Stochastic Newton methods with enhanced Hessian estimation

D. Sai Koti Reddy†, Prashanth L.A.♯, Shalabh Bhatnagar‡

***Abstract*—We propose enhancements to the Hessian estimation scheme used in two recently proposed stochastic Newton methods, based on the ideas of random directions stochastic approximation (N-RDSA-3) [1] and simultaneous perturbation stochastic approximation (N-SPSA-3) [2], respectively[1]. The proposed scheme, inspired by [3], reduces the error in the Hessian estimate by (i) incorporating a zero-mean feedback term; and (ii) optimizing the step-sizes used in the Hessian recursion. We prove that N-RDSA3 and N-SPSA-3 with our Hessian improvement scheme converges asymptotically to the true Hessian. The advantage with N-RDSA-3 and N-SPSA-3 is that it requires only 75% of the simulation cost per-iteration for N-SPSA-4 with improved Hessian estimation (N-SPSA-4-IH) [3]. Numerical experiments show that N-RDSA-3-IH outperforms both N-SPSA-4-IH and N-RDSA-3 without the improved Hessian estimation scheme.**

***Index Terms*—Stochastic optimization, stochastic approximation, random directions stochastic approximation (RDSA), simultaneous perturbation stochastic approximation (SPSA).**

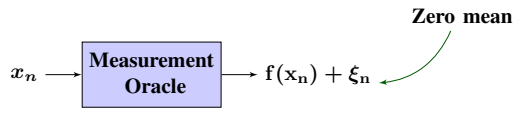


Fig. 1: Simulation optimization

TABLE I: A taxonomy of proposed algorithms.

Algorithm	Function evaluations
N-RDSA-3	$x_n, x_n \pm \delta_n d_n$
N-SPSA-3	$x_n, x_n + \delta_n \Delta_n + \delta_n \widehat{\Delta}_n, x_n - \delta_n \Delta_n - \delta_n \widehat{\Delta}_n$
N-SPSA-4	$x_n \pm \delta_n \Delta_n, x_n \pm \delta_n \Delta_n + \widehat{\delta}_n \widehat{\Delta}_n$

I. Introduction

We consider the following *simulation optimization* problem:

$$\text{Find } x^* = \arg\min_{x \in \mathbb{R}^N} f(x). \tag{1}$$

The aim is to find an algorithm that solves (1) given only noise-corrupted measurements of the objective f, as illustrated in Figure 1. A natural solution approach is to devise an algorithm that incrementally updates the parameter, say x_n, in the descent direction using the gradient and/or Hessian of the objective f. However, gradient/Hessian of f are unavailable directly and need to be estimated from samples of f. To overcome this, we adopt the simultaneous perturbation (SP) approach - a popular and efficient idea especially in high dimensional problems - see [4] for a comprehensive treatment of this subject matter.

Simultaneous perturbation stochastic approximation (SPSA) is a popular SP method. The first-order SPSA algorithm, henceforth referred to as 1SPSA, was proposed in [5]. A closely related algorithm is random directions stochastic approximation (RDSA) [6, pp. 58-60]. The gradient estimate in RDSA differs from that in SPSA, both in the construction as well as in the choice of random perturbations. In [6], the random perturbations for 1RDSA were generated by picking uniformly on the surface of a sphere and the resulting 1RDSA scheme was found to be inferior to 1SPSA from an asymptotic convergence rate viewpoint - see [7]. Recent work in [1] attempts to bridge the gap between 1RDSA and 1SPSA by incorporating random perturbations based on an asymmetric Bernoulli distribution. However, 1SPSA was found to be still marginally better than 1RDSA. On the other hand, results in [1] show that a second order RDSA approach (referred to as N-RDSA-3 hereafter) can considerably outperform the corresponding higher order SPSA algorithm [8] (referred to as N-SPSA-4 hereafter).

Our work in this paper is centered on improving the N-RDSA-3 scheme of [1] by
(i) reducing the error in the Hessian estimate through a feedback term; and
(ii) optimizing the step-sizes used in the Hessian estimation recursion, again with the objective of improving the quality of the Hessian estimate.

Items (i) and (ii) are inspired by the corresponding improvements to the Hessian estimation recursion in N-SPSA-4 - see [3]. While item (ii) is a relatively straightforward migration to the N-RDSA-3 setting, item (i) is a non-trivial contribution, primarily because the Hessian estimate in N-RDSA-3 is entirely different from that in N-SPSA-4 and the feedback term that we incorporate in N-RDSA-3 to improve the Hessian estimate neither correlates with that in N-SPSA-4 nor follows from the analysis in [3]. The advantage with N-RDSA-3-IH is that it requires only 75% of the simulation cost per-iteration for N-SPSA-4-IH.

We establish that the proposed improvements to Hessian estimation in N-RDSA-3 are such that the resulting N-RDSA-3-IH algorithm is provably convergent to the true Hessian. Further, we show empirically that N-RDSA-3-IH that we propose here outperforms both N-SPSA-4-IH of [3] and regular N-RDSA-3 of [1]. Our contribution is important because N-RDSA-3-IH, like N-RDSA-3, has lower simulation cost per iteration than N-SPSA-4 and unlike N-RDSA-3, has an

† Department of Computer Science and Automation, Indian Institute of Science, Bangalore, E-Mail: danda.reddy@csa.iisc.ernet.in
♯ Institute for Systems Research, University of Maryland, College Park, Maryland, E-Mail: prashla@isr.umd.edu.
‡ Department of Computer Science and Automation, Indian Institute of Science, Bangalore, E-Mail: shalabh@csa.iisc.ernet.in

[1]The 3 in the abberviation of the algorithms is used to indicate that the algorithms in [1], [2] require 3 function evaluations per iteration.

Algorithm 1 Structure of N-RDSA-3-IH algorithm.

Input: initial parameter $x_0 \in \mathbb{R}^N$, perturbation constants $\delta_n > 0$, step-sizes $\{a_n, b_n\}$, operator Υ.

for $n = 0, 1, 2, \ldots$ **do**

Generate $\{d_n^i, i = 1, \ldots, N\}$, independent of $\{d_m, m = 0, 1, \ldots, n-1\}$.

For any $i = 1, \ldots, N$, d_n^i is distributed either as an asymmetric Bernoulli (see (20)) or Uniform $U[-\eta, \eta]$ for some $\eta > 0$ (see Remark 2).

Function evaluation 1

Obtain $y_n^+ = f(x_n + \delta_n d_n) + \xi_n^+$.

Function evaluation 2

Obtain $y_n^- = f(x_n - \delta_n d_n) + \xi_n^-$.

Function evaluation 3

Obtain $y_n = f(x_n) + \xi_n$.

Newton step

Update the parameter and Hessian as follows:

$$x_{n+1} = x_n - a_n \Upsilon(\overline{H}_n)^{-1} \widehat{\nabla} f(x_n),$$
$$\overline{H}_n = (1 - b_n)\overline{H}_{n-1} + b_n(\widehat{H}_n - \widehat{\Psi}_n),$$

where $\widehat{H}_n$ is chosen either according to (21) or (23).

end for

Return x_n.

improved Hessian estimation scheme.

II. Second-order RDSA with improved Hessian estimation (N-RDSA-3-IH)

The second-order RDSA with improved Hessian estimate performs an update iteration as follows:

$$x_{n+1} = x_n - a_n \Upsilon(\overline{H}_n)^{-1} \widehat{\nabla} f(x_n), \tag{2}$$
$$\overline{H}_n = (1 - b_n)\overline{H}_{n-1} + b_n(\widehat{H}_n - \widehat{\Psi}_n), \tag{3}$$

where $\widehat{\nabla} f(x_n)$ is the estimate of $\nabla f(x_n)$, $\overline{H}_n$ is an estimate of the true Hessian $\nabla^2 f(\cdot)$, $\Upsilon(\cdot)$ projects any matrix onto the set of positive definite matrices and $\{a_n, n \geq 0\}$ is a step-size sequence that satisfies standard stochastic approximation conditions. There are standard procedures such as Cholesky factorization, see [9], for projecting a given square matrix to set of positive definite matrices. Moreover, in the vicinity of a local minimum, one expects the Hessian to be positive definite. In such a case, Υ will represent the identity operator.

The recursion (2) is identical to that in N-RDSA-3, while the Hessian estimation recursion (3) differs as follows:
(i) $\widehat{\Psi}_n$ is a zero-mean feedback term that reduces the error in Hessian estimate; and
(ii) b_n is a general step-size that we optimize to improve the Hessian estimate.

On the other hand, $\widehat{H}_n$ is identical to that in N-RDSA-3, i.e., it estimates the true Hessian in each iteration using 3 function evaluations. For the sake of completeness, we provide below the construction for $\widehat{\nabla} f(x_n)$ and $\widehat{H}_n$ using asymmetric Bernoulli perturbations, before we present the feedback term that reduces the error in $\widehat{H}_n$.

Algorithm 1 presents the pseudocode

*a) **Function evaluations**:* Let y_n, y_n^+ and y_n^- denote the function evaluations at x_n, $x_n + \delta_n d_n$ and $x_n - \delta_n d_n$ respectively, i.e., $y_n = f(x_n) + \xi_n$, $y_n^+ = f(x_n + \delta_n d_n) + \xi_n^+$ and $y_n^- = f(x_n - \delta_n d_n) + \xi_n^-$, where the noise terms ξ_n, ξ_n^+, ξ_n^- satisfy $\mathbb{E}\left[\xi_n^+ + \xi_n^- - 2\xi_n \middle| \mathcal{F}_n, d_n\right] = 0$ with $\mathcal{F}_n = \sigma(x_m, m < n)$ denoting the underlying sigma-field.

Further, $\delta_n, n \geq 0$ is a sequence of diminishing positive real numbers and $d_n = (d_n^1, \ldots, d_n^N)^\intercal$ is the perturbation random vector at instant n with $d_n^i, i = 1, \ldots, N$ given in (20).

