Adaptive Sampling Quasi-Newton Methods for Derivative-Free Stochastic Optimization

Raghu Bollapragada
Argonne National Laboratory
Lemont, IL, USA
raghu.bollapragada@utexas.edu

Stefan M. Wild
Argonne National Laboratory
Lemont, IL, USA
wild@anl.gov

Abstract

We consider stochastic zero-order optimization problems, which arise in settings from 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 employ modified versions of a *norm test* and an *inner product quasi-Newton test* to control the sample sizes used in the stochastic approximations. We provide preliminary numerical experiments to illustrate potential performance benefits of the proposed method.

1 Introduction

We consider unconstrained stochastic optimization problems of the form

$$\min_{x \in \Re^d} F(x) = E_\zeta \left[f(x, \zeta) \right], \tag{1}$$

where one has access only to a zero-order oracle (i.e., a black-box procedure that outputs realizations of the stochastic function values $f(x, \zeta)$ and cannot access explicit estimates of the gradient $\nabla_x f(x, \zeta)$). Such stochastic optimization problems arise in a plethora of science and engineering applications, from simulation optimization [4, 9, 11, 15] to reinforcement learning [3, 13, 17]. Several methods have been proposed to solve such "derivative-free" problems [1, 12].

We propose finite-difference 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

$$\left[\nabla F_i^{\text{FD}}(x) \right]_j \quad := \quad \frac{f(x + \nu e_j, \zeta_i) - f(x, \zeta_i)}{\nu}, \ j = 1, \dots, d \tag{2}$$

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

which employ forward differences for the i.i.d. samples of ζ in the set S_k and whereby the parameter ν needs to account only for numerical errors. This gradient estimation has two sources of error: error due to the finite-difference approximation and error due to the stochastic approximation. The error due to stochastic approximation 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 [5, 6, 7, 16]. 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 simple test problems.

2 A Derivative-Free Stochastic Quasi-Newton Algorithm

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

$$x_{k+1} = x_k - \alpha_k H_k \nabla F_{S_k}^{\text{FD}}(x_k), \tag{4}$$

where $\alpha_k > 0$ is the steplength, H_k is a positive-definite quasi-Newton matrix, and $\nabla F_{S_k}^{\text{FD}}(x_k)$ is the (e.g., forward) finite-difference subsampled (or batch) gradient estimate defined by (3). We propose to control the sample sizes $|S_k|$ over the course of optimization to achieve fast convergence by using two different strategies adapted to the setting where no gradient information is available.

Norm test. The first test is based on the norm condition [7, 10]:

$$\mathbb{E}_{S_k}[\|\nabla F_{S_k}^{\text{FD}}(x_k) - \nabla F(x_k)\|^2] \leq \theta^2 \|\nabla F(x_k)\|^2, \tag{5}$$

which is used for controlling the sample sizes in subsampled gradient methods. We note that the finite-difference approximation error (first pair of terms) in $\nabla F_{S_k}^{\text{FD}}(x_k) - \nabla F(x_k) = \nabla F_{S_k}^{\text{FD}}(x_k) - \nabla F_{S_k}(x_k) + \nabla F_{S_k}(x_k) - \nabla F(x_k)$ is nonzero and can be upper bounded (independent of S_k) for any function $f(\cdot, \zeta)$ with L-Lipschitz continuous gradients:

$$\|\nabla F_{S_k}^{\text{FD}}(x_k) - \nabla F_{S_k}(x_k)\|^2 \leq \left(\frac{L\nu\sqrt{d}}{2}\right)^2, \qquad \text{where } \nabla F_{S_k}(x_k) := \frac{\sum_{\zeta_i \in S_k} \nabla_x f(x_k, \zeta_i)}{|S_k|}; \tag{6}$$

the proof is given in the supplementary material. Therefore, one cannot always satisfy (5); moreover, satisfying (5) might be too restrictive. Instead, we propose to look at the norm condition based on the finite-difference subsampled gradient estimators. That is, we use the condition

