10^{-5}$ (bottom row). For each solver, the mean across five random trials is shown; the shaded region indicates the range of performance across these five trials.

the fact that these are low-dimensional problems and the computational savings obtained by sampling only few random directions (recall from Table 1 that $\frac{d}{T}$ 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. 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 Figure 2.

We also report the step lengths chosen at each iteration by the two variants of our algorithm in Figure 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.

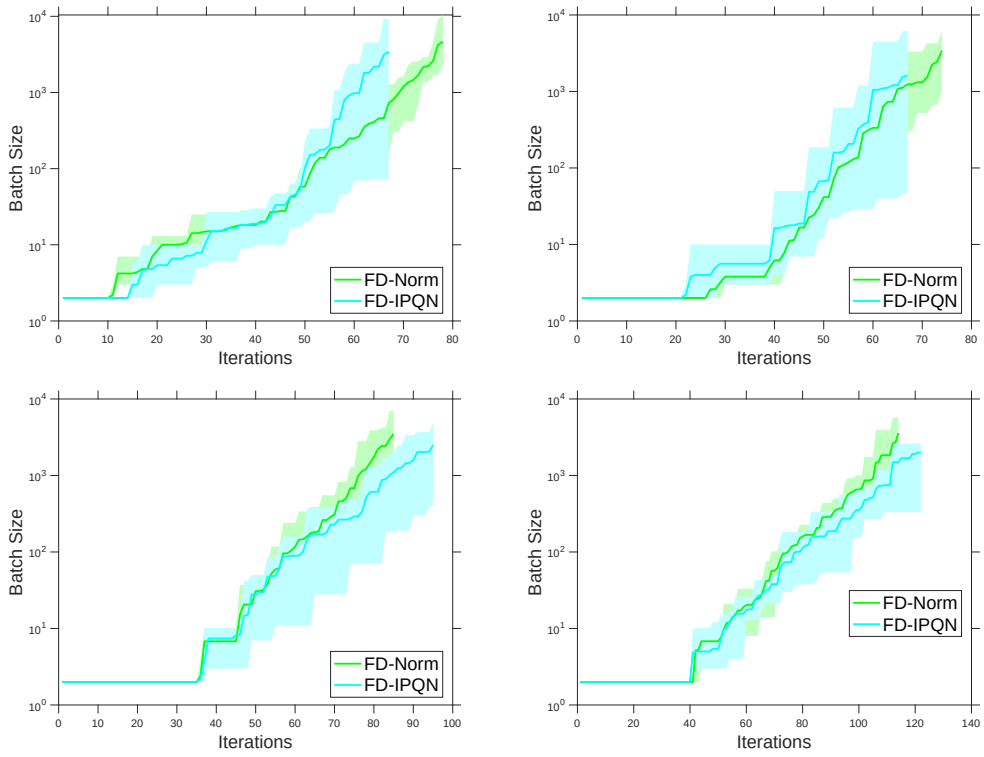

Figure 2: Chebyquad function results showing how the batch size grows over the iterations for which all five trials were running: Using f_{abs} (left column) and f_{rel} (right column) with $\sigma = 10^{-3}$ (top row) and $\sigma = 10^{-5}$ (bottom row).

Results for the other problems listed in Table 1 are given in Appendix B.

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

$$f(x, \zeta) = \|Ax - b - \zeta\|_1 = \sum_{i=1}^{p} \left| a_i^T x - b_i - \zeta_i \right| \tag{43}$$

where $\zeta \in \mathbb{R}^p$ is a uniform random vector $[-1, 1]^p$. We note that the expected function $\mathbb{E}_\zeta [f(\cdot, \zeta)]$ is continuously differentiable and strongly convex; see Appendix C for details. 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 $F^* = 25$.

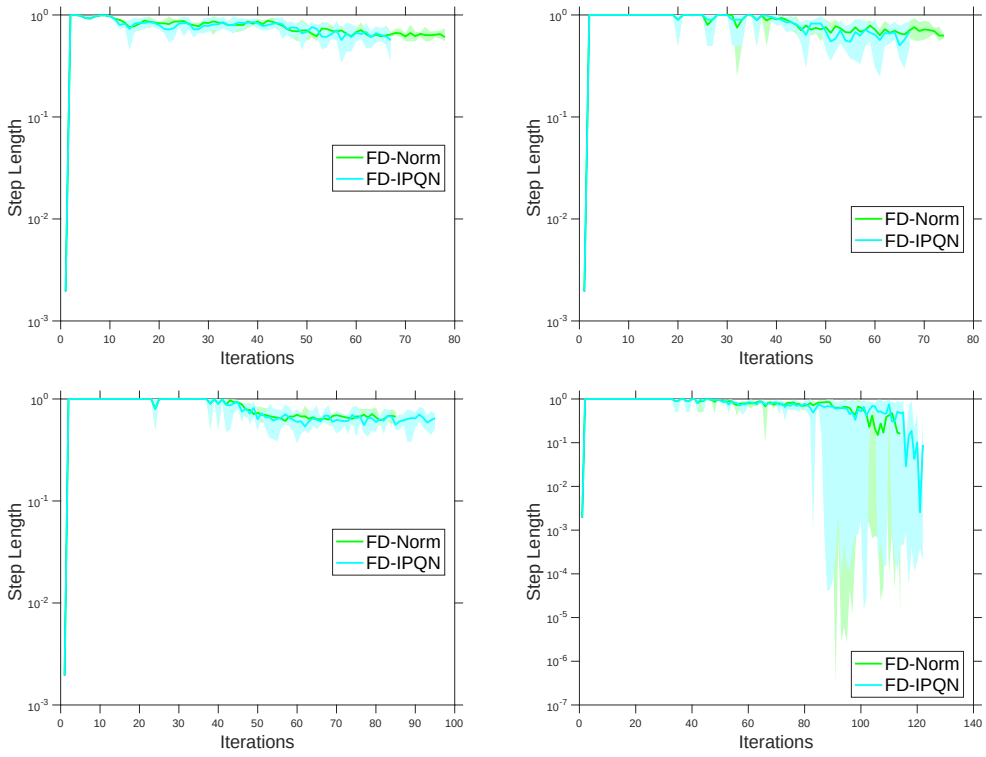

Figure 3: Chebyquad function results showing the accepted step length over the iterations for which all five trials were running: Using f_{abs} (left column) and f_{rel} (right column) with $\sigma = 10^{-3}$ (top row) and $\sigma = 10^{-5}$ (bottom row).

Figure 4 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

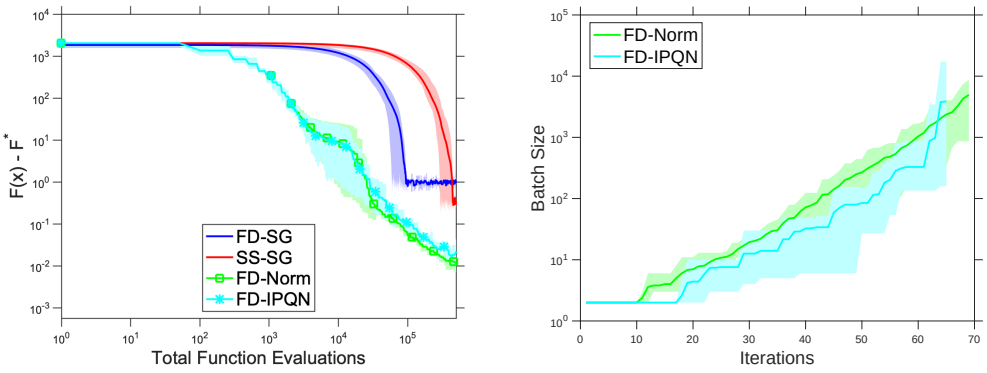

Figure 4: Results for a random instance of the nonsmooth function (43) with $d = p = 50$.

severely limited [14].

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.

Acknowledgments

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 No. DE-AC02-06CH11357.

