

Generalized Deterministic Perturbations For Stochastic Gradient Search

Chandramouli K.¹, Prabuchandran K.J.^{1,2}, D. Sai Koti Reddy³, and Shalabh Bhatnagar^{1,4}

Abstract—Stochastic optimization (SO) considers the problem of optimizing an objective function in the presence of noise. Most of the solution techniques in SO estimate gradients from the noise corrupted observations of the objective and adjust parameters of the objective along the direction of the estimated gradients to obtain locally optimal solutions. Two prominent algorithms in SO namely Random Direction Kiefer-Wolfowitz (RDKW) and Simultaneous Perturbation Stochastic Approximation (SPSA) obtain noisy gradient estimate by randomly perturbing all the parameters simultaneously. This forces the search direction to be random in these algorithms and causes them to suffer additional noise on top of the noise incurred from the samples of the objective. Owing to this additional noise, the idea of using deterministic perturbations instead of random perturbations for gradient estimation has also been studied. Two specific constructions of the deterministic perturbation sequence using lexicographical ordering and Hadamard matrices have been explored and encouraging results have been reported in the literature. In this paper, we characterize the class of deterministic perturbation sequences that can be utilized in the RDKW algorithm. This class expands the set of known deterministic perturbation sequences available in the literature. Using our characterization, we propose construction of a deterministic perturbation sequence that has the least cycle length among all deterministic perturbations. Through simulations we illustrate the performance gain of the proposed deterministic perturbation sequence in the RDKW algorithm over the Hadamard and the random perturbation counterparts. We also establish the convergence of the RDKW algorithm for the generalized class of deterministic perturbations.

I. INTRODUCTION

Stochastic optimization (SO) problems frequently arise in engineering disciplines such as transportation systems, machine learning, service systems, manufacturing etc. Practical limitations, lack of model information and the large dimensionality of these problems prohibit analytic solutions to these problems. Simulation is often employed to evaluate the performance of the current parameters of the system. Simulating and evaluating the system's performance is generally expensive and one is typically constrained by a simulation budget. In such scenarios, owing to the simulation budget one aims to drive the system to optimal parameter settings using as few simulations as possible.

Under the SO framework, we have a system that gives noise-corrupted feedback of the performance for the currently set parameters, i.e., given the system parameter vector θ , the feedback that is available is the noisy evaluation $h(\theta, \xi)$ of the performance $J(\theta) = \mathbb{E}_\xi[h(\theta, \xi)]$ where ξ is the noise

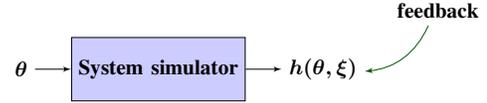


Fig. 1: Stochastic Optimization Model

term inherent in the system and $J(\theta)$ denotes the expected performance of the system for the parameter θ . The pictorial description of such a system is shown in Figure 1. The objective in the SO problem then is to determine a parameter θ^* that gives the optimal expected performance of the system, i.e.,

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^p} J(\theta). \quad (1)$$

Analogous to solutions for deterministic optimization problems where the explicit analytic gradient of the objective function is used to adjust the parameters along the negative gradient directions, many of the solution approaches in SO mimic the familiar gradient descent algorithm. However, unlike the deterministic setting, the SO setting only has access to noise corrupted samples of the objective. Thus, in the SO setting, one essentially aims at estimating the gradient of the objective function using noisy cost samples. In the pioneering work by Kiefer and Wolfowitz [1], the gradient is estimated by approximating each of the partial derivatives using either a two-sided or a one-sided finite difference approximation (FDSA) algorithm. This algorithm requires $2p$ objective function evaluations (or simulations) per iteration for the two-sided gradient approximation scheme and $p + 1$ simulations per iteration for the one-sided scheme (for a p -dimensional parameter problem, see [2]). As the number of simulations per iteration required for gradient estimation scales linearly with the dimension of the problem, FDSA algorithm is expensive to deploy under high-dimensional parameter settings.

In [3], Random Direction Kiefer-Wolfowitz (RDKW) algorithm that uses only two simulations per iteration for obtaining gradient estimates has been proposed. In the RDKW algorithm, all the parameters are randomly perturbed simultaneously using two parallel simulations and function evaluations at those perturbed parameters are used to obtain the gradient estimate. In the RDKW algorithm, the random perturbation vector as well as the random direction vector involved in estimating the gradient have been kept the same. For the choice of random direction (or perturbation), various distributions like spherical uniform distribution [3], uniform distribution [4], normal and Cauchy distribution [5], asymmetric Bernoulli [6] have been explored. The number

¹ Department of Computer Science and Automation, Indian Institute of Science (IISc)

² Supported by Amazon-IISc Postdoctoral fellowship

³ IBM Research, Bangalore

⁴ Robert Bosch Centre for Cyber-Physical Systems, IISc

of simulations required for estimating the gradients in the RDKW algorithm is significantly less compared to the FDSA algorithm and the algorithm is seen to perform empirically better than FDSA.