$$\mathbb{E}_{S_k}[\|\nabla F_{S_k}^{\text{FD}}(x_k) - \nabla F^{\text{FD}}(x_k)\|^2] \leq \theta^2 \|\nabla F^{\text{FD}}(x_k)\|^2, \quad \nabla F^{\text{FD}}(x_k) := \left[\frac{F(x_k + \nu e_j) - F(x_k)}{\nu}\right]_{j=1}^d, \tag{7}$$

for which $\mathbb{E}_{S_k}[\nabla F_{S_k}^{\text{FD}}(x_k)] = \nabla F^{\text{FD}}(x_k)$ and which corresponds to a norm condition where the right-hand side of (5) is relaxed. The left-hand side of (7) is difficult to compute but can be bounded by the true variance of individual finite-difference gradient estimators; this results in

$$\frac{\mathbb{E}_i[\|\nabla F_i^{\text{FD}}(x_k) - \nabla F^{\text{FD}}(x_k)\|^2]}{|S_k|} \leq \theta^2 \|\nabla F^{\text{FD}}(x_k)\|^2. \tag{8}$$

Approximating the true expected gradient and variance with sample gradient and variance estimates, respectively, yields the *practical finite-difference norm test*:

$$\frac{\frac{1}{|S_k^v|-1} \sum_{\zeta_i \in S_k^v} \|\nabla F_i^{\text{FD}}(x_k) - \nabla F_{S_k}^{\text{FD}}(x_k)\|^2}{|S_k|} \leq \theta^2 \|\nabla F_{S_k}^{\text{FD}}(x_k)\|^2, \tag{9}$$

where $S_k^v \subseteq S_k$ is a subset of the current sample (batch). In our algorithm, we test condition (9); whenever it is not satisfied, we control $|S_k|$ such that the condition is satisfied.

Inner product quasi-Newton test. The norm condition controls the variance in the gradient estimation and does not utilize observed quasi-Newton information to control the sample sizes. Recently, Bollapragada et al. [6] 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,

$$(H_k \nabla F_{S_k}^{\text{FD}}(x_k))^T (H_k \nabla F(x_k)) > 0 \tag{10}$$

holds with high probability. This condition can be satisfied in expectation at points that are sufficiently far away from the stationary points; that is, for points x_k such that $\|\nabla F(x_k)\| > \frac{\lambda_{\max}(H_k) L\nu\sqrt{d}}{2\lambda_{\min}(H_k)}$, where $\lambda_{\max}(H_k)$ and $\lambda_{\min}(H_k) > 0$, are the largest and smallest eigenvalues of H_k, respectively (see supplementary material). Hence, the condition (10) can be satisfied with high probability at points x_k that are sufficiently far away from being stationary points. We must control the variance in the left-hand side of (10) to achieve this objective. We note that the quantity $H_k \nabla F(x_k)$ cannot be computed directly; however, it can be approximated by $H_k \nabla F^{\text{FD}}(x_k)$. The condition is given as

$$\mathbb{E}_{S_k}\left[\left((H_k \nabla F_{S_k}^{\text{FD}}(x_k))^T (H_k \nabla F^{\text{FD}}(x_k)) - \|H_k \nabla F^{\text{FD}}(x_k)\|^2\right)^2\right] \leq \theta^2 \|H_k \nabla F^{\text{FD}}(x_k)\|^4, \tag{11}$$

2

where $\mathbb{E}_{S_k}\left[H_k\nabla F_{S_k}^{\text{FD}}(x_k)\right] = H_k\nabla F^{\text{FD}}(x_k)$. The left-hand side of (11) can be bounded by the true variance, as done before. Therefore, the following condition is sufficient for ensuring that (11) holds:

$$\frac{\mathbb{E}_i\left[\left((H_k\nabla F_i^{\text{FD}}(x_k))^T(H_k\nabla F^{\text{FD}}(x_k)) - \|H_k\nabla F^{\text{FD}}(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta^2\|H_k\nabla F^{\text{FD}}(x_k)\|^4. \quad (12)$$