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). URL 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. Computational Optimization and Applications **79**(1), 1–34 (2021). doi:10.1007/s10589-020-00249-0

[3] Audet, C., Hare, W.L.: Derivative-Free and Blackbox Optimization. Springer (2017). doi:10.1007/978-3-319-68913-5

[4] Balasubramanian, K., Ghadimi, S.: Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds.) Advances in Neural Information Processing Systems 31, pp. 3455–3464. Curran Associates, Inc. (2018). URL http://papers.nips.cc/paper/7605-zeroth-order-non-convex-stochastic-optimization-via-conditional-gradient-and-gpdf

[5] Balasubramanian, K., Ghadimi, S.: Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. Foundations of Computational Mathematics (2021). doi:10.1007/s10208-021-09499-8. To appear

[6] Berahas, A.S., Byrd, R.H., Nocedal, J.: Derivative-free optimization of noisy functions via quasi-Newton methods. SIAM Journal on Optimization **29**(2), 965–993 (2019). doi: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. Foundations of Computational Mathematics (2021). doi:10.1007/s10208-021-09513-z. To appear

[8] Berahas, A.S., Nocedal, J., Takáč, M.: A multi-batch L-BFGS method for machine learning. In: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems 29, pp. 1055–1063. Curran Associates, Inc. (2016). URL http://papers.nips.cc/paper/6145-a-multi-batch-l-bfgs-method-for-machine-learning.pdf

[9] Bertsekas, D.P.: Reinforcement Learning and Optimal Control. Athena Scientific (2019)

[10] Bertsekas, D.P., Nedić, A., Ozdaglar, A.E.: Convex Analysis and Optimization. Athena Scientific (2003)

[11] Blanchet, J., Cartis, C., Menickelly, M., Scheinberg, K.: Convergence rate analysis of a stochastic trust-region method via supermartingales. INFORMS Journal on Optimization **1**(2), 92–119 (2019). doi:10.1287/ijoo.2019.0016

[12] Blum, J.R.: Multidimensional stochastic approximation methods. The Annals of Mathematical Statistics **25**(4), 737–744 (1954). doi:10.1214/aoms/1177728659

[13] Bollapragada, R., Byrd, R., Nocedal, J.: Adaptive sampling strategies for stochastic optimization. SIAM Journal on Optimization **28**(4), 3312–3343 (2018). doi:10.1137/17m1154679

[14] 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. Journal of Physics G: Nuclear and Particle Physics **48**(2), 024001 (2021). doi:10.1088/1361-6471/abd009

[15] 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). URL http://proceedings.mlr.press/v80/bollapragada18a.html

[16] Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Mathematical Programming **134**(1), 127–155 (2012). doi:10.1007/s10107-012-0572-5

[17] Cartis, C., Scheinberg, K.: Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. Mathematical Programming **169**(2), 337–375 (2018)

[18] Chang, K.H.: Stochastic Nelder-Mead simplex method - A new globally convergent direct search method for simulation optimization. European Journal of Operational Research **220**(3), 684–694 (2012). doi:10.1016/j.ejor.2012.02.028

[19] Chen, R., Menickelly, M., Scheinberg, K.: Stochastic optimization using a trust-region method and random models. Mathematical Programming **169**(2), 447–487 (2018). doi:10.1007/s10107-017-1141-8

[20] Chen, X., Kelley, C.T.: Optimization with hidden constraints and embedded Monte Carlo computations. Optimization and Engineering **17**(1), 157–175 (2016). doi:10.1007/s11081-015-9302-1

[21] Chen, X., Kelley, C.T., Xu, F., Zhang, Z.: A smoothing direct search method for Monte Carlo-based bound constrained composite nonsmooth optimization. SIAM Journal on Scientific Computing **40**(4), A2174–A2199 (2018). doi:10.1137/17m1116714

[22] 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: H. Wallach, H. Larochelle, A. Beygelzimer,

F. d'Alché-Buc, E. Fox, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 7204–7215. Curran Associates, Inc. (2019). URL https://proceedings.neurips.cc/paper/2019/file/576d026223582a390cd323bef4bad026-Paper.pdf

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

[24] Deng, G., Ferris, M.C.: Variable-number sample-path optimization. Mathematical Programming **117**, 81–109 (2009). doi:10.1007/s10107-007-0164-y

[25] 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 Transactions on Information Theory **61**(5), 2788–2806 (2015). doi:10.1109/TIT.2015.2409256

[26] 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). doi:10.1109/wsc.2005.1574242

[27] 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. Automation and Remote Control **78**(2), 224–234 (2017). doi:10.1134/S0005117917020035

[28] Ghadimi, S.: Conditional gradient type methods for composite nonlinear and stochastic optimization. Mathematical Programming **173**(1–2), 431–464 (2019). doi:10.1007/s10107-017-1225-5

[29] Ghadimi, S., Lan, G.: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization **23**(4), 2341–2368 (2013). doi:10.1137/120880811

[30] Gould, N.I.M., Orban, D., Toint, P.L.: CUTEr and SifDec: A constrained and unconstrained testing environment, revisited. ACM Transactions on Mathematical Software **29**(4), 373–394 (2003). doi:10.1145/962437.962439

[31] 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 **33**, 1503–1510 (2019). doi:10.1609/aaai.v33i01.33011503

[32] 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). URL http://proceedings.mlr.press/v119/huang20j.html

[33] Kelley, C.T.: Users Guide for imfil version 0.5. Available at www4.ncsu.edu/~ctk/imfil.html

[34] Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics **22**(3), 462–466 (1952). doi:10.1214/aoms/1177729392

[35] Kim, S., Pasupathy, R., Henderson, S.G.: A guide to sample average approximation. In: M. Fu (ed.) Handbook of Simulation Optimization, *International Series in Operations Research & Management Science*, vol. 216, pp. 207–243. Springer (2015). doi:10.1007/978-1-4939-1384-8_8

[36] Kleinman, N.L., Spall, J.C., Naiman, D.Q.: Simulation-based optimization with stochastic approximation using common random numbers. Management Science **45**(11), 1570–1578 (1999). doi:10.1287/mnsc.45.11.1570

[37] Larson, J., Billups, S.C.: Stochastic derivative-free optimization using a trust region framework. Computational Optimization and Applications **64**(3), 619–645 (2016). doi:10.1007/s10589-016-9827-z

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

[39] L'Ecuyer, P., Yin, G.: Budget-dependent convergence rate of stochastic approximation. SIAM Journal on Optimization **8**(1), 217–247 (1998). doi:10.1137/S1052623495270723

[40] Liu, S., Kailkhura, B., Chen, P.Y., Ting, P., Chang, S., Amini, L.: Zeroth-order stochastic variance reduction for nonconvex optimization. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds.) Advances in Neural Information Processing Systems 31, pp. 3731–3741. Curran Associates, Inc. (2018). URL http://papers.nips.cc/paper/7630-zeroth-order-stochastic-variance-reduction-for-nonconvex-optimization.pdf

[41] Mania, H., Guy, A., Recht, B.: Simple random search of static linear policies is competitive for reinforcement learning. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett

(eds.) Advances in Neural Information Processing Systems 31, pp. 1800–1809. Curran Associates, Inc. (2018). URL http://papers.nips.cc/paper/7451-simple-random-search-of-static-linear-policies-is-competitive-for-reinforceme pdf

[42] Moré, J.J., Wild, S.M.: Benchmarking derivative-free optimization algorithms. SIAM Journal on Optimization **20**(1), 172–191 (2009). doi:10.1137/080724083

[43] Moré, J.J., Wild, S.M.: Estimating computational noise. SIAM Journal on Scientific Computing **33**(3), 1292–1314 (2011). doi:10.1137/100786125

[44] Moré, J.J., Wild, S.M.: Estimating derivatives of noisy simulations. ACM Transactions on Mathematical Software **38**(3), 19:1–19:21 (2012). doi:10.1145/2168773.2168777

[45] Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Foundations of Computational Mathematics **17**(2), 527–566 (2017). doi:10.1007/s10208-015-9296-2

[46] Nocedal, J., Wright, S.J.: Numerical Optimization, second edn. Springer (2006). doi:10.1007/978-0-387-40065-5

[47] Pasupathy, R., Ghosh, S.: Simulation optimization: A concise overview and implementation guide. In: Theory Driven by Influential Applications, pp. 122–150. INFORMS (2013). doi:10.1287/educ.2013.0118

[48] Pasupathy, R., Glynn, P., Ghosh, S., Hashemi, F.S.: On sampling rates in simulation-based recursions. SIAM Journal on Optimization **28**(1), 45–73 (2018). doi:10.1137/140951679

[49] Powell, M.J.: Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In: R.W. Cottle, C.E. Lemke (eds.) Nonlinear Programming, *SIAM-AMS Proceedings*, vol. 9, pp. 53–72 (1976)

[50] Robbins, H., Monro, S.: A stochastic approximation method. The Annals of Mathematical Statistics **22**(3), 400–407 (1951). doi:10.1214/aoms/1177729586

[51] 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). URL http://proceedings.mlr.press/v89/sahu19a.html

[52] Salimans, T., Ho, J., Chen, X., Sidor, S., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. Tech. Rep. 1703.03864, ArXiv (2017). URL https://arxiv.org/abs/1703.03864

[53] Shamir, O.: An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. Journal of Machine Learning Research **18**(52), 1–11 (2017). URL http://jmlr.org/papers/v18/16-632.html

[54] Shashaani, S., Hashemi, F.S., Pasupathy, R.: ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. SIAM Journal on Optimization **28**(4), 3145–3176 (2018). doi:10.1137/15m1042425

[55] 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). doi:10.1109/wsc.2016.7822121

[56] Wibisono, A., Wainwright, M.J., Jordan, M.I., Duchi, J.C.: Finite sample convergence rates of zero-order stochastic optimization methods. In: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 25, pp. 1439–1447. Curran Associates, Inc. (2012). URL http://papers.nips.cc/paper/4550-finite-sample-convergence-rates-of-zero-order-stochastic-optimization-methods.pdf

[57] Xie, Y., Byrd, R.H., Nocedal, J.: Analysis of the BFGS method with errors. SIAM Journal on Optimization **30**(1), 182–209 (2020). doi:10.1137/19m1240794

A Supplementary Proofs

Here we collect proofs of several intermediate results.