*b) **Gradient estimate**:* The RDSA estimate of the gradient $\nabla f(x_n)$ is given by

$$\widehat{\nabla} f(x_n) = \frac{1}{\lambda} d_n \left[\frac{y_n^+ - y_n^-}{2\delta_n}\right], \tag{4}$$

where $\lambda = \mathbb{E}(d_n^i)^2$ and the perturbations d_n^i, $i = 1, \ldots, N$ are i.i.d.

*c) **Hessian estimate**:*

$$\widehat{H}_n = M_n \left(\frac{y_n^+ + y_n^- - 2y_n}{\delta_n^2}\right), \text{ where} \tag{5}$$
$$M_n = \begin{bmatrix} \frac{1}{\kappa}\left((d_n^1)^2 - \lambda\right) & \cdots & \frac{1}{2\lambda^2} d_n^1 d_n^N \\ \frac{1}{2\lambda^2} d_n^2 d_n^1 & \cdots & \frac{1}{2\lambda^2} d_n^2 d_n^N \\ \cdots & \cdots & \cdots \\ \frac{1}{2\lambda^2} d_n^N d_n^1 & \cdots & \frac{1}{\kappa}\left((d_n^N)^2 - \lambda\right) \end{bmatrix},$$

where $\lambda = \mathbb{E}(d_n^i)^2$, $\tau = E(d_n^i)^4$, and $\kappa = \left(\tau - \lambda^2\right)$ for any $i = 1, \ldots, N$.

*d) **Feedback term** $\widehat{\Psi}_n$:* The Hessian estimate $\widehat{H}_n$ can be simplified as follows:

$$\begin{aligned} \widehat{H}_n &= M_n \left(\frac{y_n^+ + y_n^- - 2y_n}{\delta_n^2}\right) \\ &= M_n \left[\left(\frac{f(x_n + \delta_n d_n) + f(x_n - \delta_n d_n) - 2f(x_n)}{\delta_n^2}\right) \right. \\ &\qquad \left. + \left(\frac{\xi_n^+ + \xi_n^- - 2\xi_n}{\delta_n^2}\right)\right] \\ &= M_n \left(d_n^\intercal \nabla^2 f(x_n) d_n + O(\delta_n^2) + \left(\frac{\xi_n^+ + \xi_n^- - 2\xi_n}{\delta_n^2}\right)\right). \end{aligned} \tag{6}$$

For the first term on the RHS above, note that

$$\mathbb{E}[M_n \left(d_n^\intercal \nabla^2 f(x_n) d_n\right) \mid \mathcal{F}_n] = \mathbb{E}\left[M_n \times \left(\sum_{i=1}^{N-1} (d_n^i)^2 \nabla_{ii}^2 f(x_n) + 2\sum_{i=1}^{N} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n)\right) \middle| \mathcal{F}_n\right]. \tag{7}$$

In analyzing the lth diagonal term of the above expression, the following zero-mean term appears (see the proof of Lemma 4 in [1]):

$$\mathbb{E}\left[([M_n]_D)_{l,l} \left(2\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n)\right) \middle| \mathcal{F}_n\right] = 0, \tag{8}$$

where for any matrix M, $[M]_D$ refers to a matrix that retains only the diagonal entries of M and replaces all the remaining entries with zero, and $([M]_D)_{i,j}$ refers to the (i,j)th entry in $[M]_D$. We shall also use $[M]_N$ to refer to a matrix that

retains only the off-diagonal entries of M, while replaces all the diagonal entries with zero.

The term on the LHS in (8), denoted by $\Psi_n^1(\nabla^2 f(x_n))$, can be written in matrix form as follows:

$$\Psi_n^1(\nabla^2 f(x_n)) = [M_n]_D \left(d_n^{\mathrm{T}} [\nabla^2 f(x_n)]_N \, d_n\right). \quad (9)$$

In analyzing the off-diagonal term ((k,l) where $k < l$) of (40), the following zero-mean term appears:

$$\mathbb{E}\left[([M_n]_D)_{k,l} \left(\sum_{i=1}^{N} (d_n^i)^2 \nabla_{ii}^2 f(x_n) \right) \middle| \mathcal{F}_n \right] = 0. \quad (10)$$

The term on the LHS above, denoted by $\Psi_n^2(\nabla^2 f(x_n))$, can be written in matrix form as follows:

$$\Psi_n^2(\nabla^2 f(x_n)) = [M_n]_N \left(d_n^{\mathrm{T}} [\nabla^2 f(x_n)]_D \, d_n\right). \quad (11)$$

From the foregoing, the per-iteration Hessian estimate $\widehat{H}_n$ can be re-written as follows:

$$\begin{aligned}\mathbb{E}\left[\widehat{H}_n \middle| \mathcal{F}_n\right] =& \nabla^2 f(x_n) + \mathbb{E}\left[\Psi_n(\nabla^2 f(x_n)) \middle| \mathcal{F}_n\right] + O(\delta_n^2) \\ &+ \mathbb{E}\left[\left(\frac{\xi_n^+ + \xi_n^- - 2\xi_n}{\delta_n^2} \right) \middle| \mathcal{F}_n \right], \end{aligned} \quad (12)$$

where, for any matrix H,

$$\begin{aligned}\Psi_n(H) &= \Psi_n^1(H) + \Psi_n^2(H) \\ &= [M_n]_D \left(d_n^{\mathrm{T}} [H]_N \, d_n\right) + [M_n]_N \left(d_n^{\mathrm{T}} [H]_D \, d_n\right). \end{aligned} \quad (13)$$

In the RHS of (12), it is easy to see that the second term involving Ψ_n and the last term involving the noise are zero-mean. Moreover, since the noise is bounded by assumption, the last term in (12) vanishes asymptotically at the rate $O(\delta_n^{-2})$. So, the error in estimating the Hessian is due to the second term, which involves the perturbations d_n. This motivates the term $\widehat{\Psi}_n$ in the update rule (2).

Given that we operate in a simulation optimization setting, which implies $\nabla^2 f$ is not known, we construct the feedback term $\widehat{\Psi}_n$ in (2) by using $\overline{H}_{n-1}$ as a proxy for $\nabla^2 f$, i.e.,

$$\widehat{\Psi}_n = \Psi_n(\overline{H}_{n-1}). \quad (14)$$

*e) **Optimizing the step-sizes** b_n:* Unlike the feedback term, adapting the idea of optimizing the step-sizes for N-RDSA-3 is relatively straightforward from the corresponding approach for N-SPSA-4 in [3]. The difference here is that there exists only one N-dimensional perturbation vector d_n in our setting, while N-SPSA-4 required two such vectors. This in turn implies that only the perturbation constant δ_n is needed in optimizing b_n.

The optimal choice for b_n in (3) is the following:

$$b_i = \delta_i^4 / \sum_{j=0}^{i} \delta_j^4. \quad (15)$$

The main idea behind the above choice is provided below. From (12), we can infer that

$$\mathbb{E} \parallel \widehat{H}_n \parallel^2 \leq \frac{C}{\delta_n^4} \text{ for some } C < \infty.$$

This is because the third term in (12) vanishes asymptotically, while the fourth term there dominates asymptotically. Moreover, the noise factors in the fourth term in (12) are bounded above due to (C9) and independent of n, leaving the δ_n^2 term in the denominator there.

So, the optimization problem to be solved at instant n is as follows:

$$\sum_{i=0}^{n} (\tilde{b}_k)^2 \delta_i^{-4}, \text{ subject to} \quad (16)$$

$$\tilde{b}_i \geq 0 \ \forall i \text{ and } \sum_{i=0}^{n} \tilde{b}_i = 1. \quad (17)$$

The optimization variable $\tilde{b}_i$ from the above is related to the Hessian recursion (3) as follows:

$$\overline{H}_n = \sum_{i=0}^{n} \tilde{b}_k (\widehat{H}_i - \widehat{\Psi}_i). \quad (18)$$

The solution to (16) is achieved for $\tilde{b}_i^* = \delta_i^4 / \sum_{j=0}^{n} \delta_j^4, i = 1, \ldots, n$. The optimal choice $\tilde{b}_i^*$ can be translated to the step-sizes b_i, leading to (15).

In [1], the authors suggest two alternatives for the distribution of random perturbations d_n: the asymmetric Bernoulli and uniform.

Remark 1. ***(Asymmetric Bernoulli)***

*f) **Gradient estimate**: The RDSA estimate of the gradient $\nabla f(x_n)$ is given by*

$$\widehat{\nabla} f(x_n) = \frac{1}{\lambda} d_n \left[\frac{y_n^+ - y_n^-}{2\delta_n} \right], \quad (19)$$

where $\lambda = \mathbb{E}(d_n^i)^2 = (1+\epsilon)$ and the perturbations d_n^i, $i = 1, \ldots, N$ are i.i.d. and distributed as follows:

$$d_n^i = \begin{cases} -1 & \text{w.p. } \dfrac{1+\epsilon}{(2+\epsilon)}, \\ 1+\epsilon & \text{w.p. } \dfrac{1}{(2+\epsilon)}, \end{cases} \quad (20)$$

with $\epsilon > 0$ being a constant that can be chosen to be arbitrarily small.

*g) **Hessian estimate**:*

$$\widehat{H}_n = M_n \left(\frac{y_n^+ + y_n^- - 2y_n}{\delta_n^2} \right), \text{ where} \quad (21)$$

$$M_n = \begin{bmatrix} \frac{1}{\kappa}\left((d_n^1)^2 - \lambda\right) & \cdots & \frac{1}{2\lambda^2} d_n^1 d_n^N \\ \frac{1}{2\lambda^2} d_n^2 d_n^1 & \cdots & \frac{1}{2\lambda^2} d_n^2 d_n^N \\ \cdots & \cdots & \cdots \\ \frac{1}{2\lambda^2} d_n^N d_n^1 & \cdots & \frac{1}{\kappa}\left((d_n^N)^2 - \lambda\right) \end{bmatrix},$$

where $\lambda = \mathbb{E}(d_n^i)^2 = (1+\epsilon)$, $\tau = E(d_n^i)^4 = \dfrac{(1+\epsilon)(1+(1+\epsilon)^3)}{(2+\epsilon)}$, and $\kappa = \left(\tau - \lambda^2\right)$ for any $i = 1, \ldots, N$.