Approximating the true expected gradient and variance with sample gradient and variance estimates results in the *practical finite-difference inner product quasi-Newton test:*

$$\frac{\frac{1}{|S_k^v|-1}\sum_{\zeta_i\in S_k^v}\left((H_k\nabla F_{S_k}^{\text{FD}}(x_k))^T H_k\nabla F_i^{\text{FD}}(x_k) - \|H_k\nabla F_{S_k}^{\text{FD}}\|^2\right)^2}{|S_k|} \leq \theta^2\|H_k\nabla F_{S_k}^{\text{FD}}(x_k)\|^4, \quad (13)$$

where $S_k^v \subseteq S_k$ is a subset of the current sample (batch). This variance computation requires only one additional Hessian-vector product (i.e., the product of H_k with $H_k\nabla F_{S_k}^{\text{FD}}(x_k)$). In our algorithm, we test the condition (13); whenever it is not satisfied, we control $|S_k|$ to satisfy the condition.

Finite-difference parameter and steplength selection. We select the finite-difference parameter ν by minimizing an upper bound on the error in the gradient approximation. Assuming that numerical errors in computing $f(x,\zeta)$ are uniformly bounded by ϵ_m yields the parameter value $\nu^* := 2\sqrt{\frac{\epsilon_m}{L}}$. This finite-difference parameter is analogous to the one derived in [14], which depends on the variance in stochastic noise; however, in the CRN setting we need to account only for numerical errors.

We employ a stochastic line search to choose the steplength α_k by using a sufficient decrease condition based on the sampled function values. In particular, we would like α_k to satisfy

$$F_{S_k}\left(x_k - \alpha_k H_k\nabla F_{S_k}^{\text{FD}}(x_k)\right) \leq F_{S_k}(x_k) - c_1\alpha_k(\nabla F_{S_k}^{\text{FD}}(x_k))^T H_k\nabla F_{S_k}^{\text{FD}}(x_k), \quad (14)$$

where $F_{S_k}(x_k) = \frac{1}{|S_k|}\sum_{\zeta_i\in S_k}f(x_k,\zeta_i)$ and $c_1\in(0,1)$ is a user-specified parameter. We employ a backtracking procedure wherein a trial α_k that does not satisfy (14) is reduced by a fixed fraction $\tau < 1$ (i.e., $\alpha_k \leftarrow \tau\alpha_k$). One cannot always satisfy (14), since the quasi-Newton direction $-H_k\nabla F_{S_k}^{\text{FD}}(x_k)$ may not be a descent direction for the sampled function F_{S_k} at x_k. Intuitively, at points where the error in the sample gradient estimation error $\nabla F_{S_k}^{\text{FD}}(x_k) - \nabla F_{S_k}(x_k)$ dominates the sample gradient itself, measured in terms of the matrix norm $\|\cdot\|_{H_k}$, this condition may not be satisfied, and the line search fails. In practice, stochastic line search failure is an indication that the method has converged to a neighborhood of a solution for the sampled function F_{S_k}, and the solution cannot be further improved. Therefore, one can use the line search failure as an early stopping criterion.

We also note that because of the stochasticity in the function values, it is not guaranteed that a decrease in stochastic function realizations can ensure decrease in the true function. A conservative strategy to address this issue is to choose the initial trial steplength to be small enough such that the increase in function values (when the stochastic approximations are not good) is controlled. Following the strategy proposed in [6], we derive a heuristic to choose the initial steplength as

$$\hat{\alpha}_k = \left(1 + \frac{\text{Var}_{\zeta_i\in S_k^v}\left(\nabla F_i^{FD}(x_k)\right)}{|S_k|\|\nabla F_{S_k}^{\text{FD}}(x_k)\|^2}\right)^{-1}, \quad (15)$$

where $\text{Var}_{\zeta_i\in S_k^v}\left(\nabla F_i^{FD}(x_k)\right) := \frac{\sum_{i\in S_k^v}\|\nabla F_i^{\text{FD}}(x_k)-\nabla F_{S_k}^{\text{FD}}(x_k)\|^2}{|S_k^v|-1}$ is the sample variance used in (9).