A.1 Bounded Variances in (12)

The left-hand side of (12) is difficult to compute but can be bounded by the true variance of individual finite-difference gradient estimators; that is,

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

This bound requires that the true variance is bounded, which is Assumption D. The proof follows from

$$\begin{aligned}
&\mathbb{E}_{\zeta_i}\left[\left\|\nabla^{\mathrm{FD}} F_{\zeta_i}(x_k) - \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right] \\
&\quad= \sum_{j=1}^{d} \mathbb{E}_{\zeta_i}\left[\left(\frac{f(x_k+\nu e_j,\zeta_i) - f(x_k,\zeta_i)}{\nu} - \frac{F(x_k+\nu e_j) - F(x_k)}{\nu}\right)^2\right] \\
&\quad\leq \sum_{j=1}^{d} \mathbb{E}_{\zeta_i}\left[2\left(\frac{f(x_k+\nu e_j,\zeta_i) - F(x_k+\nu e_j)}{\nu}\right)^2 + 2\left(\frac{f(x_k,\zeta_i) - F(x_k)}{\nu}\right)^2\right] \\
&\quad= 2\sum_{j=1}^{d}\left(\mathbb{E}_{\zeta_i}\left[\left(\frac{f(x_k+\nu e_j,\zeta_i) - F(x_k+\nu e_j)}{\nu}\right)^2\right] + \mathbb{E}_{\zeta_i}\left[\left(\frac{f(x_k,\zeta_i) - F(x_k)}{\nu}\right)^2\right]\right) \\
&\quad\leq 2\sum_{j=1}^{d}\left(\frac{\omega_1^2 + \omega_2^2\|\nabla F(x_k+\nu e_j)\|^2}{\nu^2} + \frac{\omega_1^2 + \omega_2^2\|\nabla F(x_k)\|^2}{\nu^2}\right) \\
&\quad\leq 2\sum_{j=1}^{d}\frac{\omega_1^2 + 2\omega_2^2\left(\|\nabla F(x_k+\nu e_j) - \nabla F(x_k)\|^2 + \|\nabla F(x_k)\|^2\right)}{\nu^2} + 2d\frac{\omega_1^2 + \omega_2^2\|\nabla F(x_k)\|^2}{\nu^2} \\
&\quad\leq \frac{4\omega_1^2 d}{\nu^2} + \frac{3\omega_2^2 d\|\nabla F(x_k)\|^2}{\nu^2} + 2\omega_2^2 L_{\nabla F}^2, \qquad (44)
\end{aligned}$$

where the first and third inequalities are due to the fact $(a+b)^2 \leq 2(a^2+b^2)$, the second inequality is due to Assumption D, and the last inequality is due to Assumption B. Therefore, for all iterations k where $\|\nabla F(x_k)\| < \infty$, we have

$$\mathbb{E}_{\zeta_i}\left[\left\|\nabla^{\mathrm{FD}} F_{\zeta_i}(x_k) - \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right] < \infty.$$

In a similar manner, we can show that the true variance of the inner product quasi-Newton condition is also bounded. That is,

$$\begin{aligned}
&\mathbb{E}_{\zeta_i}\left[\left(\left(H_k\nabla^{\mathrm{FD}}F_{\zeta_i}(x_k)\right)^T H_k\nabla^{\mathrm{FD}}F(x_k) - \left\|H_k\nabla^{\mathrm{FD}}F(x_k)\right\|^2\right)^2\right] \\
&\quad= \mathbb{E}_{\zeta_i}\left[\left(\left(H_k\nabla^{\mathrm{FD}}F_{\zeta_i}(x_k) - H_k\nabla^{\mathrm{FD}}F(x_k)\right)^T H_k\nabla^{\mathrm{FD}}F(x_k)\right)^2\right] \\
&\quad\le \mathbb{E}_{\zeta_i}\left[\left\|H_k\left(\nabla^{\mathrm{FD}}F_{\zeta_i}(x_k) - \nabla^{\mathrm{FD}}F(x_k)\right)\right\|^2\left\|H_k\nabla^{\mathrm{FD}}F(x_k)\right\|^2\right] \\
&\quad\le \lambda_{\max}^4(H_k)\,\mathbb{E}_{\zeta_i}\left[\left\|\nabla^{\mathrm{FD}}F_{\zeta_i}(x_k) - \nabla^{\mathrm{FD}}F(x_k)\right\|^2\right]\left\|\nabla^{\mathrm{FD}}F(x_k)\right\|^2 \\
&\quad\le 2\lambda_{\max}^4(H_k)\,\mathbb{E}_{\zeta_i}\left[\left\|\nabla^{\mathrm{FD}}F_{\zeta_i}(x_k) - \nabla^{\mathrm{FD}}F(x_k)\right\|^2\right]\left(\left\|\nabla^{\mathrm{FD}}F(x_k) - \nabla F(x_k)\right\|^2 + \left\|\nabla F(x_k)\right\|^2\right) \\
&\quad\le \lambda_{\max}^4(H_k)\,\mathbb{E}_{\zeta_i}\left[\left\|\nabla^{\mathrm{FD}}F_{\zeta_i}(x_k) - \nabla^{\mathrm{FD}}F(x_k)\right\|^2\right]\left(\left(\frac{L_{\nabla F}\nu}{2}\right)^2 d + \left\|\nabla F(x_k)\right\|^2\right),
\end{aligned}$$

where the third inequality is due to $(a+b)^2 \le 2(a^2+b^2)$, the fifth inequality is due to (7), and $\lambda_{\max}(H_k)$ is the largest eigenvalue of H_k. Therefore, from (44), for all iterations k where $\|\nabla F(x_k)\|^2 < \infty$, we have

$$\mathbb{E}_{\zeta_i}\left[\left(\left(H_k\nabla^{\mathrm{FD}}F_{\zeta_i}(x_k)\right)^T H_k\nabla^{\mathrm{FD}}F(x_k) - \left\|H_k\nabla^{\mathrm{FD}}F(x_k)\right\|^2\right)^2\right] < \infty.$$

Hence,

$$\begin{aligned}
&\mathbb{E}_{S_k}\left[\left(\left(H_k\nabla^{\mathrm{FD}}F_{S_k}(x_k)\right)^T H_k\nabla^{\mathrm{FD}}F(x_k) - \left\|H_k\nabla^{\mathrm{FD}}F(x_k)\right\|^2\right)^2\right] \\
&\qquad\le \frac{\mathbb{E}_{\zeta_i}\left[\left(\left(H_k\nabla^{\mathrm{FD}}F_{\zeta_i}(x_k)\right)^T H_k\nabla^{\mathrm{FD}}F(x_k) - \left\|H_k\nabla^{\mathrm{FD}}F(x_k)\right\|^2\right)^2\right]}{|S_k|}.
\end{aligned}$$

A.2 Nonconstant Step Lengths

Generalizations of Lemma 5, and subsequent lemmas and theorems, that allow for step lengths α_k that vary by iteration are readily available. Below we provide one such generalization of Lemma 5.

Lemma 11. *Suppose Assumption I is satisfied. For any x_0, let $\{x_k : k \in \mathbb{Z}_{++}\}$ be generated by iteration (8) with $\alpha_k > 0$, and with $|S_k|$ chosen such that*

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

for some constants $a_1 > 0$ and $a_2 > 0$. Then,

$$\mathbb{E}\left[F(x_k) - F(x^*)\right] \le \prod_{i=1}^{k}(1-\mu a_1\alpha_k)\left(F(x_0) - F(x^*) - \frac{a_2}{\mu a_1}\right) + \frac{a_2}{\mu a_1} \qquad \forall k \in \mathbb{Z}_{++}.$$

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

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

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

$$\begin{aligned}\mathbb{E}\left[F(x_{k+1}) - F(x^*)\right] - \frac{a_2}{\mu a_1} &\leq (1 - \mu a_1 \alpha_k)\,\mathbb{E}\left[F(x_k) - F(x^*)\right] + a_2\alpha_k - \frac{a_2}{\mu a_1} \\ &= (1 - \mu a_1 \alpha_k)\left(\mathbb{E}\left[F(x_k) - F(x^*)\right] - \frac{a_2}{\mu a_1}\right). \quad (45)\end{aligned}$$

The lemma follows by applying (45) repeatedly through iteration $k \in \mathbb{Z}_+$. □

A.3 Initial Heuristic Step Length Derivation

Because of the stochasticity of the function values f, it is not guaranteed that a decrease in stochastic function realizations can ensure decrease in the true function F. A conservative strategy to address this issue is to choose the initial trial step length to be small enough such that the increase in function values when the stochastic approximations are not good is controlled. Bollapragada et al. [15] proposed a heuristic to choose the initial trial estimate for α_k such that there is a decrease in the expected function value. Following a similar strategy, we derive a heuristic to choose the initial step length as