Remark 2. ***(Uniform perturbations)*** *Choose d_n^i, $i = 1, \ldots, N$ to be i.i.d. $U[-\eta, \eta]$ for some $\eta > 0$, where $U[-\eta, \eta]$ denotes*

the uniform distribution on the interval $[-\eta, \eta]$. *Then, the RDSA estimate of the gradient is given by*

$$\widehat{\nabla} f(x_n) = \frac{1}{\lambda} d_n \left[\frac{y_n^+ - y_n^-}{2\delta_n} \right]. \tag{22}$$

The Hessian estimate in this case is given by

$$\widehat{H}_n = M_n \left(\frac{y_n^+ + y_n^- - 2y_n}{\delta_n^2} \right), \text{ where} \tag{23}$$

$$M_n = \begin{bmatrix} \frac{1}{\kappa}\left((d_n^1)^2 - \lambda\right) & \cdots & \frac{1}{2\lambda^2} d_n^1 d_n^N \\ \frac{1}{2\lambda^2} d_n^2 d_n^1 & \cdots & \frac{1}{2\lambda^2} d_n^2 d_n^N \\ \cdots & \cdots & \cdots \\ \frac{1}{2\lambda^2} d_n^N d_n^1 & \cdots & \frac{1}{\kappa}\left((d_n^N)^2 - \lambda\right) \end{bmatrix}.$$

where $\lambda = \mathbb{E}(d_n^i)^2 = \frac{\eta^2}{3}$, $\tau = E(d_n^i)^4 = \frac{\eta^4}{5}$ *and* $\kappa = \left(\tau - \lambda^2\right)$ *for any* $i = 1, \ldots, N$.

The feedback term that we proposed is applicable to both asymmetric Bernoulli case and uniform perturbations.

III. Second-order SPSA with improved Hessian estimation (N-SPSA-3-IH)

a) ***Function evaluations****:* Let y_n, y_n^+ and y_n^- denote the function evaluations at x_n, $x_n + \delta_n\Delta_n + \delta_n\widehat{\Delta}_n$ and $x_n - \delta_n\Delta_n - \delta_n\widehat{\Delta}_n$ respectively, i.e., $y_n = f(x_n) + \xi_n$, $y_n^+ = f(x_n + \delta_n\Delta_n + \delta_n\widehat{\Delta}_n) + \xi_n^+$ and $y_n^- = f(x_n - \delta_n\Delta_n - \delta_n\widehat{\Delta}_n) + \xi_n^-$, where the noise terms ξ_n, ξ_n^+, ξ_n^- satisfy $\mathbb{E}\left[\xi_n^+ + \xi_n^- - 2\xi_n | \mathcal{F}_n, \Delta_n, \widehat{\Delta}_n\right] = 0$ with $\mathcal{F}_n = \sigma(x_m, m < n)$ denoting the underlying sigma-field.

Further, $\delta_n, n \geq 0$ is a sequence of diminishing positive real numbers and $\Delta_n = (\Delta_n^1, \ldots, \Delta_n^N)^\intercal$, $\widehat{\Delta}_n = (\widehat{\Delta}_n^1, \ldots, \widehat{\Delta}_n^N)^\intercal$ are the perturbation random vector at instant n with $\Delta_n^i, i = 1, \ldots, N$, and $\widehat{\Delta}_n^i, i = 1, \ldots, N$ are independent identically distributed (i.i.d), mean-zero random varibles having finite inverse movements of order greater than 2 and satisfy the conditions (C16), C(21).

b) ***Gradient estimate****:* The SPSA estimate of the gradient $\nabla f(x_n)$ is given by

$$\widehat{\nabla}_{(i)} f(x_n) = \left[\frac{y_n^+ - y_n^-}{2\delta_n \Delta_n^{(i)}} \right], \tag{24}$$

where the perturbations Δ_n^i, $i = 1, \ldots, N$ are i.i.d. and as mentioned above.

c) ***Hessian estimate****:* The i, jth entry of the Hessian estimate in this case is given by

$$\left(\widehat{H}_n\right)_{ij} = \left(\frac{y_n^+ + y_n^- - 2y_n}{2\delta_n^2 \Delta_n^{(i)} \widehat{\Delta}_n^{(j)}} \right), \tag{25}$$

d) ***Feedback term*** $\widehat{\Psi}_n$*:* The i, jth term of the Hessian estimate $\widehat{H}_n$ can be simplified as follows:

$$\begin{aligned} \left(\widehat{H}_n\right)_{ij} &= \left(\frac{y_n^+ + y_n^- - 2y_n}{2\delta_n^2 \Delta_n^{(i)} \widehat{\Delta}_n^{(j)}} \right) \\ &= \left(\frac{(\Delta_n + \widehat{\Delta}_n)^\intercal \nabla^2 f(x_n)(\Delta_n + \widehat{\Delta}_n)}{2\Delta_n^{(i)} \widehat{\Delta}_n^{(j)}} + O(\delta_n^2) + \right. \\ &\qquad \left. + \left(\frac{\xi_n^+ + \xi_n^- - 2\xi_n}{2\delta_n^2 \Delta_n^{(i)} \widehat{\Delta}_n^{(j)}} \right) \right). \end{aligned} \tag{26}$$

After expanding the first term on the RHS above, we get

$$\begin{aligned} \sum_{l=1}^N \sum_{m=1}^N \frac{\Delta_n^l \nabla_{lm}^2 f(x_n) \Delta_n^m}{2\Delta_n^i \widehat{\Delta}_n^j} + \sum_{l=1}^N \sum_{m=1}^N \frac{\Delta_n^l \nabla_{lm}^2 f(x_n) \widehat{\Delta}_n^m}{\Delta_n^i \widehat{\Delta}_n^j} \\ + \sum_{l=1}^N \sum_{m=1}^N \frac{\widehat{\Delta}_n^l \nabla_{lm}^2 f(x_n) \widehat{\Delta}_n^m}{2\Delta_n^i \widehat{\Delta}_n^j} \end{aligned} \tag{27}$$

It is easy to see that

$$\begin{aligned} \mathbb{E}\left[\sum_{l=1}^N \sum_{m=1}^N \frac{\Delta_n^l \nabla_{lm}^2 f(x_n) \Delta_n^m}{\Delta_n^i \widehat{\Delta}_n^j} \middle| \mathcal{F}_n \right] &= 0 \\ \mathbb{E}\left[\sum_{l=1}^N \sum_{m=1}^N \frac{\widehat{\Delta}_n^l \nabla_{lm}^2 f(x_n) \widehat{\Delta}_n^m}{\Delta_n^i \widehat{\Delta}_n^j} \middle| \mathcal{F}_n \right] &= 0 \\ \mathbb{E}\left[\sum_{l=1}^N \sum_{m=1}^N \frac{\Delta_n^l \nabla_{lm}^2 f(x_n) \widehat{\Delta}_n^m}{\Delta_n^i \widehat{\Delta}_n^j} \middle| \mathcal{F}_n \right] &= \nabla_{ij}^2 f(x_n) \quad a.s.. \end{aligned} \tag{28}$$

From the above equation we know that first and second terms are mean zero terms, they can be represented in matrix form as below

$$\Psi_n^1(\nabla^2 f(x_n)) = \frac{1}{2} M_n \left[\Delta_n^\intercal \nabla^2 f(x_n) \Delta_n + \widehat{\Delta}_n^\intercal \nabla^2 f(x_n) \widehat{\Delta}_n \right] \tag{29}$$

where $M_n = [1/\Delta_n^1, \ldots, 1/\Delta_n^N]^\intercal [1/\Delta_n^1, \ldots, 1/\Delta_n^N]$. Now consider the second term in the (27). It can be written as following

$$\nabla_{ij}^2 f(x_n) + \frac{1}{\Delta_n^i \widehat{\Delta}_n^j} \sum_{\substack{l=1 \\ lm \neq ij}}^N \sum_{m=1}^N \Delta_n^l \nabla_{lm}^2 f(x_n) \widehat{\Delta}_n^m \tag{30}$$

Now in the above equation the second term is mean zero term and it can be represented as follows

$$\begin{aligned} \Psi_n^2(\nabla^2 f(x_n)) = \widehat{N}_n^\intercal \nabla^2 f(x_n) N_n + \widehat{N}_n^\intercal \nabla^2 f(x_n) \\ + \nabla^2 f(x_n) N_n \end{aligned} \tag{31}$$

where $N_n, \widehat{N}_n$ are defined as follows $N_n = \Delta_n \left[\frac{1}{\Delta_n^1}, \ldots, \frac{1}{\Delta_n^N} \right] - I_N$ and $\widehat{N}_n = \widehat{\Delta}_n \left[\frac{1}{\widehat{\Delta}_n^1}, \ldots, \frac{1}{\widehat{\Delta}_n^N} \right] - I_N$. Where I_N is $N \times N$ identity matrix. From the foregoing, the per-iteration Hessian estimate $\widehat{H}_n$ can be re-written as follows:

$$H_n = \nabla^2 f(x_n) + \Psi_n(\nabla^2 f(x_n)) + O(\delta_n^2) + O(\delta_n^{-2}) \tag{32}$$

where, for any matrix H,

$$\begin{aligned} \Psi_n(H) &= \Psi_n^1(H) + \Psi_n^2(H) \\ &= \frac{1}{2} M_n \left[\Delta_n^\intercal H \Delta_n + \widehat{\Delta}_n^\intercal H \widehat{\Delta}_n \right] + \widehat{N}_n^\intercal H N_n \\ &\qquad + \widehat{N}_n^\intercal H + H N_n. \end{aligned} \tag{33}$$

Given that we operate in a simulation optimization setting, which implies $\nabla^2 f$ is not known, we construct the feedback term $\widehat{\Psi}_n$ in (2) by using $\overline{H}_{n-1}$ as a proxy for $\nabla^2 f$, i.e.,

$$\widehat{\Psi}_n = \Psi_n(\overline{H}_{n-1}). \tag{34}$$

*e) **Optimizing the step-sizes** b_n:* The optimal choice for b_n in (3) is the following:

$$b_i = \delta_i^4 / \sum_{j=0}^{i} \delta_j^4. \tag{35}$$

The main idea behind the above choice is provided below. From (32), we can infer that

$$\mathbb{E} \parallel \widehat{H}_n \parallel^2 \leq \frac{C}{\delta_n^4} \text{ for some } C < \infty.$$

This is because the third term in (32) vanishes asymptotically, while the fourth term there dominates asymptotically. Moreover, the noise factors in the fourth term in (32) are bounded above due to (C21) and independent of n, leaving the δ_n^2 term in the denominator there.