Stable quasi-Newton update. In BFGS and L-BFGS methods, the inverse Hessian approximation is updated by using the formulae

$$H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T, \qquad \rho_k = (y_k^T s_k)^{-1}, \qquad V_k = I - \rho_k y_k s_k^T, \quad (16)$$

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, several recent works [2, 6, 18] define y_k as the difference in gradients measured on the same sample S_k to ensure stability in the quasi-Newton approximation. We follow the same approach and define

$$y_k := \nabla F_{S_k}^{\text{FD}}(x_{k+1}) - \nabla F_{S_k}^{\text{FD}}(x_k). \quad (17)$$

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. Therefore, in our tests we skip the quasi-Newton update if the condition $y_k^T s_k > \beta\|s_k\|^2$ is not satisfied with $\beta = 10^{-2}$.
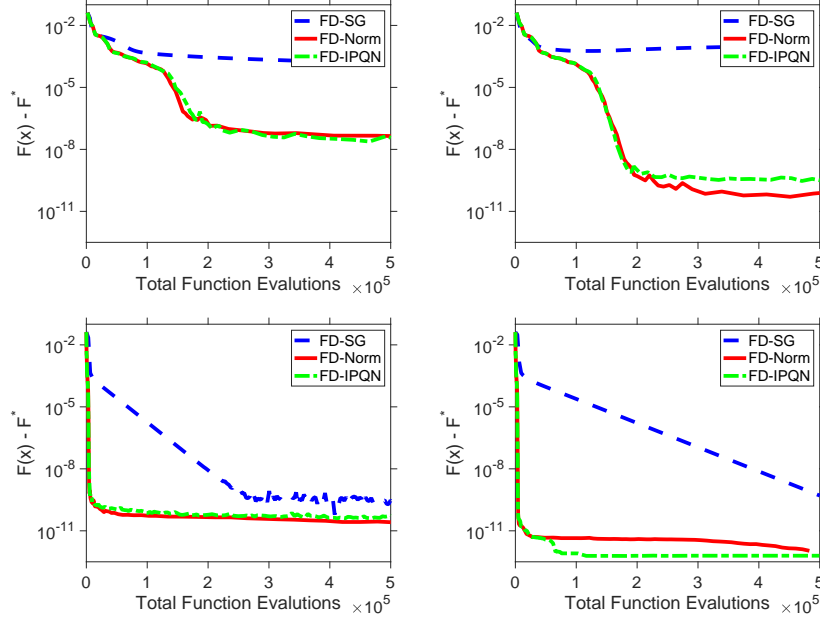
Figure 1: Chebyquad function results for a single run: Using f_{abs} (left column) and f_{rel} (right column) with $\sigma = 10^{-3}$ (top row) and $\sigma = 10^{-5}$ (bottom row).

Convergence results. We leave to future work the settings under which one can establish global convergence results to a neighborhood of an optimal solution for the proposed methods.

3   Numerical Experiments

As a demonstration, we conducted preliminary 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:

$$ f_{\text{rel}}(x, \zeta) := \frac{1}{1+\sigma^2} \sum_{j=1}^{p} \phi_j^2(x)(1+\zeta_j)^2 \qquad \text{and} \qquad f_{\text{abs}}(x, \zeta) := \sum_{j=1}^{p} (\phi_j(x) + \zeta_j)^2 - \sigma^2, \ (18) $$

where $\zeta \sim \mathcal{N}(0, \sigma^2 I_p)$ and $\mathbb{E}_\zeta[f(x, \zeta)] = \sum_{j=1}^{p} \phi_j^2(x)$. In both cases, the function $f(\cdot, \zeta)$ and the expected function $E_\zeta[f(\cdot, \zeta)]$ are twice continuously differentiable. Here we report results only for ϕ defined by the Chebyquad function from [8] with $d = 30$, $p = 45$, and an approximate noise-free value $F^* = 0.0174$.