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

By Assumptions A, B, and D and Lemma 1, for any deterministic α_k we have that

$$\begin{aligned}\mathbb{E}_{S_k}\left[F(x_{k+1})\right] &\leq F(x_k) - \mathbb{E}_{S_k}\left[\alpha_k \left(H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right)^T \nabla F(x_k)\right] \\ &\quad + \frac{L_{\nabla F}}{2}\,\mathbb{E}_{S_k}\left[\alpha_k^2 \left\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k)\right\|^2\right] \\ &= F(x_k) - \alpha_k \nabla^{\mathrm{FD}} F(x_k)^T H_k \nabla F(x_k) + \frac{L_{\nabla F}\alpha_k^2}{2}\left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2 \\ &\quad + \frac{L_{\nabla F}\alpha_k^2}{2}\,\mathbb{E}_{S_k}\left[\left\|H_k \nabla^{\mathrm{FD}} F_{S_k}(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right] \\ &\leq F(x_k) - \alpha_k \nabla^{\mathrm{FD}} F(x_k)^T H_k \nabla F(x_k) \\ &\quad + \frac{L_{\nabla F}\alpha_k^2}{2}\left(1 + \frac{\mathbb{E}_{\zeta_i}\left[\left\|H_k \nabla^{\mathrm{FD}} F_{\zeta_i}(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right]}{|S_k| \|H_k \nabla^{\mathrm{FD}} F(x_k)\|^2}\right)\left\|H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2.\end{aligned}$$

By using $\delta_k = \nabla^{\mathrm{FD}} F(x_k) - \nabla F(x_k)$, $R_k = \frac{\mathbb{E}_{\zeta_i}\left[\left\|H_k \nabla^{\mathrm{FD}} F_{\zeta_i}(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right]}{|S_k| \|H_k \nabla^{\mathrm{FD}} F(x_k)\|^2}$, $\hat{L}_k = L_{\nabla F}(1 + R_k)$, and Assumption G, we have that

$$\begin{aligned}
\mathbb{E}_{S_k}\left[F(x_{k+1})\right] &\leq F(x_k) - \alpha_k (\nabla F(x_k) + \delta_k)^T H_k \nabla F(x_k) + \frac{\hat{L}_k \alpha_k^2}{2} \|H_k(\nabla F(x_k) + \delta_k)\|^2 \\
&= F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\hat{L}_k \alpha_k^2}{2} (\|H_k \nabla F(x_k)\|^2 + \|H_k \delta_k\|^2) \\
&\quad - \alpha_k (H_k^{1/2} \delta_k)^T (I - \hat{L}_k \alpha_k H_k)(H_k^{1/2} \nabla F(x_k)).
\end{aligned}$$

If

$$W_k = I - L_{\nabla F}\left(1 + \frac{\mathbb{E}_{\zeta_i}\left[\left\|H_k \nabla^{\mathrm{FD}} F_{\zeta_i}(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right]}{|S_k| \|H_k \nabla^{\mathrm{FD}} F(x_k)\|^2}\right) \alpha_k H_k$$

is a positive-definite matrix, then we have

$$\begin{aligned}
\mathbb{E}_{S_k}\left[F(x_{k+1})\right] &\leq F(x_k) - \alpha_k \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\tilde{L}_{\nabla F} \alpha_k^2}{2} (\|H_k \nabla F(x_k)\|^2 + \|H_k \delta_k\|^2) \\
&\quad + \frac{\alpha_k}{2}\left((H_k^{1/2} \nabla F(x_k))^T (I - \tilde{L}_{\nabla F} \alpha_k H_k)(H_k^{1/2} \nabla F(x_k))\right) \\
&\quad + \frac{\alpha_k}{2}\left((H_k^{1/2} \delta_k)^T (I - \tilde{L}_{\nabla F} \alpha_k H_k)(H_k^{1/2} \delta_k)\right) \\
&= F(x_k) - \frac{\alpha_k}{2} \nabla F(x_k)^T H_k \nabla F(x_k) + \frac{\alpha_k}{2} \delta_k^T H_k \delta_k,
\end{aligned}$$

where the first inequality is due to the assumption that W_k is a positive-definite matrix, and $2|x^T A y| \leq x^T A x + y^T A y$ for any positive-definite matrix A. Therefore, to obtain a decrease in the expected function value (to a certain neighborhood), the matrix W_k must be positive definite. The only difference between the deterministic case and the stochastic case is the presence of the additional variance term in the matrix W_k. In the deterministic case, for a reasonably good quasi-Newton matrix H_k, one expects that $\alpha_k = 1$ will result in a decrease in the function (to a certain neighborhood), and therefore the initial trial step-length parameter should be chosen to be 1. In the stochastic case, the initial trial value

$$\hat{\alpha}_k = \left(1 + \frac{\mathbb{E}_{\zeta_i}\left[\left\|H_k \nabla^{\mathrm{FD}} F_{\zeta_i}(x_k) - H_k \nabla^{\mathrm{FD}} F(x_k)\right\|^2\right]}{|S_k| \|H_k \nabla^{\mathrm{FD}} F(x_k)\|^2}\right)^{-1}$$

will most likely result in the decrease in expected function value (to a certain neighborhood). However, since this formula involves the expensive computation of the individual matrix-vector products $H_k \nabla^{\mathrm{FD}} F_{\zeta_i}(x_k)$, we approximate the

variance-bias ratio as follows:

$$\hat{\alpha}_k = \left(1 + \frac{\mathrm{Var}_{i \in S_k^v}\left[\nabla^{\mathrm{FD}} F_{\zeta_i}(x_k)\right]}{|S_k| \|\nabla^{\mathrm{FD}} F_{S_k}(x_k)\|^2}\right)^{-1},$$

where $S_k^v \subseteq S_k$.

A.4 Assumption G can be Guaranteed to Hold Algorithmically

Assumption G can be shown to hold both for convex and nonconvex functions by updating H_k only when $y_k^T s_k \geq \beta \|s_k\|_2^2$, where $\beta > 0$ is a predetermined constant [8]. We first provide the following technical lemma, which is similar to Lemma 3.1 in [8].

Lemma 12. *If Assumption C is satisfied, and the quasi-Newton matrix update is skipped whenever one of (26) and (28) is not satisfied, then there exist constants $\Lambda_2 \geq \Lambda_1 > 0$ such that*

$$\Lambda_1 I \preceq H_k \preceq \Lambda_2 I, \qquad \forall k \in \mathbb{Z}_{++}.$$

Proof. From (27) and (28), we have

$$\frac{\|y_k\|^2}{y_k^T s_k} \leq \frac{3L^2}{\beta_1} + \frac{3\nu^2 d}{2\beta_1 \|s_k\|^2} \leq \frac{3L^2}{\beta_1} + \frac{3\nu^2 d}{2\beta_1 \beta_2^2}. \tag{46}$$

From (26), we have

$$\beta_1 \|s_k\|^2 \leq y_k^T s_k \leq \|y_k\| s_k\|,$$

and therefore

$$\|s_k\| \leq \frac{1}{\beta_1} \|y_k\|.$$

It follows that

$$y_k^T s_k \leq \|y_k\| \|s_k\| \leq \frac{1}{\beta_1} \|y_k\|^2$$

and hence

$$\frac{\|y_k\|^2}{y_k^T s_k} \geq \beta_1. \tag{47}$$

Let $\Lambda_l = \beta_1$ and $\Lambda_u = \frac{3L^2}{\beta_1} + \frac{3\nu^2 d}{2\beta_1 \beta_2^2}$. Combining upper bound (46) and lower bound (47), we get

$$\Lambda_l \leq \frac{\|y_k\|^2}{y_k^T s_k} \leq \Lambda_u. \tag{48}$$

The rest of the proof follows directly from the proof of Lemma 3.1 in [8]. We provide it here for the sake of completeness. Now, consider the direct Hessian approximation $B_k = H_k^{-1}$. The limited memory quasi-Newton updating formula is given as follows

1. Set $B_k^{(0)} = \frac{y_k^T y_k}{s_k^T y_k} I$ and $\tilde{m} = \min\{k, m\}$; where m is the memory in L-BFGS.
2. For $i = 0, \ldots, \tilde{m} - 1$ set $j = k - \tilde{m} + i$ and compute

$$B_k^{(i+1)} = B_k^{(i)} - \frac{B_k^{(i)} s_j s_j^T B_k^{(i)}}{s_j^T B_k^{(i)} s_j} + \frac{y_j y_j^T}{y_j^T s_j}.$$

3. Set $B_{k+1} = B_k^{(\tilde{m})}$.

Due to (48), the eigenvalues of the matrices $B_k^{(0)} = \frac{y_k^T y_k}{s_k^T y_k} I$ at the start of the L-BFGS update cycles are bounded above and away from zero, for all k. We now use a Trace-Determinant argument to show that the eigenvalues of B_k are bounded above and away from zero.