So, the optimization problem to be solved at instant n is as follows:

$$\sum_{i=0}^{n} (\tilde{b}_k)^2 \delta_i^{-4}, \text{ subject to} \tag{36}$$

$$\tilde{b}_i \geq 0 \ \forall i \text{ and } \sum_{i=0}^{n} \tilde{b}_i = 1. \tag{37}$$

The optimization variable $\tilde{b}_i$ from the above is related to the Hessian recursion (3) as follows:

$$\overline{H}_n = \sum_{i=0}^{n} \tilde{b}_k (\widehat{H}_i - \widehat{\Psi}_i). \tag{38}$$

The solution to (36) is achieved for $\tilde{b}_i^* = \delta_i^4 / \sum_{j=0}^{n} \delta_j^4, i = 1, \ldots, n$. The optimal choice $\tilde{b}_i^*$ can be translated to the step-sizes b_i, leading to (35).

IV. Convergence analysis for N-RDSA-3-IH

We make the same assumptions as those used in the analysis of [1], with a few minor alterations. The assumptions are listed below:

- **(C1)** The function f is four-times differentiable[2] with $\left|\nabla^4_{i_1 i_2 i_3 i_4} f(x)\right| < \infty$, for $i_1, i_2, i_3, i_4 = 1, \ldots, N$ and for all $x \in \mathbb{R}^N$.
- **(C2)** For each n and all x, there exists a $\rho > 0$ not dependent on n and x, such that $(x - x^*)^\intercal \bar{f}_n(x) \geq \rho \|x_n - x\|$, where $\bar{f}_n(x) = \Upsilon(\overline{H}_n)^{-1} \nabla f(x)$.
- **(C3)** $\{\xi_n, \xi_n^+, \xi_n^-, n = 1, 2, \ldots\}$ are such that, for all n, $\mathbb{E}\left[\xi_n^+ + \xi_n^- - 2\xi_n | \mathcal{F}_n, d_n\right] = 0$, where $\mathcal{F}_n = \sigma(x_m, m < n)$ denotes the underlying sigma-field..
- **(C4)** $\{d_n^i, i = 1, \ldots, N, n = 1, 2, \ldots\}$ are i.i.d. and independent of $\mathcal{F}_n$.
- **(C5)** The step-sizes a_n and perturbation constants δ_n are positive, for all n and satisfy

$$a_n, \delta_n \to 0 \text{ as } n \to \infty, \sum_n a_n = \infty \text{ and } \sum_n \left(\frac{a_n}{\delta_n}\right)^2 < \infty.$$

- **(C6)** For each $i = 1, \ldots, N$ and any $\rho > 0$, $P(\{\bar{f}_{ni}(x_n) \geq 0 \text{ i.o}\} \cap \{\bar{f}_{ni}(x_n) < 0 \text{ i.o}\} \mid \{|x_{ni} - x_i^*| \geq \rho \quad \forall n\}) = 0$.
- **(C7)** The operator Υ satisfies $\delta_n^2 \Upsilon(H_n)^{-1} \to 0$ a.s. and $E(\left\|\Upsilon(H_n)^{-1}\right\|^{2+\zeta}) \leq \rho$ for some $\zeta, \rho > 0$.
- **(C8)** For any $\tau > 0$ and nonempty $S \subseteq \{1, \ldots, N\}$, there exists a $\rho'(\tau, S) > \tau$ such that

$$\limsup_{n \to \infty} \left| \frac{\sum_{i \notin S} (x - x^*)_i \bar{f}_{ni}(x)}{\sum_{i \in S} (x - x^*)_i \bar{f}_{ni}(x)} \right| < 1 \text{ a.s.}$$

  for all $|(x - x^*)_i| < \tau$ when $i \notin S$ and $|(x - x^*)_i| \geq \rho'(\tau, S)$ when $i \in S$.
- **(C9)** For some $\alpha_0, \alpha_1 > 0$ and for all n, $\mathbb{E}{\xi_n}^2 \leq \alpha_0$, $\mathbb{E}\xi_n^{\pm 2} \leq \alpha_0$, $\mathbb{E} f(x_n)^2 \leq \alpha_1$, $\mathbb{E} f(x_n \pm \delta_n d_n)^2 \leq \alpha_1$ and $\mathbb{E}\left(\left\|\Upsilon(\overline{H}_n)\right\|^2 \mid \mathcal{F}_n\right) \leq \alpha_1$.
- **(C10)** $\delta_n = \frac{\delta_0}{(n+1)^\varsigma}$, where $\delta_0 > 0$ and $0 < \varsigma \leq 1/8$.

The reader is referred to Section II-B of [1] for a detailed discussion of the above assumptions. We remark here that (C1)-(C8) are identical to that in [1], while (C9) and (C10) introduce minor additional requirements on $\left\|\Upsilon(\overline{H}_n\right\|^2$ and δ_n, respectively and these are inspired by [3].

Lemma 1. ***(Bias in Hessian estimate)*** *Under (C1)-(C10), with $\widehat{H}_n$ defined according to either* (23) *or* (21)*, we have a.s. that*[3]*, for $i, j = 1, \ldots, N$,*

$$\left| \mathbb{E}\left[\widehat{H}_n(i,j) \middle| \mathcal{F}_n \right] - \nabla^2_{ij} f(x_n) \right| = O(\delta_n^2). \tag{39}$$

Proof. See Lemma 4 in [1]. The following provides general proof for both Uniform and Asymmetric Bernoulli cases

By a Taylor's series expansion, we obtain

$$\begin{aligned} f(x_n \pm \delta_n d_n) &= f(x_n) \pm \delta_n d_n^\intercal \nabla f(x_n) + \frac{\delta_n^2}{2} d_n^\intercal \nabla^2 f(x_n) d_n \\ &\pm \frac{\delta_n^3}{6} \nabla^3 f(x_n)(d_n \otimes d_n \otimes d_n) \\ &+ \frac{\delta_n^4}{24} \nabla^4 f(\tilde{x}_n^+)(d_n \otimes d_n \otimes d_n \otimes d_n). \end{aligned}$$

The fourth-order term in each of the expansions above can be shown to be of order $O(\delta_n^4)$ using (C1) and arguments similar to that in Lemma 1 of [1] . Hence,

$$\begin{aligned} &\frac{f(x_n + \delta_n d_n) + f(x_n - \delta_n d_n) - 2f(x_n)}{\delta_n^2} \\ &= d_n^\intercal \nabla^2 f(x_n) d_n + O(\delta_n^2) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} d_n^i d_n^j \nabla^2_{ij} f(x_n) + O(\delta_n^2) \\ &= \sum_{i=1}^{N} (d_n^i)^2 \nabla^2_{ii} f(x_n) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla^2_{ij} f(x_n) + O(\delta_n^2). \end{aligned}$$

[2]Here $\nabla^4 f(x) = \dfrac{\partial^4 f(x)}{\partial x^\intercal \partial x^\intercal \partial x^\intercal \partial x^\intercal}$ denotes the fourth derivate of f at x and $\nabla^4_{i_1 i_2 i_3 i_4} f(x)$ denotes the $(i_1 i_2 i_3 i_4)$th entry of $\nabla^4 f(x)$, for $i_1, i_2, i_3, i_4 = 1, \ldots, N$.

[3]Here $\widehat{H}_n(i,j)$ and $\nabla^2_{ij} f(\cdot)$ denote the (i,j)th entry in the Hessian estimate $\widehat{H}_n$ and the true Hessian $\nabla^2 f(\cdot)$, respectively.

Now, taking the conditional expectation of the Hessian estimate $\widehat{H}_n$ and observing that $\mathbb{E}[\xi_n^+ + \xi_n^- - 2\xi_n \mid \mathcal{F}_n, d_n] = 0$ by (C3), we obtain the following:

$$\mathbb{E}[\widehat{H}_n \mid \mathcal{F}_n] = \mathbb{E}\left[M_n \left(\sum_{i=1}^{N-1} (d_n^i)^2 \nabla_{ii}^2 f(x_n) + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n) + O(\delta_n^2) \right) \middle| \mathcal{F}_n \right]. \tag{40}$$

Note that the $O(\delta_n^2)$ term inside the conditional expectation above remains $O(\delta_n^2)$ even after the multiplication with M_n. We analyse the diagonal and off-diagonal terms in the multiplication of the matrix M_n with the scalar above, ignoring the $O(\delta_n^2)$ term.

Diagonal terms in (40)*:*

Consider the lth diagonal term inside the conditional expectation in (40):

$$\begin{aligned}
&\frac{1}{\kappa}\left((d_n^l)^2 - \lambda\right)\left(\sum_{i=1}^{N} (d_n^i)^2 \nabla_{ii}^2 f(x_n) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n) \right) \\
=&\frac{1}{\kappa}(d_n^l)^2 \sum_{i=1}^{N} (d_n^i)^2 \nabla_{ii}^2 f(x_n) \\
&+ \frac{2}{\kappa}(d_n^l)^2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n) \\
&- \frac{\lambda}{\kappa} \sum_{i=1}^{N} (d_n^i)^2 \nabla_{ii}^2 f(x_n) - \frac{2\lambda}{\kappa} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n).
\end{aligned} \tag{41}$$

From the distributions of d_n^i, d_n^j and the fact that d_n^i is independent of d_n^j for $i < j$, it is easy to see that $\mathbb{E}\left((d_n^l)^2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n) \middle| \mathcal{F}_n \right) = 0$ and $\mathbb{E}\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n) \middle| \mathcal{F}_n \right) = 0$. Thus, the conditional expectations of the second and fourth terms on the RHS of (41) are both zero.