We implemented two variants, "FD-Norm" and "FD-IPQN", of the proposed algorithm using L-BFGS with $|S_k|$ chosen based on either (9) or (13), respectively. We also implemented a *finite-difference stochastic gradient* method ("FD-SG") $x_{k+1} = x_k - \alpha \nabla F_{S_k}^{FD}(x_k)$, where $\nabla F_{S_k}^{FD}(x_k)$ is defined in (3) and $|S_k| = |S_0|$. We report results for the best version of FD-SG based on tuning the constant steplength for each problem (e.g., considering $\alpha_0 = 2^j$, for $j \in \{-20, -19, \ldots, 9, 10\}$). We use the initial sample $|S_0| = 64$ for larger variance ($\sigma = 10^{-3}$) problems and $|S_0| = 2$ for smaller variance problems ($\sigma = 10^{-5}$) in all the methods.

Figure 1 measures the error in the function, $F(x) - F^*$, 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. Furthermore, the stochastic gradient method converged to a significantly larger neighborhood of the solution as compared with the quasi-Newton variants. No significant difference in performance was observed between the norm test and inner product quasi-Newton test. These preliminary numerical results show that the modified tests have potential for stochastic problems where the CRN approach is feasible.

4

References

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

[2] A. S. Berahas, J. Nocedal, and M. Takác. A multi-batch L-BFGS method for machine learning. In *Advances in Neural Information Processing Systems*, pages 1055–1063, 2016.

[3] D. P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

[4] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust region method via submartingales. *INFORMS Journal on Optimization*, 2019. URL https://arxiv.org/abs/1609.07428. To appear.

[5] R. Bollapragada, R. Byrd, and J. Nocedal. Adaptive sampling strategies for stochastic optimization. *SIAM Journal on Optimization*, 28(4):3312–3343, 2018.

[6] R. Bollapragada, J. Nocedal, D. Mudigere, H. Shi, and P. Tang. A progressive batching l-BFGS method for machine learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 620–629, 2018.

[7] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu. Sample size selection in optimization methods for machine learning. *Mathematical Programming*, 134(1):127–155, 2012.

[8] R. Fletcher. Function minimization without evaluating derivatives – a review. *The Computer Journal*, 8:33–41, 1965. doi: 10.1093/comjnl/8.1.33.

[9] M. C. Fu, F. W. Glover, and J. April. Simulation optimization: A review, new developments, and applications. In *Proceedings of the Winter Simulation Conference*. IEEE, 2005. doi: 10.1109/wsc.2005.1574242.

[10] F. S. Hashemi, S. Ghosh, and R. Pasupathy. On adaptive sampling rules for stochastic recursions. In *Proceedings of the Winter Simulation Conference 2014*, pages 3959–3970. IEEE, 2014.

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

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

[13] H. Mania, A. Guy, and B. Recht. Simple random search of static linear policies is competitive for reinforcement learning. In *Advances in Neural Information Processing Systems 31*, pages 1800–1809. Curran Associates, Inc., 2018.

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

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

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

[17] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. Technical Report 1703.03864, ArXiv, 2017. URL https://arxiv.org/abs/1703.03864.

[18] N. N. Schraudolph, J. Yu, and S. Günter. A stochastic quasi-Newton method for online convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 436–443, 2007.