In a seminal work [7], the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm that uses two simulations similar to RDKW has been proposed. Unlike the RDKW algorithm, SPSA employs different choices for parameter perturbations and the random direction of movement, in particular, the random perturbation direction and the random direction of movement have been chosen to be inverses of each other. In [7], symmetric Bernoulli distribution has been shown to be the best choice for random perturbations among all the distributions and the proposed SPSA scheme has been proven to perform asymptotically better compared to FDSA. In [8], a comprehensive comparative study of the stochastic optimization algorithms namely FDSA, RDKW and SPSA has been provided. Further, under a general third order cross derivative assumption on the loss function, RDKW with symmetric Bernoulli distribution has been shown to be the best choice for random directions. In [9], an example of a loss function that does not satisfy the third order cross derivative condition in [8] has been constructed. For such a loss function, it has been shown that the optimal distribution choice for random directions need not be symmetric Bernoulli.

In [3] and [10], to further reduce simulation cost per iteration, extensions of the RDKW and SPSA algorithms that estimate the gradient with only one simulation or measurement of the objective have been considered. However, it is observed that the one-simulation gradient estimate has higher bias compared to the two-simulation gradient estimate. In [11] and [12], deterministic conditions for the perturbation and noise sequences required to obtain almost sure convergence of the iterates have been discussed. In [13], to enhance the performance of one-sided SPSA scheme, deterministic perturbations based on lexicographical ordering and Hadamard matrices have been proposed. Further, the numerical results in [13], illustrate the benefit of Hadamard matrix based perturbation sequences as it has been shown to improve the performance of SPSA empirically for the case of one sided measurements. In [14], a unified view of both RDKW and SPSA is presented and a binary deterministic perturbation sequence using orthogonal arrays [15] for obtaining gradient estimate in both of the algorithms has been discussed.

In this paper, we generalize the class of deterministic perturbation sequences that can be utilized in the RDKW algorithm. Based on this characterization, we provide a construction of a deterministic perturbation sequence using a specially chosen circulant matrix. We empirically study the performance of the constructed sequence against the afore mentioned Hadamard matrix based deterministic perturbations and the randomized perturbations. We expect with our generalization the study of rate of convergence for the RDKW algorithm based on deterministic perturbation sequences would be possible. We now summarize our con-

tributions:

- We generalize the class of deterministic perturbation sequences that can be applied in the RDKW algorithm.
- We provide a special construction of deterministic perturbation sequence with smaller cycle length compared to Hadamard perturbation sequence.
- We illustrate the performance gain of the proposed deterministic perturbations over the Hadamard matrix based perturbations as well as random perturbations.
- We prove the convergence of the RDKW algorithm for the class of deterministic perturbations.

II. CONDITIONS ON DETERMINISTIC PERTURBATIONS

In this section, we describe the classical RDKW algorithm and motivate the necessary conditions that a deterministic perturbation sequence should satisfy for almost sure convergence of the iterates in the deterministic perturbation version of RDKW algorithm.

The standard RDKW algorithm iteratively updates the parameter vector along the direction of the negative estimated gradient, i.e.,

$$\theta_{n+1} = \theta_n - a_n \widehat{\nabla J}(\theta_n), \quad (2)$$

where a_n is the step-size that satisfies standard stochastic approximation conditions (see Assumption **A2** in section IV) and $\widehat{\nabla J}$ is the estimate of the gradient of the objective function J at the current parameter.

In the case of two-simulation RDKW algorithm, the gradient estimate at θ is obtained as

$$\widehat{\nabla J}(\theta) = \frac{J(\theta + \delta d) - J(\theta - \delta d)}{2\delta} d, \quad (3)$$

where d is the random perturbation direction chosen according to a specific probability distribution. The properties that the specific distribution on d should satisfy can be obtained as explained below. The Taylor series expansion of $J(\theta \pm \delta d)$ around θ is given by

$$J(\theta \pm \delta d) = J(\theta) \pm \delta d^T \nabla J(\theta) + o(\delta^2). \quad (4)$$

From (4), the error between the estimate and the true gradient at θ can be obtained as

$$\begin{aligned} & \frac{J(\theta + \delta d) - J(\theta - \delta d)}{2\delta} d - \nabla J(\theta) \\ &= (dd^T - I) \nabla J(\theta) + o(\delta). \end{aligned} \quad (5)$$

Note that the term $(dd^T - I) \nabla J(\theta)$ constitutes the bias in the gradient estimate. For the error estimate in (5) to be negligible, we require

$$\mathbb{E} \left[dd^T \right] = I. \quad (6)$$

Here, the expectation $\mathbb{E}[\cdot]$ is taken over the random perturbation distribution.