The first term on the RHS of (41) with the conditional expectation can be simplified as follows:

$$\begin{aligned}
&\frac{1}{\kappa}\mathbb{E}\left((d_n^l)^2 \sum_{i=1}^{N} (d_n^i)^2 \nabla_{ii}^2 f(x_n) \middle| \mathcal{F}_n \right) \\
=&\frac{1}{\kappa}\mathbb{E}\left((d_n^l)^4 \nabla_{ll}^2 f(x_n) + \sum_{i=1, i\neq l}^{N} (d_n^l)^2 (d_n^i)^2 \nabla_{ii}^2 f(x_n) \right) \\
=&\frac{1}{\kappa}\left(\tau \nabla_{ll}^2 f(x_n) + \lambda^2 \sum_{i=1, i\neq l}^{N} \nabla_{ii}^2 f(x_n) \right), \text{ a.s.}
\end{aligned} \tag{42}$$

For the second equality above, we have used the fact that $\mathbb{E}[(d_n^l)^4] = \tau$ and $\mathbb{E}[(d_n^l)^2 (d_n^i)^2] = \mathbb{E}[(d_n^l)^2]\mathbb{E}[(d_n^i)^2] = \lambda^2$, $\forall l \neq i$.

The third term in (41) with the conditional expectation and without the negative sign can be simplified as follows:

$$\begin{aligned}
&\frac{\lambda}{\kappa}\mathbb{E}\left(\sum_{i=1}^{N} (d_n^i)^2 \nabla_{ii}^2 f(x_n) \middle| \mathcal{F}_n \right) \\
=&\frac{\lambda}{\kappa} \sum_{i=1}^{N} \mathbb{E}\left[(d_n^i)^2 \right] \nabla_{ii}^2 f(x_n) = \frac{\lambda^2}{\kappa} \sum_{i=1}^{N} \nabla_{ii}^2 f(x_n), \text{ a.s.}
\end{aligned} \tag{43}$$

Combining the above followed by some algebra, i.e., (42) − (43), we obtain

$$\frac{1}{\kappa}(\tau - \lambda^2)\nabla_{ll}^2 f(x_n)$$

By using the fact that $\kappa = \tau - \lambda^2$, we obtain

$$\frac{1}{\kappa}\mathbb{E}\left[\left((d_n^l)^2 - \lambda\right)\left(\sum_{i=1}^{N} (d_n^i)^2 \nabla_{ii}^2 f(x_n) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n) \right) \middle| \mathcal{F}_n \right] = \nabla_{ll}^2 f(x_n), \text{ a.s.}$$

Off-diagonal terms in (40)*:*

We now consider the (k, l)th term in (40): Assume w.l.o.g that $k < l$. Then,

$$\begin{aligned}
&\frac{1}{2\lambda^2}\mathbb{E}\left[d_n^k d_n^l \left(\sum_{i=1}^{N} (d_n^i)^2 \nabla_{ii}^2 f(x_n) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_n^i d_n^j \nabla_{ij}^2 f(x_n) \right) \middle| \mathcal{F}_n \right] \\
=&\frac{1}{2\lambda^2} \sum_{i=1}^{N} \mathbb{E}\left(d_n^k d_n^l (d_n^i)^2 \right) \nabla_{ii}^2 f(x_n) \\
&+ \frac{1}{\lambda^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbb{E}\left(d_n^k d_n^l d_n^i d_n^j \right) \nabla_{ij}^2 f(x_n) \\
=&\nabla_{kl}^2 f(x_n).
\end{aligned} \tag{44}$$

The last equality follows from the fact that the first term in (44) is 0 since $k \neq l$, while the second term in (44) can be seen to be equal to $\frac{1}{\lambda^2}\mathbb{E}\left((d_n^k)^2 (d_n^l)^2\right) \nabla_{kl}^2 f(x_n) = \nabla_{kl}^2 f(x_n)$. The claim follows for the case of both for uniform and asymmetric Bernoulli perturbations perturbations.

□

Theorem 2. ***(Strong Convergence of Hessian)*** *Under (C1)-(C10), we have that*

$$\overline{H}_n \to \nabla^2 f(x^*) \text{ a.s. as } n \to \infty.$$

In the above, $\overline{H}_n$ *is updated according to* (3)*.* $\widehat{H}_n$ *defined according to either* (21) *or* (23) *and the step-sizes* b_n *are chosen as suggested in* (15)*.*

Proof. For proving the main claim regarding $\overline{H}_n$, we closely follow the approach used to prove a corresponding result for N-SPSA-4 (see Theorem 1 in [3]). The first step is to prove the following:

$$\sum_{k=0}^{n} \frac{\delta_k^4 \left(\widehat{H}_k - \widehat{\Psi}_k - \mathbb{E}(\widehat{H}_k \mid \mathcal{F}_k)\right)}{\sum_{i=0}^{n} \delta_i^4} \to 0. \tag{45}$$

By a completely parallel argument to that used in the proof of Theorem 1 in [3], we obtain: For any $i, j = 1, \ldots, N$,

$$\mathbb{E}\left[\left((\widehat{H}_k)_{i,j} - (\widehat{\Psi}_k)_{i,j} - \mathbb{E}((\widehat{H}_k)_{i,j} \mid \mathcal{F}_k)\right)^2\right] = O(\delta_k^{-4}).$$

Now (45) follows by an application of Kronecker's Lemma along with the martingale convergence theorem (see Theorem 6.2.1 of [10]).

From Lemma 1, we have

$$\mathbb{E}[\widehat{H}_k \mid \mathcal{F}_k] = \nabla^2 f(x_n) + O(\delta_n^2) \text{ a.s.}$$

Since the Hessian is continuous near x_n and x_n converges almost surely to x^*, we have [4]

$$\begin{aligned}\sum_{k=0}^{n} \frac{\delta_k^4 \left(\mathbb{E}(\widehat{H}_k \mid \mathcal{F}_k)\right)}{\sum_{i=0}^{n} \delta_i^4} &= \sum_{k=0}^{n} \frac{\delta_k^4 \left(\nabla^2 f(x_n) + O(\delta_n^2)\right)}{\sum_{i=0}^{n} \delta_i^4} \\ &= \sum_{k=0}^{n} \frac{\delta_k^4 \left(\nabla^2 f(x^*) + o(1)\right)}{\sum_{i=0}^{n} \delta_i^4} \\ &\to \nabla^2 f(x^*) \text{ a.s. as } n \to \infty.\end{aligned}$$

The last step above follows from Toeplitz Lemma (see p. 89 of [10]) after observing that $\sum_{i=0}^{n} \delta_i^4 \to \infty$ due to (C10). The main claim now follows since

$$\overline{H}_n = \sum_{k=0}^{n} \frac{\delta_k^4 \left(\widehat{H}_k - \Psi_k\right)}{\sum_{i=0}^{n} \delta_i^4}.$$

□

We next present a convergence rate result for the special case of a quadratic objective function under the following additional assumptions:

(C11) f is quadratic and $\nabla^2 f(x^*) > 0$.

(C12) The operator Υ is chosen such that $\mathbb{E}(\| \Upsilon(\overline{H}_n) - \overline{H}_n \|^2) = o(e^{-2wn^{1-r}/(1-r)})$ and $\| \Upsilon(H) - H \|^2 / (1 + \| H \|^2)$ is uniformly bounded.

Theorem 3. ***(Quadratic case - Convergence rate)*** *Assume (C4), (C10), (C11) and (C12) and also that the setting is noise-free. Let $b_n = b_0/n^r$, $n = 1, 2, \ldots, k$, where $1/2 < r < 1$ and $0 < b_0 \leq 1$. For notational simplicity, let $H^* = \nabla^2 f(x^*)$. Letting $\Lambda_k = \overline{H}_k - H^*$, we have*

$$\textit{trace}[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)] = O(e^{-2b_0 n^{1-r}/1-r}). \tag{46}$$

Proof. Since the setting is noise-free with a quadratic objective, we can rewrite (12) as follows:

$$\widehat{H}_n = \Phi_n(\nabla^2 f(x_n)) + \Psi_n(\nabla^2 f(x_n)), \tag{47}$$

[4]Since assumptions (C1)-(C10) here are similar to that in [1], Theorem 5 of [1] holds here as well. In particular, this implies almost sure convergence of x_n to x^*.

where $\Psi_n(H)$ is defined in (13) and for any matrix H,

$$\Phi_n(H) = [M_n]_D \left(d_n^\intercal [H]_D \, d_n\right) + [M_n]_N \left(d_n^\intercal [H]_N \, d_n\right).$$

The proof involves the following steps:

Step 1: Here we prove the MSE convergence of $\overline{H}_k$, i.e., $\mathbb{E}[\Lambda_n^\intercal \Lambda_n] \to 0$ a.s. as $n \to \infty$.

Step 2: We unroll the recursion (3) and then derive convenient representation for trace$[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)]$.

Step 3: We derive the main result in (46) using a proof by contradiction.

Step 1: MSE convergence of $\overline{H}_n$

This part in exactly the same manner as part (i) in Theorem 3 of [3].