Let $Tr(B)$ and $det(B)$ denote the trace and determinant of matrix B, respectively, and set $j_i = k - \tilde{m} + i$. The trace of the matrix B_{k+1} can be expressed as

$$\begin{aligned} Tr(B_{k+1}) &= Tr(B_k^{(0)}) - Tr \sum_{i=1}^{\tilde{m}} \left(\frac{B_k^{(i)} s_{j_i} s_{j_i}^T B_k^{(i)}}{s_{j_i}^T B_k^{(i)} s_{j_i}} \right) + Tr \sum_{i=1}^{\tilde{m}} \frac{y_{j_i} y_{j_i}^T}{y_{j_i}^T s_{j_i}} \\ &\leq Tr(B_k^{(0)}) + \sum_{i=1}^{\tilde{m}} \frac{\|y_{j_i}\|^2}{y_{j_i}^T s_{j_i}} \\ &\leq Tr(B_k^{(0)}) + \tilde{m} \Lambda_u \\ &\leq C_1, \end{aligned} \tag{49}$$

for some constant $C_1 > 0$, where the first inequality is due to positive semi-definiteness of $B_k^{(i)}$ update formula, the second inequality is due to (48) and the last inequality is due to the fact that the eigenvalues of the initial L-BFGS matrix $B_k^{(0)}$ are bounded above and away from zero.

Using a result due to Powell [49], the determinant of the matrix B_{k+1} generated

by the proposed algorithm can be expressed as,

$$\begin{aligned} det(B_{k+1}) &= det(B_k^0)\Pi_{i=1}^{\tilde{m}} \frac{y_{j_i}^T s_{j_i}}{s_{j_i}^T B_k^{(i-1)} s_{j_i}} \\ &= det(B_k^0)\Pi_{i=1}^{\tilde{m}} \frac{y_{j_i}^T s_{j_i}}{s_{j_i}^T s_{j_i}} \frac{s_{j_i}^T s_{j_i}}{s_{j_i}^T B_k^{(i-1)} s_{j_i}} \\ &\geq det(B_k^0) \left(\frac{\beta_1}{C_1}\right)^{\tilde{m}} \\ &\geq C_2, \end{aligned} \tag{50}$$

for some constant $C_2 > 0$, where the first inequality is due to (26) and the fact that the largest eigenvalue of $B_k^{(i)}$ is less than C_1, and the last inequality is due to the fact that the eigenvalues of the initial L-BFGS matrix $B_k^{(0)}$ are bounded above and away from zero.

The trace (49) and determinant (50) inequalities derived above imply that largest eigenvalues of all matrices B_k are bounded above, uniformly, and the smallest eigenvalues of all matrices B_k are bounded away from zero, uniformly. Therefore, the inverse Hessian approximation H_k also has eigenvalues bounded above and away from zero. □

B Additional Numerical Results

Here we include numerical results for the smooth problems in Table 1; see Section 5.1 for further details.

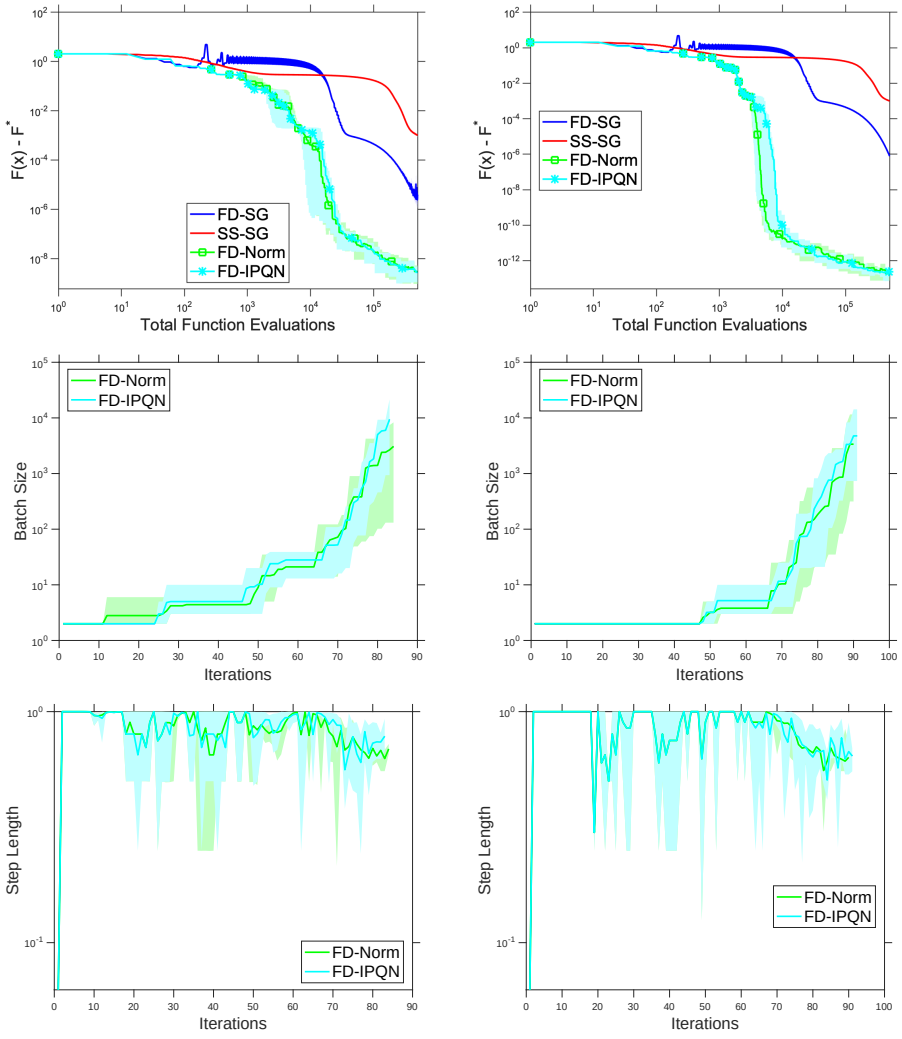

Figure 5: Osborne function ($d = 11$, $p = 65$) results: Using f_{abs} with $\sigma = 10^{-3}$ (left column) and $\sigma = 10^{-5}$ (right column). Top row: $F - F^*$ value versus number of f evaluations. Middle row: Batch size versus number of iterations. Bottom row: Step length versus number of iterations.

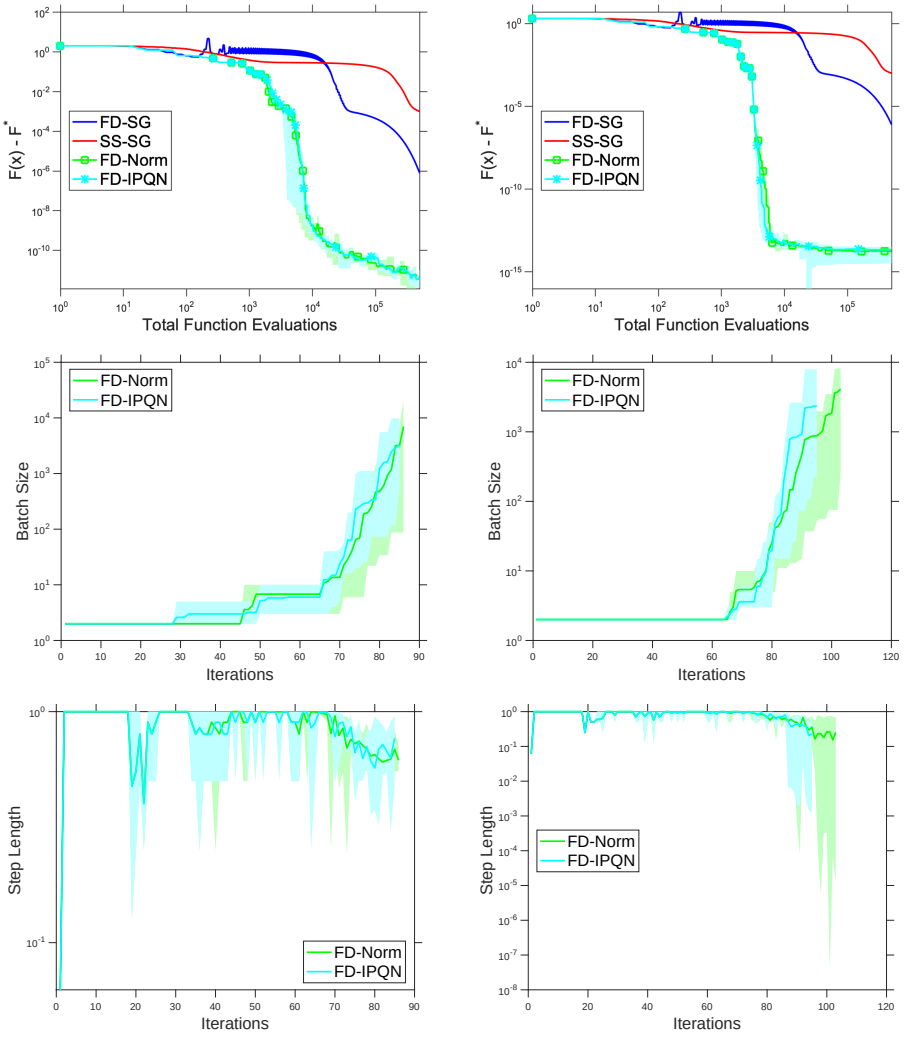

Figure 6: Osborne function ($d = 11$, $p = 65$) results: Using f_{rel} with $\sigma = 10^{-3}$ (left column) and $\sigma = 10^{-5}$ (right column). Top row: $F - F^*$ value versus number of f evaluations. Middle row: Batch size versus number of iterations. Bottom row: Step length versus number of iterations.