In the one-simulation version of the RDKW algorithm, the gradient estimate at θ is obtained as

$$\widehat{\nabla J}(\theta) = \frac{J(\theta + \delta d)}{\delta} d. \quad (7)$$

By analogous Taylor series argument, we obtain the error between the estimate and the true gradient as

$$\begin{aligned} & \frac{J(\theta + \delta d)}{\delta} d - \nabla J(\theta) \\ &= \frac{J(\theta)}{\delta} d + (d d^T - I) \nabla J(\theta) + O(\delta). \end{aligned} \quad (8)$$

From (8), we require the following to hold in addition to (6) in the case of random perturbations for the one simulation version of RDKW algorithm, i.e.,

$$\mathbb{E}[d] = 0. \quad (9)$$

For the random perturbations, $d \sim F$, F is any distribution that satisfies (6) and (9), the noise in the gradient estimates gets averaged asymptotically. An example distribution for F is symmetric Bernoulli where each component of the perturbation vector is ± 1 with equal probability.

From (6) and (9) clearly one is motivated to look for perturbations that satisfy similar properties. In what follows, the sequence of deterministic perturbations (that will be used in either (3) or (7)) will be denoted by $\{d_n\}_{n \geq 1}$ and we require the following two properties to hold for the perturbation sequence d_n for the almost sure convergence of the iterates to a local minima.

P1. Let $D_n := d_n d_n^T - I_{p \times p}$. For any $s \in \mathbb{N}$ there exists a $P \in \mathbb{N}$ such that $\sum_{n=s+1}^{s+P} D_n = 0$ and,

P2. $\sum_{n=s+1}^{s+P} d_n = 0$.

Remark 1. The properties **P1** and **P2** are the deterministic analogues of (6) and (9). For the properties **P1** and **P2** to hold, it is sufficient to determine a finite sequence $\{d_1, d_2, \dots, d_P\}$ such that $\sum_{n=1}^P d_n d_n^T = PI$ and $\sum_{n=1}^P d_n = 0$ and for $n \geq P+1$, periodically cycle through this sequence, i.e., set $d_n = d_{n \% P+1}$. We will refer the length of the deterministic perturbation sequence P as the cycle length.

III. CONSTRUCTION OF DETERMINISTIC PERTURBATIONS

In section III-A, following Remark 1, we first characterize the finite sequences $\{d_1, d_2, \dots, d_P\}$ that satisfy properties **P1** and **P2** by providing a matrix equation whose solution gives the deterministic perturbations. In Section III-B, we then construct a specific sequence using a circulant matrix that has the least possible cycle length among all the deterministic perturbation sequences. Finally in section III-C, we completely describe the RDKW algorithm that uses the deterministic perturbation sequence constructed using the circulant matrix approach.

A. Matrix condition for Deterministic Perturbations

The properties **P1** and **P2** can be satisfied individually. For example, to satisfy property **P1**, let $P = p$ and $d_n = \sqrt{p} e_n$, $n \in \{1, \dots, P\}$, the scaled canonical basis vectors, then $\sum_{n=1}^P d_n d_n^T = \sum_{n=1}^P p e_n e_n^T = pI$. To satisfy

property **P2**, consider any set of linearly dependent vectors $\{v_0, \dots, v_P\}$. Then there exists scalars $\alpha_1, \dots, \alpha_P$ such that $\sum_{n=1}^P \alpha_n v_n = 0$. Now for the choice $d_n = \alpha_n v_n$ the property **P2**, $\sum_{n=1}^P d_n = \sum_{n=1}^P \alpha_n v_n = 0$ is trivially satisfied. A natural question would be to determine sequences $\{d_n\}_{1 \leq n \leq P}$ that satisfy both the properties simultaneously.

To address this problem, let us consider a $p \times P$ matrix Y as follows: $Y := \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ d_1 & d_2 & \dots & d_P \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix}$. Let $u = [1, 1, \dots, 1]^T$ be a $P \times 1$ dimension vector. The perturbations that satisfy properties **P1** and **P2** essentially solve the two matrix equations $Yu = 0$ and $YY^T = PI$. These equations can be compactly written in a single matrix equation as

$$XX^T = PI_{(p+1) \times (p+1)}, \quad (10)$$

where $X = \begin{bmatrix} u^T \\ Y \end{bmatrix}$. Note that $Y_{p \times P}$ and P are the unknowns here.