Step 2: Representation of trace$[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)]$

From (3) and (47), we have

$$\begin{aligned}\Lambda_n &= \Lambda_{n-1} - b_n(\overline{H}_{n-1} - \hat{H}_n + \hat{\Psi}_n) \\ &= (1 - b_n)\Lambda_{n-1} - b_n(H^* + \hat{\Psi}_n - \hat{H}_n) \\ &= (1 - b_n)\Lambda_{n-1} - b_n(H^* + \hat{\Psi}_n - \Phi_n(H^*) - \Psi_n(H^*) \\ &= (1 - b_n)\Lambda_{n-1} - b_n \Psi_n(\Upsilon(\overline{H}_{n-1})) \\ &\qquad + b_n(\Phi_n(H^*) - H^*) \\ &= (1 - b_n)\Lambda_{n-1} - b_n \Psi_n(\Lambda'_{n-1}) + b_n(\Phi_n(H^*) - H^*),\end{aligned} \tag{48}$$

where $\Lambda'_n = \Upsilon(\overline{H}_n) - H^*$. The equality in (48) follows from (13). Unrolling the recursion (48), we obtain

$$\begin{aligned}\Lambda_n = &\left[\prod_{k=1}^{n}(1 - b_k)\right] \Lambda_0 \\ &- \sum_{k=1}^{n} \left[\prod_{j=k+1}^{n} (1 - b_j)\right] b_k \Psi_k(\Lambda'_{k-1}) \\ &+ \sum_{k=1}^{n} \left[\prod_{j=k+1}^{n} (1 - b_j)\right] b_k(\Phi_k(H^*) - H^*) \ \ a.s.\end{aligned} \tag{49}$$

Squaring on both sides and taking expectations, we obtain

$$\begin{aligned}\mathbb{E}(\Lambda_n^\intercal \Lambda_n) = &\left[\prod_{k=1}^{n}(1 - b_k)\right]^2 \mathbb{E}(\Lambda_0^\intercal \Lambda_0) \\ &+ \sum_{k=1}^{n} \left[\prod_{j=k+1}^{n} (1 - b_j)\right]^2 b_k^2 \mathbb{E}(\Psi_k(\Lambda'_{k-1})^\intercal \Psi_k(\Lambda'_{k-1})) \\ &+ \sum_{k=1}^{n} \left[\prod_{j=k+1}^{n} (1 - b_j)\right]^2 b_k^2 \\ &\qquad \times \mathbb{E}((\Phi_k(H^*) - H^*)^\intercal (\Phi_k(H^*) - H^*)) \\ &- 2 \sum_{k=1}^{n} \left[\prod_{j=k+1}^{n} (1 - b_j)\right]^2 b_k^2 \mathbb{E}\left(\Psi_k(\Lambda_{k-1})^\intercal (\Phi_k(H^*) - H^*)\right).\end{aligned} \tag{50}$$

The equality above uses the fact that $\mathbb{E}(\Psi_k(\Lambda'_{k-1})) = 0$, $\mathbb{E}(\Phi_k(H^*) - H^*) = 0$, which get rid of the corresponding cross terms with the first term on RHS of (49). Notice that there are only n terms in the last term in the RHS of (50)

because d_k are identically distributed for each k and across k. We now characterize trace$[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)]$ using (50) as follows:

$$\begin{aligned}
\text{trace}\left[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)\right] &= \left[\prod_{k=1}^{n}(1-b_k)\right]^2 \text{trace}\left[\mathbb{E}(\Lambda_0^\intercal \Lambda_0)\right] \\
&+ \sum_{k=1}^{n}\left[\prod_{j=k+1}^{n}(1-b_j)\right]^2 b_k^2\ \tau\left(\mathbb{E}(\Lambda'_{k-1}\otimes\Lambda'_{k-1})\right) \\
&+ \sum_{k=1}^{n}\left[\prod_{j=k+1}^{n}(1-b_j)\right]^2 b_k^2 \\
&\quad\times \text{trace}\left[\mathbb{E}((\Phi_k(H^*)-H^*)^\intercal(\Phi_k(H^*)-H^*))\right] \\
&- 2\sum_{k=1}^{n}\left[\prod_{j=k+1}^{n}(1-b_j)\right]^2 b_k^2 \\
&\quad\times \text{trace}\left[\mathbb{E}\left(\Psi_k(\Lambda_{k-1})^\intercal(\Phi_k(H^*)-H^*)\right)\right]. \qquad (51)
\end{aligned}$$

As in the proof of Theorem 3 of [3], observe that $1-b_k = e^{-b_k}(1-O(b_k^2))$ and since $0<b_k<1$, we have that the $O(b_k^2)$ term is strictly positive. Letting $\Gamma_{ij}=\sum_{k=i}^{j} b_k$ with $\Gamma_{nn}=1$ and $\beta_{kn}=\left[\prod_{i=k+1}^{n}(1-O(b_i^2))\right]^2$, we can simplify (51) as follows:

$$\begin{aligned}
\text{trace}\left[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)\right] &= e^{-2\Gamma_{1n}}\beta_{0n}\text{trace}\left[\mathbb{E}(\Lambda_0^\intercal \Lambda_0)\right] \\
&+ e^{-2\Gamma_{1n}}\sum_{k=1}^{n} e^{2\Gamma_{1k}}\beta_{kn}b_k^2\ \tau\left(\mathbb{E}(\Lambda'_{k-1}\otimes\Lambda'_{k-1})\right) \\
&+ e^{-2\Gamma_{1n}}\sum_{k=1}^{n} e^{2\Gamma_{1k}}\beta_{kn}b_k^2 \\
&\quad\times \text{trace}\left[\mathbb{E}((\Phi_k(H^*)-H^*)^\intercal(\Phi_k(H^*)-H^*))\right] \\
&- 2e^{-2\Gamma_{1n}}\sum_{k=1}^{n} e^{2\Gamma_{1k}}\beta_{kn}b_k^2 \\
&\quad\times \text{trace}\left[\mathbb{E}\left(\Psi_k(\Lambda_{k-1})^\intercal(\Phi_k(H^*)-H^*)\right)\right]. \qquad (52)
\end{aligned}$$

Comparing the sum with integrals, we obtain

$$\Gamma_{ij} = \int_i^j \frac{b_0}{x^r}dx + O(1) = \left(\frac{b_0}{1-r}\right)(j^{1-r}-i^{1-r}) + O(1),$$

where we have used the facts that $0<b_k<1, \forall k\geq 2$ and $\sum_{k=i}^{j} b_k \to \infty$ as $j-i\to\infty$ since $b_k = b_0/k^r$ with $r>0.5$. Observing that β_{kn} are uniformly upper-bounded, say by $\bar{\beta}_n$, we have

$$\begin{aligned}
\text{trace}\left[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)\right] &= e^{-2\Gamma_{1n}}\beta_{0n}\text{trace}\left[\mathbb{E}(\Lambda_0^\intercal \Lambda_0)\right] \\
&+ \bar{\beta}_n e^{-2b_0 n^{1-r}/(1-r)}\sum_{k=1}^{n} e^{2b_0 k^{1-r}/(1-r)}\times\frac{b_0^2}{k^{2r}} \\
&\qquad\times\tau\left(\mathbb{E}(\Lambda'_{k-1}\otimes\Lambda'_{k-1})\right) \\
&+ \bar{\beta}_n e^{-2b_0 n^{1-r}/(1-r)}\sum_{k=1}^{n} e^{2b_0 k^{1-r}/(1-r)}\times\frac{b_0^2}{k^{2r}} \\
&\quad\times \text{trace}\left[\mathbb{E}((\Phi_k(H^*)-H^*)^\intercal(\Phi_k(H^*)-H^*))\right] \\
&- 2\bar{\beta}_n e^{-2b_0 n^{1-r}/(1-r)}\sum_{k=1}^{n} e^{2b_0 k^{1-r}/(1-r)}\times\frac{b_0^2}{k^{2r}} \\
&\quad\times \text{trace}\left[\mathbb{E}\left(\Psi_k(\Lambda_{k-1})^\intercal(\Phi_k(H^*)-H^*)\right)\right]. \qquad (53)
\end{aligned}$$

$$\begin{aligned}
\text{trace}\left[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)\right] &= e^{-2\Gamma_{1n}}\beta_{0n}\text{trace}\left[\mathbb{E}(\Lambda_0^\intercal \Lambda_0)\right] \\
&+ \bar{\beta}_n e^{-2b_0 n^{1-r}/(1-r)}\sum_{k=1}^{n} e^{2b_0 k^{1-r}/(1-r)}\times\frac{b_0^2}{k^{2r}} \\
&\qquad\times\tau\left(\mathbb{E}(\Lambda'_{k-1}\otimes\Lambda'_{k-1})\right) \\
&+ \bar{\beta}_n e^{-2b_0 n^{1-r}/(1-r)}\sum_{k=1}^{n} e^{2b_0 k^{1-r}/(1-r)}\times\frac{b_0^2}{k^{2r}} \\
&\times \text{trace}\left[\mathbb{E}[((\Phi_k(H^*)-H^*)-2\Psi_k(\Lambda_{k-1}))^\intercal(\Phi_k(H^*)-H^*)]\right]. \qquad (54)
\end{aligned}$$

Step 3: The big-O result on trace$[\mathbb{E}(\Lambda_n^\intercal \Lambda_n)]$ convergence In comparison to Eq. (7.6) in [3] that corresponds to (53) for the N-SPSA-4 setting, there are two extra terms - the third and the fourth - in the RHS of (53). Consider the third term on the RHS of (53). As a consequence of (C9) and the construction of random perturbations d_k, it can be seen that trace $[\mathbb{E}((\Phi_k(H^*)-H^*)^\intercal(\Phi_k(H^*)-H^*))]$ is uniformly bounded above, say by $\top$, independent of k. Thus, the third term is equivalent to the following:

$$\bar{c}_n e^{-2b_0 n^{1-r}/(1-r)}\sum_{k=1}^{n} e^{2b_0 k^{1-r}/(1-r)}\frac{b_0^2}{k^{2r}}\top. \qquad (55)$$

Now by moving out side exponential term to inside and changing the summation to integration, we can rewrite the above equation as follows

$$\overline{\overline{c_n}}\int_1^n \frac{1}{k^{2r}}. \qquad (56)$$

Where $\overline{\overline{c_n}}$ is a constant. By above equation we can conclude that the third term is of $O(\frac{1}{k^{2r-1}})$

Part (iii) needed fix □

V. Convergence analysis for N-SPSA-3-IH

We make the same assumptions as those used in the analysis of [1], with a few minor alterations. The assumptions are listed below:

- **(C13)** The function f is four-times differentiable[5] with $\left|\nabla^4_{i_1 i_2 i_3 i_4} f(x)\right| < \infty$, for $i_1, i_2, i_3, i_4 = 1,\ldots,N$ and for all $x\in\mathbb{R}^N$.
- **(C14)** For each n and all x, there exists a $\rho>0$ not dependent on n and x, such that $(x-x^*)^\intercal \bar{f}_n(x) \geq \rho\,\|x_n - x\|$, where $\bar{f}_n(x) = \Upsilon(\overline{H}_n)^{-1}\nabla f(x)$.
- **(C15)** $\{\xi_n, \xi_n^+, \xi_n^-, n = 1,2,\ldots\}$ are such that, for all n, $\mathbb{E}\left[\xi_n^+ + \xi_n^- - 2\xi_n | \mathcal{F}_n, \Delta_n, \widehat{\Delta}_n\right] = 0$, where $\mathcal{F}_n = \sigma(x_m, m<n)$ denotes the underlying sigma-field..
- **(C16)** $\{\Delta_n^i, \widehat{\Delta}_n^i, i = 1,\ldots,N, n = 1,2,\ldots\}$ are i.i.d independent of $\mathcal{F}_n$ and for some $\alpha_0>0$ and for all n,l $|\Delta_{nl}|\leq\alpha_0$, Δ_{nl} is symmetrically distributed about 0, Δ_{nl} are mutually independent across n and l and they satisfy $\mathbb{E}(\Delta_{nl}^{-2})$,$\mathbb{E}(\widehat{\Delta}_{nl}^{-2})\leq\alpha_0$.