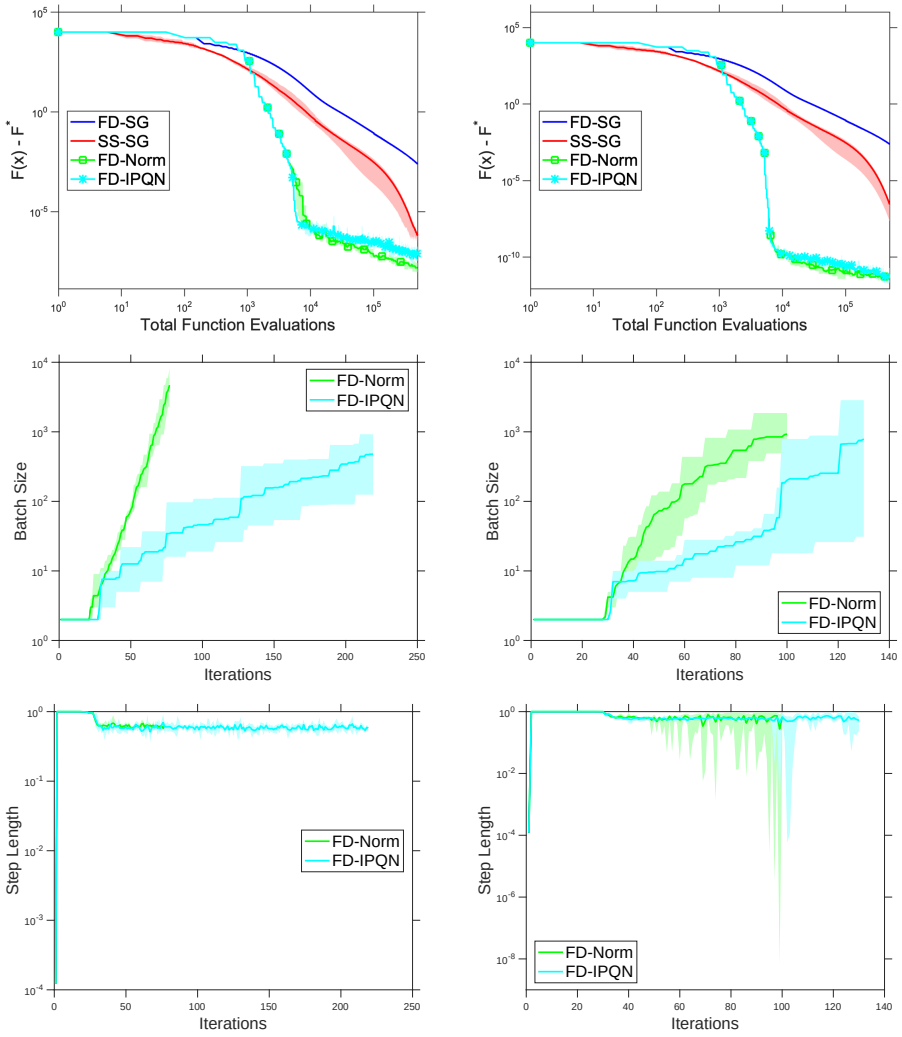

Figure 7: Bdqrtic function ($d = 50$, $p = 92$) results: Using f_{abs} with $\sigma = 10^{-3}$ (left column) and $\sigma = 10^{-5}$ (right column). Top row: $F - F^*$ value versus number of f evaluations. Middle row: Batch size versus number of iterations. Bottom row: Step length versus number of iterations.

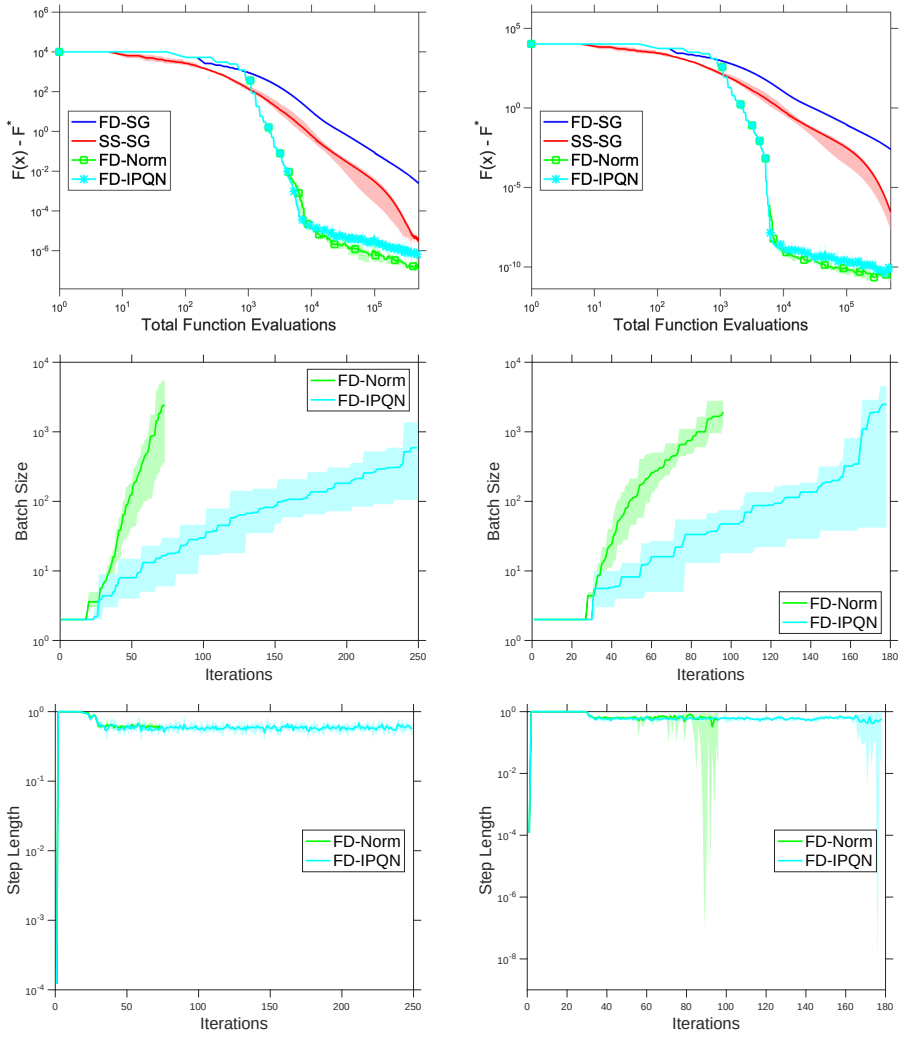

Figure 8: Bdqrtic function ($d = 50$, $p = 92$) results: Using f_{rel} with $\sigma = 10^{-3}$ (left column) and $\sigma = 10^{-5}$ (right column). Top row: $F - F^*$ value versus number of f evaluations. Middle row: Batch size versus number of iterations. Bottom row: Step length versus number of iterations.