Supplementary Material

Finite-difference approximation error. For any $\nu > 0$ and any function $f(\cdot, \zeta)$ with L-Lipschitz continuous gradients, we have that

$$\|\nabla F_{S_k}^{\text{FD}}(x_k) - \nabla F_{S_k}(x_k)\|^2 = \sum_{j=1}^{d} \left(\frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \left(\frac{f(x_k + \nu e_j, \zeta_i) - f(x_k, \zeta_i)}{\nu} - [\nabla_x f(x_k, \zeta_i)]_j \right) \right)^2$$

$$\leq \sum_{j=1}^{d} \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \left(\frac{f(x_k + \nu e_j, \zeta_i) - f(x_k, \zeta_i)}{\nu} - [\nabla_x f(x_k, \zeta_i)]_j \right)^2$$

$$\leq \sum_{j=1}^{d} \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \left(\frac{L\nu}{2} \right)^2 = \left(\frac{L\nu}{2} \right)^2 d,$$

where the first equality is by the definitions of $\nabla F_{S_k}^{\text{FD}}(x_k)$ and $\nabla F_{S_k}(x_k) = \frac{1}{|S_k|} \sum_{\zeta_i \in S_k} \nabla_x f(x, \zeta_i)$. The first inequality is due to the fact that $(a_1 + a_2 + \cdots + a_n)^2 \leq n(a_1^2 + a_2^2 + \cdots + a_n^2)$, and the second inequality is due to

$$f(y, \zeta) \leq f(x, \zeta) + (y - x)^T \nabla_x f(x, \zeta) + \frac{L}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

In a similar manner, we can show that, for the true gradients, the error due to finite-difference approximation is bounded. That is,

$$\|\nabla F^{\text{FD}}(x_k) - \nabla F(x_k)\|^2 = \sum_{j=1}^{d} \left(\frac{F(x + \nu e_j) - F(x)}{\nu} - [\nabla F(x_k)]_j \right)^2$$

$$\leq \left(\frac{L\nu}{2} \right)^2 d. \tag{19}$$

Inner product quasi-Newton expected condition. Consider the following:

$$\mathbb{E}_{S_k}[(H_k \nabla F_{S_k}^{\text{FD}}(x_k))^T (H_k \nabla F(x_k))] = (H_k \nabla F^{\text{FD}}(x_k))^T (H_k \nabla F(x_k))$$

$$= (H_k (\nabla F^{\text{FD}}(x_k) - \nabla F(x_k)))^T (H_k \nabla F(x_k)) + \|H_k \nabla F(x_k)\|^2,$$

where we used

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

and we have,

$$(H_k (\nabla F^{\text{FD}}(x_k) - \nabla F(x_k)))^T (H_k \nabla F(x_k))$$

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

$$\leq \|H_k \nabla F(x_k)\| \|H_k\| \frac{L\nu\sqrt{d}}{2},$$

where the last inequality is due to (19). Therefore,

$$\mathbb{E}_{S_k}[(H_k \nabla F_{S_k}^{\text{FD}}(x_k))^T (H_k \nabla F(x_k))] \geq \|H_k \nabla F(x_k)\|^2 - \|H_k \nabla F(x_k)\| \|H_k\| \frac{L\nu\sqrt{d}}{2}$$

$$\geq \|H_k \nabla F(x_k)\| \left(\|H_k \nabla F(x_k)\| - \|H_k\| \frac{L\nu\sqrt{d}}{2} \right).$$

For any x_k such that $\|\nabla F(x_k)\| > \frac{\lambda_{\max}(H_k) L\nu\sqrt{d}}{2\lambda_{\min}(H_k)}$, where $\lambda_{\max}(H_k), \lambda_{\min}(H_k) > 0$ are the largest and smallest eigenvalues of H_k, respectively, it thus follows that

$$\mathbb{E}_{S_k} \left[(H_k \nabla F_{S_k}^{\text{FD}}(x_k))^T (H_k \nabla F(x_k)) \right] > 0.$$

Numerical experiments setup. In the tests of the proposed algorithm, we use $\theta = 0.9$, finite-difference parameter $\nu = 10^{-8}$, L-BFGS memory parameter $m = 10$, and line search parameters ($c_1 = 10^{-4}$, $\tau = 0.5$). None of these parameters have been tuned to the problems being considered.