[5]Here $\nabla^4 f(x) = \dfrac{\partial^4 f(x)}{\partial x^\intercal \partial x^\intercal \partial x^\intercal \partial x^\intercal}$ denotes the fourth derivate of f at x and $\nabla^4_{i_1 i_2 i_3 i_4} f(x)$ denotes the $(i_1 i_2 i_3 i_4)$th entry of $\nabla^4 f(x)$, for $i_1, i_2, i_3, i_4 = 1,\ldots,N$.

(**C17**) The step-sizes a_n and perturbation constants δ_n are positive, for all n and satisfy

$$a_n, \delta_n \to 0 \text{ as } n \to \infty, \sum_n a_n = \infty \text{ and } \sum_n \left(\frac{a_n}{\delta_n}\right)^2 < \infty.$$

(**C18**) For each $i = 1, \ldots, N$ and any $\rho > 0$, $P(\{\bar{f}_{ni}(x_n) \geq 0 \text{ i.o}\} \cap \{\bar{f}_{ni}(x_n) < 0 \text{ i.o}\} \mid \{|x_{ni} - x_i^*| \geq \rho \quad \forall n\}) = 0$.

(**C19**) The operator Υ satisfies $\delta_n^2 \Upsilon(H_n)^{-1} \to 0$ a.s. and $E(\left\|\Upsilon(H_n)^{-1}\right\|^{2+\zeta}) \leq \rho$ for some $\zeta, \rho > 0$.

(**C20**) For any $\tau > 0$ and nonempty $S \subseteq \{1, \ldots, N\}$, there exists a $\rho'(\tau, S) > \tau$ such that

$$\limsup_{n \to \infty} \left| \frac{\sum_{i \notin S} (x - x^*)_i \bar{f}_{ni}(x)}{\sum_{i \in S} (x - x^*)_i \bar{f}_{ni}(x)} \right| < 1 \text{ a.s.}$$

for all $|(x - x^*)_i| < \tau$ when $i \notin S$ and $|(x - x^*)_i| \geq \rho'(\tau, S)$ when $i \in S$.

(**C21**) For some $\alpha_0, \alpha_1, \alpha_2 > 0$ and for all n, l, m, $\mathbb{E}{\xi_n}^2 \leq \alpha_0$, $\mathbb{E}{\xi_n^{\pm}}^2 \leq \alpha_0$, $\mathbb{E}f(x_n)^2 \leq \alpha_1$, $\mathbb{E}f(x_n + \delta_n\Delta_n + \delta_n\widehat{\Delta}_n)^2, \mathbb{E}f(x_n - \delta_n\Delta_n - \delta_n\widehat{\Delta}_n)^2 \leq \alpha_1$, $\mathbb{E}\left[|f(x_n + \delta_n\Delta_n + \delta_n\widehat{\Delta}_n)/(\Delta_{nl}\widehat{\Delta}_{nm})|^{2+\alpha_2} \mid \mathcal{F}_n\right]$, $\mathbb{E}\left[|f(x_n - \delta_n\Delta_n - \delta_n\widehat{\Delta}_n)/(\Delta_{nl}\widehat{\Delta}_{nm})|^{2+\alpha_2} \mid \mathcal{F}_n\right]$, $\mathbb{E}\left[(\xi_n^+ + \xi_n^- - 2\xi_n)^2/(\Delta_{nl}\widehat{\Delta}_{nm})^2 \middle| \mathcal{F}_n\right] \leq \alpha_1$ and $\mathbb{E}\left(\left\|\Upsilon(\overline{H}_n)\right\|^2 \mid \mathcal{F}_n\right) \leq \alpha_1$.

(**C22**) $\delta_n = \frac{\delta_0}{(n+1)^{\varsigma}}$, where $\delta_0 > 0$ and $0 < \varsigma \leq 1/8$.

The reader is referred to Section II-B of [1] for a detailed discussion of the above assumptions. We remark here that (C13)-(C20) are identical to that in [1], while (C21) and (C22) introduce minor additional requirements on $\left\|\Upsilon(\overline{H}_n\right\|^2$ and δ_n, respectively and these are inspired by [3].

Lemma 4. ***(N-SPSA-3-IH Bias in Hessian estimate)*** *Under* ***(C13)-(C22) equivalent****, with $\widehat{H}_n$ defined according to* (25) *, we have a.s. that*[6]*, for $i, j = 1, \ldots, N$,*

$$\left|\mathbb{E}\left[\widehat{H}_n(i,j)\middle|\mathcal{F}_n\right] - \nabla_{ij}^2 f(x_n)\right| = O(\delta_n^2). \tag{57}$$

Proof. See Proposition 4.2 in [2]. □

Theorem 5. ***(N-SPSA-3-IH Strong Convergence of Hessian)*** *Under (C13)-(C22), we have that*

$$\overline{H}_n \to \nabla^2 f(x^*) \text{ a.s. as } n \to \infty.$$

In the above, $\overline{H}_n$ is updated according to (3)*. $\widehat{H}_n$ defined according to* (25) *and the step-sizes b_n are chosen as suggested in* (35)*.*

Proof. This proof is similar to the proof of the theorem 2 □

We next present a convergence rate result for the special case of a quadratic objective function under the following additional assumptions:

(**C23**) f is quadratic and $\nabla^2 f(x^*) > 0$.

(**C24**) The operator Υ is chosen such that $\mathbb{E}(\| \Upsilon(\overline{H}_n) - \overline{H}_n \|^2) = o(e^{-2wn^{1-r}/(1-r)})$ and $\| \Upsilon(H) - H \|^2 / (1 + \| H \|^2)$ is uniformly bounded.

[6]Here $\widehat{H}_n(i,j)$ and $\nabla_{ij}^2 f(\cdot)$ denote the (i,j)th entry in the Hessian estimate $\widehat{H}_n$ and the true Hessian $\nabla^2 f(\cdot)$, respectively.

Theorem 6. ***(N-SPSA-3-IH Quadratic case - Convergence rate)*** *Assume (C16), (C22), (C23) and (C24) and also that the setting is noise-free. Let $b_n = b_0/n^r$, $n = 1, 2, \ldots, k$, where $1/2 < r < 1$ and $0 < b_0 \leq 1$. For notational simplicity, let $H^* = \nabla^2 f(x^*)$. Letting $\Lambda_k = \overline{H}_k - H^*$, we have*

$$\mathit{trace}[\mathbb{E}(\Lambda_n^{\intercal}\Lambda_n)] = O(e^{-2b_0 n^{1-r}/1-r}). \tag{58}$$

Proof. This proof follows from the proof of theorem 3 of [3] □

VI. Numerical Experiments

A. Implementation

We test the performance of N-RDSA-3-Unif, N-RDSA-3-AsymBer and N-SPSA-4, with/without improved Hessian estimation. N-SPSA-4 algorithm uses Bernoulli ± 1-valued perturbations, while N-RDSA-3/N-RDSA-3-IH come in two variants - one that uses $U[-1, 1]$ distributed perturbations (referred to as N-RDSA-3-Unif/N-RDSA-3-IH-Unif) and the other that uses asymmetric Bernoulli perturbations (referred to as N-RDSA-3-AsymBer/N-RDSA-3-IH-AsymBer)[7].

For the empirical evaluations, we use the following two loss functions in $N = 10$ dimensions:

a) Quadratic loss:

$$f(x) = x^{\intercal}Ax + b^{\intercal}x. \tag{59}$$

The optimum x^* for the above f is such that each coordinate of x^* is -0.9091, with $f(x^*) = -4.55$.

b) Fourth-order loss:

$$f(x) = x^{\intercal}A^{\intercal}Ax + 0.1\sum_{j=1}^{N}(Ax)_j^3 + 0.01\sum_{j=1}^{N}(Ax)_j^4. \tag{60}$$

The optimum x^* for above f is $x^* = 0$, with $f(x^*) = 0$.

In both functions, A is such that NA is an upper triangular matrix with each entry one, b is the N-dimensional vector of ones and the noise structure is similar to that used in [8]. For any x, the noise is $[x^{\intercal}, 1]z$, where $z \approx \mathcal{N}(0, \sigma^2 I_{11\times 11})$. We perform experiments for noisy as well as noise-less settings, with $\sigma = 0.1$ for the noisy case.

For all algorithms, we set $\delta_n = 3.8/n^{0.101}$ and $a_n = 1/n^{0.6}$, while b_n are set according to (15). These choices have been used for N-SPSA-4 implementations before (see [8]) and have demonstrated good finite-sample performance empirically, while satisfying the theoretical requirements needed for asymptotic convergence. For all the algorithms, the initial point x_0 is the N-dimensional vector of ones. For both N-SPSA-4 and N-RDSA-3/N-RDSA-3-IH, an initial 20% of the simulation budget was used up by 1SPSA/1RDSA and the resulting iterate was used to initialize N-SPSA-4/N-RDSA-3. The distribution parameter ϵ is set to 0.0001 for N-RDSA-3 and to 0.01 for 1RDSA.

[7]The implementation is available at https://github.com/prashla/RDSA/archive/master.zip.