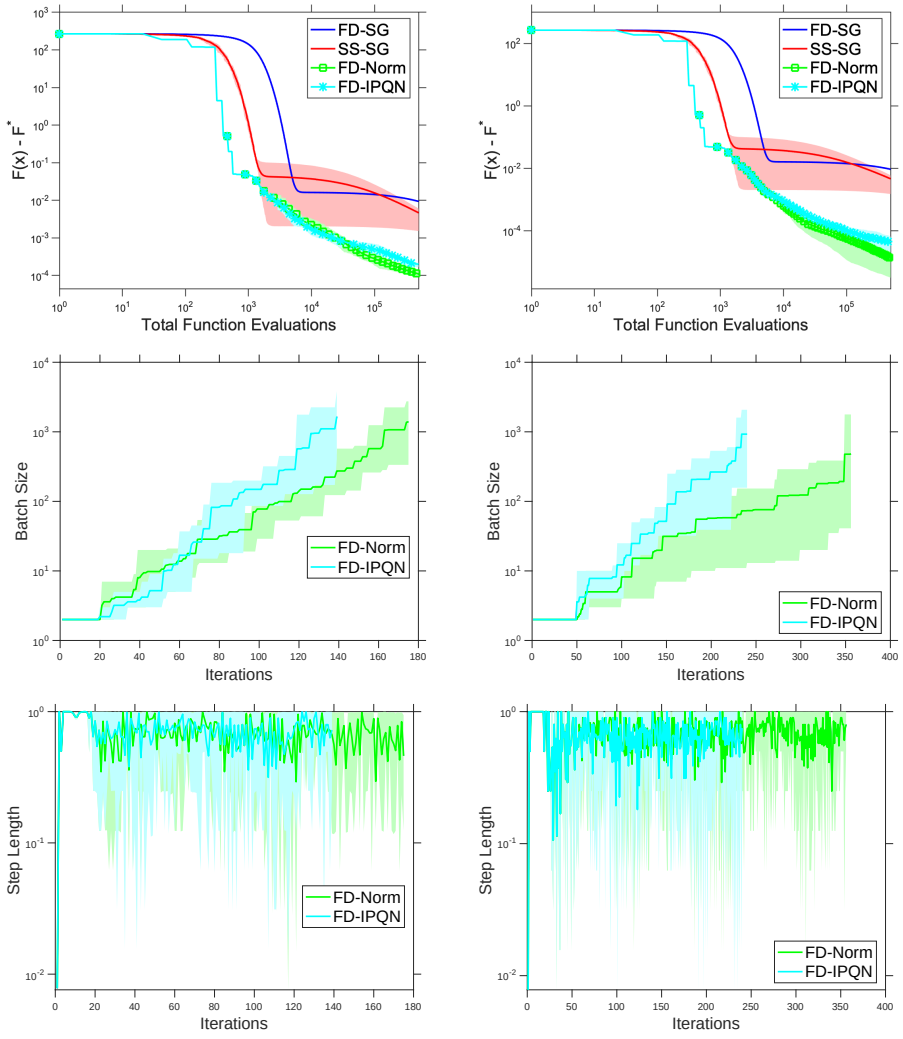

Figure 9: Cube function ($d = 20$, $p = 30$) results: Using f_{abs} with $\sigma = 10^{-3}$ (left column) and $\sigma = 10^{-5}$ (right column). Top row: $F - F^*$ value versus number of f evaluations. Middle row: Batch size versus number of iterations. Bottom row: Step length versus number of iterations.

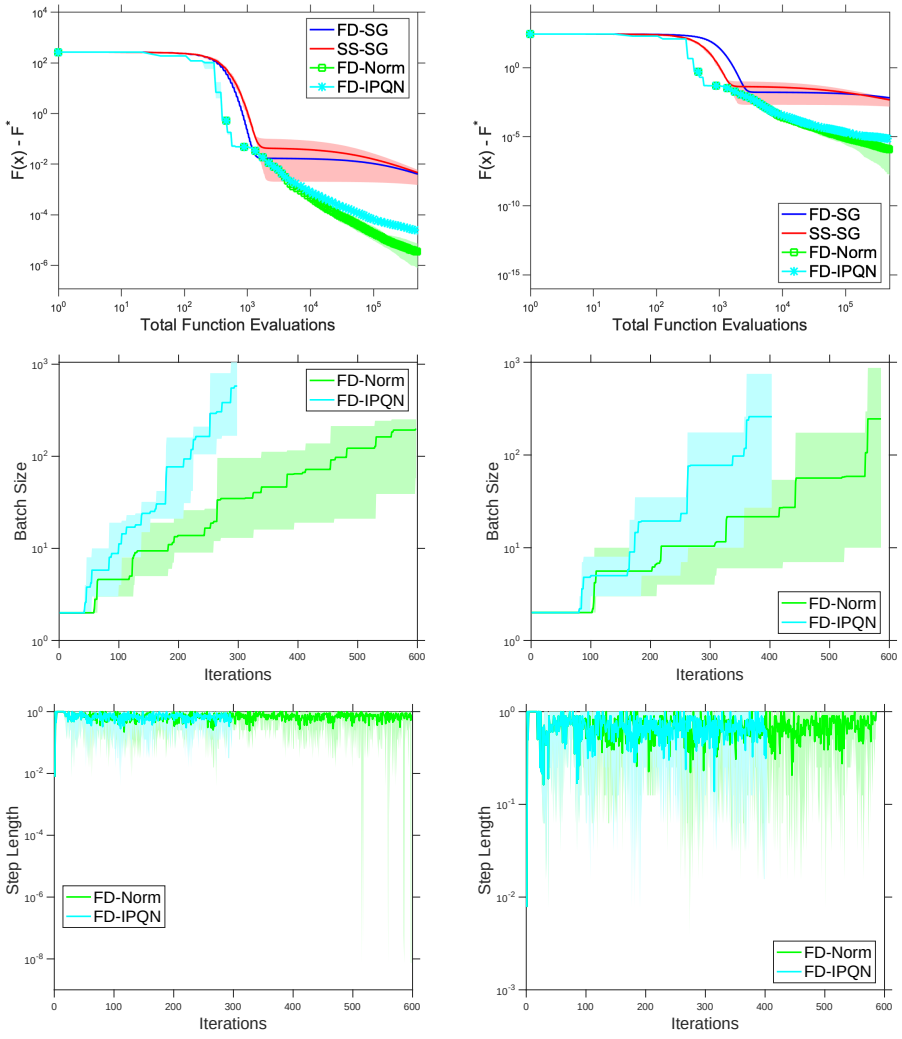

Figure 10: Cube function ($d = 20$, $p = 30$) results: Using f_{rel} with $\sigma = 10^{-3}$ (left column) and $\sigma = 10^{-5}$ (right column). Top row: $F - F^*$ value versus number of f evaluations. Middle row: Batch size versus number of iterations. Bottom row: Step length versus number of iterations.

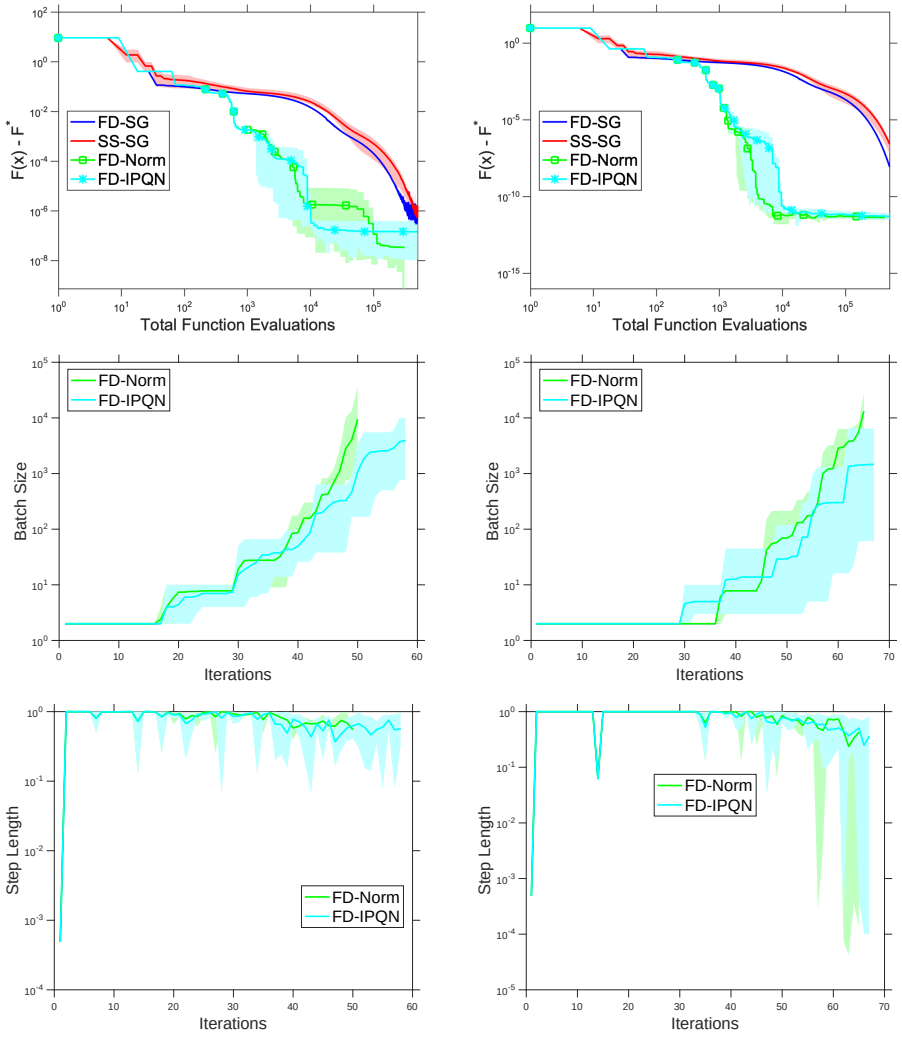

Figure 11: Heart8ls function ($d = 8$, $p = 8$) results: Using f_{abs} with $\sigma = 10^{-3}$ (left column) and $\sigma = 10^{-5}$ (right column). Top row: $F - F^*$ value versus number of f evaluations. Middle row: Batch size versus number of iterations. Bottom row: Step length versus number of iterations.

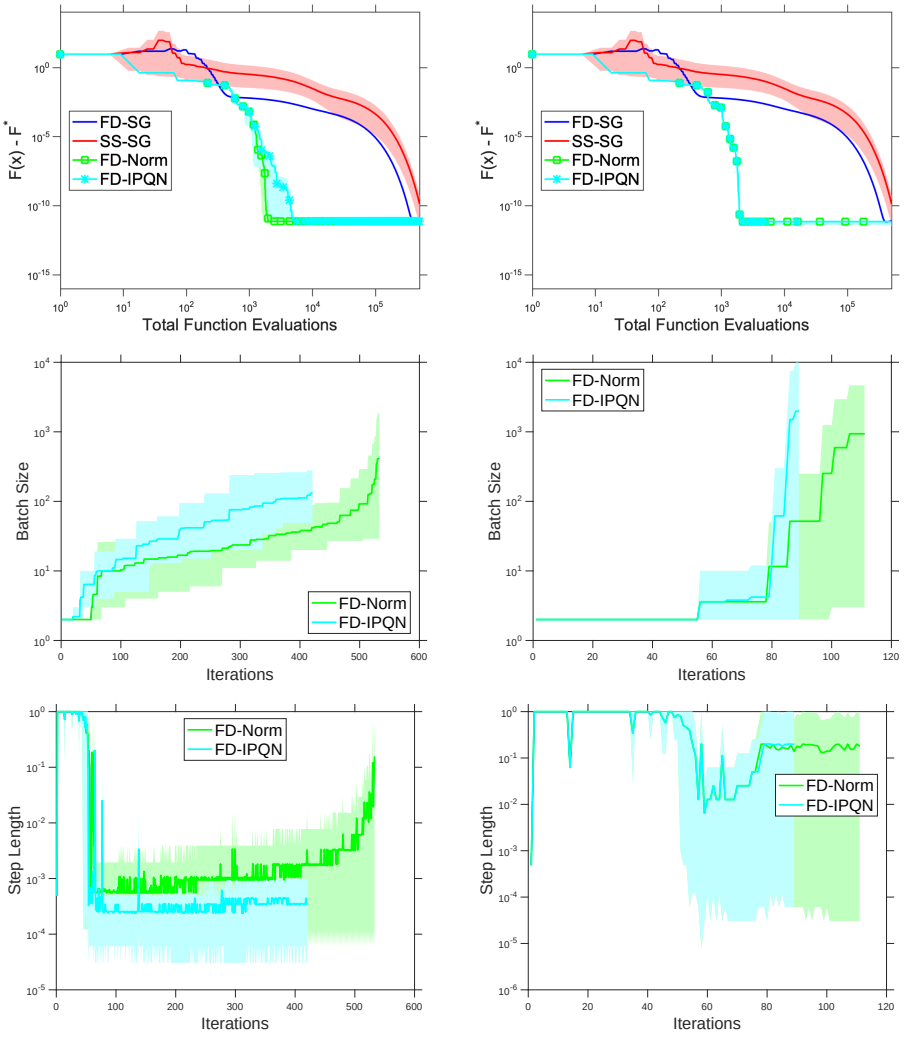

Figure 12: Heart8ls function ($d = 8$, $p = 8$) results: Using f_{rel} with $\sigma = 10^{-3}$ (left column) and $\sigma = 10^{-5}$ (right column). Top row: $F - F^*$ value versus number of f evaluations. Middle row: Batch size versus number of iterations. Bottom row: Step length versus number of iterations.

C Properties of the Nonsmooth Test Function

Here we collect properties of the nonsmooth stochastic function (43) and its expectation

$$F(x) = E_\zeta\left[f(x,\zeta)\right] = \sum_{i=1}^{p} E_{\zeta_i}\left[\left|a_i^T x - b_i - \zeta_i\right|\right] = \frac{1}{2}\sum_{i=1}^{p}\int_{-1}^{1}\left|a_i^T x - b_i - \zeta_i\right| d\zeta_i \tag{51}$$

in the case where $\zeta_1, \ldots, \zeta_p$ are i.i.d. and uniformly distributed over the interval $[-1, 1]$.

Lemma 13. *For any $c \in \mathbb{R}$, when $\zeta \sim Unif[-1,1]$, we have that:*

$$2E_\zeta\left[|c-\zeta|\right] = \int_{-1}^{1}|c-\zeta|\,d\zeta = \begin{cases} c^2+1 & \text{if } |c|\le 1 \\ 2|c| & \text{if } |c|>1. \end{cases} \tag{52}$$

Proof. If $|c| \le 1$, then

$$\begin{aligned}\int_{-1}^{1}|c-\zeta|\,d\zeta &= \int_{-1}^{c}(c-\zeta)\,d\zeta - \int_{c}^{1}(c-\zeta)\,d\zeta \\ &= \left(c^2 - \frac{1}{2}c^2 + c + \frac{1}{2}\right) - \left(c - \frac{1}{2} - c^2 + \frac{1}{2}c^2\right) = c^2+1.\end{aligned}$$

If $c < -1$, then

$$\int_{-1}^{1}|c-\zeta|\,d\zeta = -\int_{-1}^{1}(c-\zeta)\,d\zeta = -\left(c - \frac{1}{2} + c + \frac{1}{2}\right) = -2c.$$

If $c > 1$, then

$$\int_{-1}^{1}|c-\zeta|\,d\zeta = \int_{-1}^{1}(c-\zeta)\,d\zeta = \left(c - \frac{1}{2} + c + \frac{1}{2}\right) = 2c.$$

□

We observe from (52) that

$$\begin{aligned}E_{\zeta_i}\left[\left|a_i^T x - b_i - \zeta_i\right|\right] &= \frac{1}{2}\int_{-1}^{1}\left|a_i^T x - b_i - \zeta_i\right| d\zeta_i \left(\mathbb{I}_{[|a_i^T x - b_i|\le 1]} + \mathbb{I}_{[|a_i^T x - b_i|>1]}\right) \\ &= \frac{1}{2}\left(\left(a_i^T x - b_i\right)^2 + 1\right)\mathbb{I}_{[|a_i^T x - b_i|\le 1]} + \left|a_i^T x - b_i\right|\mathbb{I}_{[|a_i^T x - b_i|>1]},\end{aligned}$$

where $\mathbb{I}_{[\cdot]}$ is the Dirac delta function. Thus,

$$\begin{aligned}\nabla_x E_{\zeta_i}\left[\left|a_i^T x - b_i - \zeta_i\right|\right] &= a_i\left(a_i^T x - b_i\right)\mathbb{I}_{[|a_i^T x - b_i|\le 1]} + a_i \mathrm{sgn}\left[a_i^T x - b_i\right]\mathbb{I}_{[|a_i^T x - b_i|>1]} \\ &= a_i\left(\left(a_i^T x - b_i\right)\mathbb{I}_{[|a_i^T x - b_i|\le 1]} + \mathrm{sgn}\left[a_i^T x - b_i\right]\mathbb{I}_{[|a_i^T x - b_i|>1]}\right)\end{aligned}$$

and, for $|a_i^T x - b_i| < 1$,

$$\nabla^2_{xx} E_{\zeta_i}\left[\left|a_i^T x - b_i - \zeta_i\right|\right] \;=\; a_i a_i^T.$$

As a consequence of the above and from the definition (51) we have thus shown that

$$\begin{aligned}
F(x) &= \sum_{i:\,|a_i^T x - b_i| \le 1} \frac{\left(a_i^T x - b_i\right)^2 + 1}{2} + \sum_{i:\,|a_i^T x - b_i| > 1} \left|a_i^T x - b_i\right| \\
&= \sum_{i=1}^{p} \left(\frac{\left(a_i^T x - b_i\right)^2 + 1}{2} \mathbb{I}_{[|a_i^T x - b_i| \le 1]} + \left|a_i^T x - b_i\right| \mathbb{I}_{[|a_i^T x - b_i| > 1]} \right) \\
\nabla_x F(x) &= \sum_{i:\,|a_i^T x - b_i| \le 1} a_i \left(a_i^T x - b_i\right) + \sum_{i:\,|a_i^T x - b_i| > 1} a_i \mathrm{sgn}\left[a_i^T x - b_i\right] \\
&= \sum_{i=1}^{p} a_i \left(\left(a_i^T x - b_i\right) \mathbb{I}_{[|a_i^T x - b_i| \le 1]} + \mathrm{sgn}\left[a_i^T x - b_i\right] \mathbb{I}_{[|a_i^T x - b_i| > 1]} \right) \\
\nabla^2_{xx} F(x) &= \sum_{i:\,|a_i^T x - b_i| < 1} a_i a_i^T,
\end{aligned}$$

where the last expression is only well defined when there is no $a_i \neq 0$ for which $|a_i^T x - b_i| = 1$. We conclude that F is continuously differentiable.

Furthermore, at any x^* for which $Ax^* = b$, we have that $F(x^*) = \frac{p}{2}$.