TABLE II: Normalized loss values for fourth-order objective (60) with and without noise: standard error from 500 replications shown after $\pm$

Noise parameter $\sigma = 0.1$		
	Regular	**Improved Hessian estimation**
N-SPSA-4	0.132 ± 0.0267	0.104 ± 0.0355
N-RDSA-3-Unif	0.115 ± 0.0214	0.0271 ± 0.0538
N-RDSA-3-AsymBer	0.0471 ± 0.021	$\mathbf{0.0099 \pm 0.0014}$
Noise parameter $\sigma = 0$		
	Regular	**Improved Hessian estimation**
N-SPSA-4	0.0795 ± 0.0234	0.0628 ± 0.0234
N-RDSA-3-Unif	0.0813 ± 0.0275	0.0214 ± 0.00376
N-RDSA-3-AsymBer	0.0199 ± 0.0114	$\mathbf{0.0098 \pm 0.00147}$

TABLE III: Normalized loss values for quadratic objective (59) with and without noise: standard error from 500 replications shown after $\pm$

Noise parameter $\sigma = 0.1$		
	Regular	**Improved Hessian estimation**
N-SPSA-4	-0.0062 ± 0.1164	-0.1229 ± 0.1374
N-RDSA-3-Unif	0.0485 ± 0.1465	-0.259 ± 0.0398
N-RDSA-3-AsymBer	-0.2564 ± 0.068	$\mathbf{-0.2877 \pm 0.0051}$
Noise parameter $\sigma = 0$		
	Regular	**Improved Hessian estimation**
N-SPSA-4	-0.0785 ± 0.1178	-0.1716 ± 0.1339
N-RDSA-3-Unif	0.0326 ± 0.1599	-0.2672 ± 0.0299
N-RDSA-3-AsymBer	-0.2777 ± 0.0488	$\mathbf{-0.2881 \pm 0.0012}$

TABLE IV: NMSE values for quadratic objective (59) with and without noise: standard error from 500 replications shown after $\pm$

Noise parameter $\sigma = 0.1$		
	Regular	**Improved Hessian estimation**
N-SPSA-4	0.9491 ± 0.0131	0.5495 ± 0.0217
N-RDSA-3-Unif	1.0073 ± 0.0140	0.1953 ± 0.0095
N-RDSA-3-AsymBer	0.1667 ± 0.0095	$\mathbf{0.0324 \pm 0.0007}$
Noise parameter $\sigma = 0$		
	Regular	**Improved Hessian estimation**
N-SPSA-4	0.7325 ± 0.0180	0.3939 ± 0.0230
N-RDSA-3-Unif	0.9834 ± 0.0170	0.1623 ± 0.0086
N-RDSA-3-AsymBer	0.0686 ± 0.0078	$\mathbf{0.0316 \pm 0.0006}$

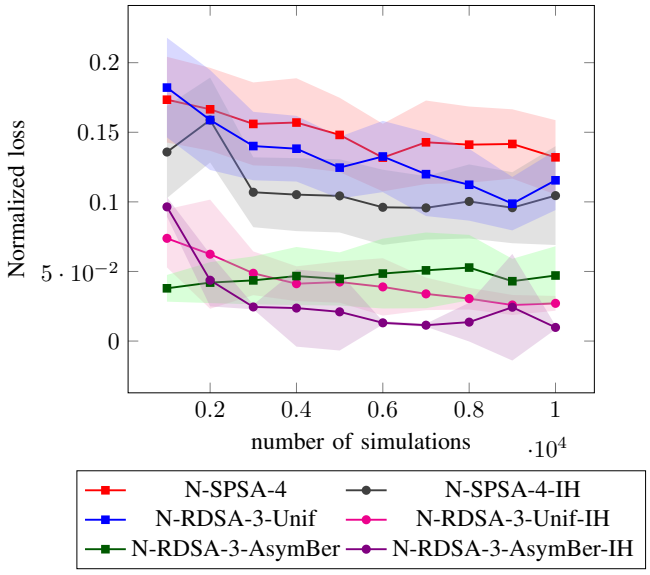


Fig. 2: Normalized loss vs. number of simulations for fourth-order loss (60) with $\sigma = 0.1$ for N-SPSA-4, N-RDSA-3-Unif and N-RDSA-3-AsymBer algorithms with/without improved Hessian estimation: bands around the curves represent standard error from 500 replications.

B. Results

We use normalized loss and normalized MSE (NMSE) as performance metrics for evaluating the algorithms. NMSE is the ratio $\|x_{n_{\text{end}}} - x^*\|^2 / \|x_0 - x^*\|^2$, while normalized loss is the ratio $f(x_{n_{\text{end}}})/f(x_0)$. Here n_{end} denotes the iteration number when the algorithm stopped updating its parameter. Note that n_{end} is a function of the simulation budget. N-RDSA-3/N-RDSA-3-IH use only three simulations per-iteration and hence, n_{end} is $1/3$rd of the simulation budget, while it is $1/4$th of the simulation budget for N-SPSA-4, since the latter algorithm uses four simulations per-iteration.

Tables II–III present the normalized loss values observed for the three algorithms - N-SPSA-4, N-RDSA-3-Unif and N-RDSA-3-AsymBer - with/without improved Hessian estimation scheme and for the fourth-order and quadratic loss functions, respectively. Table IV presents the NMSE values obtained for the aforementioned algorithms with the quadratic loss. The results in Tables II–IV are obtained after running all the algorithms with a budget of 10000 function evaluations. Figure 2 plots the normalized loss as a function of the simulation budget with the fourth-order loss objective with $\sigma = 0.1$ (see Figures 3a–3a in the appendix for similar results with $\sigma = 0$ for fourth-order loss and $\sigma = 0.1$ for quadratic loss). From the results in Tables II–IV and Fig 2, we make the following observations:

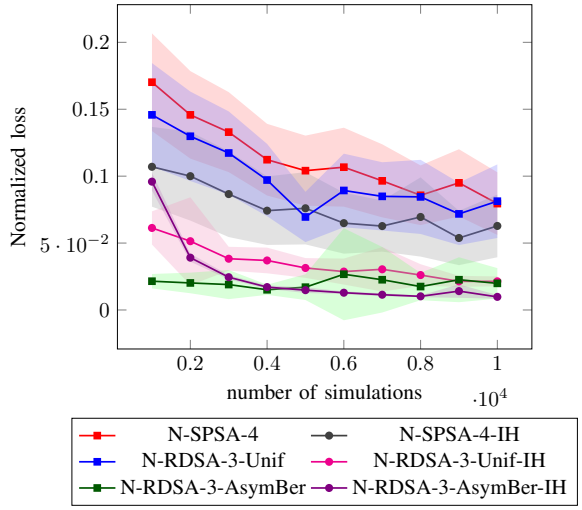
Observation 1: *Among N-RDSA-3 schemes, N-RDSA-3-IH performs better than regular N-RDSA-3, for both perturbation choices.*

Observation 2: *N-RDSA-3-IH variants outperform both N-SPSA-4 and N-SPSA-4-IH, with N-RDSA-3-IH-AsymBer performing the best overall.*

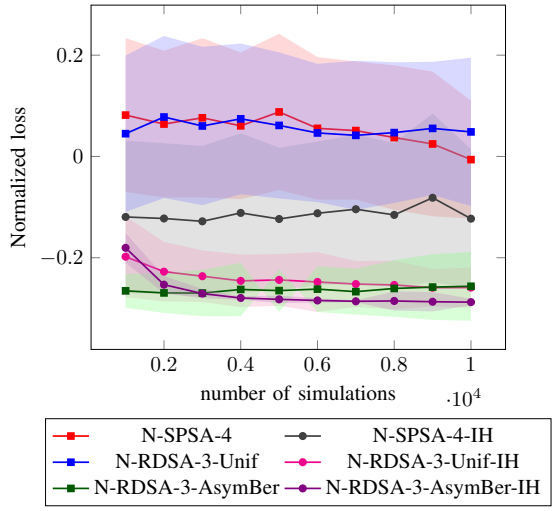
VII. Conclusions

We presented an improved Hessian estimation scheme for the second-order random directions stochastic approximation (N-RDSA-3) algorithm [1]. The proposed scheme was shown to be provably convergent to the true Hessian. The advantage with N-RDSA-3-IH is that it requires only 75% of the simulation cost per-iteration for N-SPSA-4 with Hessian estimation improvements (N-SPSA-4-IH) [3]. Numerical experiments demonstrated that N-RDSA-3-IH outperforms both N-SPSA-4-IH and N-RDSA-3 without the improved Hessian estimation scheme.

Appendix



(a) Fourth-order loss (60) with $\sigma = 0$.



(b) Quadratic loss (59) with $\sigma = 0.1$.

Fig. 3: Normalized loss vs. number of simulations in two different loss settings for all the algorithms.

References

[1] L. A. Prashanth, S. Bhatnagar, M. Fu, and S. Marcus, "Adaptive system optimization using random directions stochastic approximation," *arXiv preprint arXiv:1307.3176v2*, 2014.

[2] S. Bhatnagar and L. Prashanth, "Simultaneous perturbation newton algorithms for simulation optimization," *Journal of Optimization Theory and Applications*, vol. 164, no. 2, pp. 621–643, 2015.

[3] J. C. Spall, "Feedback and weighting mechanisms for improving Jacobian estimates in the adaptive simultaneous perturbation algorithm," *IEEE Trans. Autom. Contr.*, vol. 54, no. 6, pp. 1216–1229, 2009.

[4] S. Bhatnagar, H. L. Prasad, and L. A. Prashanth, *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods (Lecture Notes in Control and Information Sciences)*. Springer, 2013, vol. 434.

[5] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Auto. Cont.*, vol. 37, no. 3, pp. 332–341, 1992.

[6] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer Verlag, 1978.

[7] D. C. Chin, "Comparative study of stochastic algorithms for system optimization based on gradient approximations," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 27, no. 2, pp. 244–249, 1997.

[8] J. C. Spall, "Adaptive stochastic approximation by the simultaneous perturbation method," *IEEE Trans. Autom. Contr.*, vol. 45, pp. 1839–1853, 2000.

[9] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[10] R. G. Laha and V. K. Rohatgi, *Probability Theory*. Wiley, New York, 1979.