It can be observed from (10) that $\frac{X}{\sqrt{P}}$ could be treated as a $p \times P$ submatrix of a $P \times P$ orthogonal matrix with the first row being $\frac{u^T}{\sqrt{P}}$, a $1 \times P$ vector. It has been shown in [13] that columns of Hadamard matrices satisfy properties **P1** and **P2** simultaneously with $\bar{P} = 2^{\lceil \log_2 p \rceil}$, i.e., X is chosen as a $(p+1) \times 2^{\lceil \log_2 p \rceil}$ submatrix of the Hadamard matrix. It is not in general clear if the equation (10) can be solved for a smaller $P \leq \bar{P}$.

Remark 2. We note that similar analysis for matrix condition for the construction of deterministic perturbations for SPSA estimates involves solving the following matrix system. $AB = PI, Au = 0$ and $A \circ B^T = vu^T$ where A is $p \times P$, B is $P \times p$, u is $P \times 1$ vector of ones, v is $p \times 1$ vector of ones and \circ denotes the Hadamard product of the matrices A and B . It is not clear how to solve for P , A and B due to the presence of Hadamard product in this system.

B. Specific Perturbation Sequence Construction

In this section, our goal is to obtain a sequence with least cycle length. Using a simple matrix rank argument it can be shown that P is at least $p+1$. Thus, in what follows, we give a construction of deterministic perturbation sequence with cycle length $P = p+1$. We first write

$$Y = \begin{bmatrix} \uparrow & \dots & \uparrow \\ Z & & -ZU \\ \downarrow & \dots & \downarrow \end{bmatrix}$$

where Z is a $p \times p$ matrix and U is any $p \times (P-p)$ matrix with columns that sum to 1. Clearly $Yu = 0$ satisfies property **P2**.

To satisfy property **P1**, i.e., $YY^T = I$ is equivalent to

$$ZZ^T + ZUU^T Z^T = Z(I + UU^T)Z^T = PI. \quad (11)$$

Clearly construction of deterministic perturbations with smaller cycle length P is equivalent to solving for Z with an appropriate choice of U .

The simplest choice of U with column sums being 1 is $U = u$, a $p \times 1$ vector, thus $P = p+1$. Let $C = I + UU^T =$

$I + uu^T$ ($p \times p$ dimensional matrix)

$$C = \begin{bmatrix} 2 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ & & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 2 \end{bmatrix}. \quad (12)$$

Observe that C is a positive definite circulant matrix. Hence $C^{-1/2}$ is well defined and the choice $Z = C^{-1/2}$ satisfies (11) and solves the system $YY^T = I$ with $P = p + 1$, i.e.,

$$Y = \sqrt{p+1}[C^{-1/2}, -C^{-1/2}u]. \quad (13)$$

The columns of Y finally give us the deterministic perturbations. We note that in general the computation of $C^{-1/2}$ is $O(p^3)$ and can be very expensive for large p . However owing to the special structure of C , using a Sherman-Morrison type result (see Lemma 1, Section IV), $C^{-1/2}$ can be computed in $O(p^2)$ time complexity.

C. Gradient estimation

In this section, we present the RDKW algorithms that use the deterministic perturbation sequence constructed above in two-simulation and one-simulation gradient estimates of the objective. We denote the corresponding algorithms by DSPKW-2C and DSPKW-1C respectively.

Algorithm 1 Basic structure of DSPKW.

1: **Input:**

- $\theta_0 \in \mathbb{R}^p$, initial parameter vector
- $\delta_n, n \geq 0$, a sequence of sensitivity parameters to approximate gradient
- Matrix of perturbations

$$Y = \sqrt{p+1}[C^{-1/2}, -C^{-1/2}u],$$

with $u = [1, 1, \dots, 1]^T$;

- noisy measurements of cost objective J
- $a_n, n \geq 0$, step-size sequence satisfying assumption **A2**. (see section IV)
- n_{end} , the total number of iterations determined by simulation budget

2: **Output:** $\theta_{n_{end}}$, approximate local optimal solution

3: **for** $n = 1, 2, \dots, n_{end}$ **do**

4: Let d_n be the $\text{mod}(n, p + 1)$ th column of Y .

5: Update the parameter as follows:

$$\theta_{n+1} = \theta_n - a_n \widehat{\nabla J}(\theta_n)$$

$\widehat{\nabla J}(\theta_n)$ is chosen according to either (14) or (15) for DSPKW-2C and DSPKW-1C respectively.

6: **end for**

7: **Return** $\theta_{n_{end}}$

Let $\delta_n, n \geq 0$ denote a sequence of diminishing positive real numbers satisfying assumption **A2**. in section IV. Let y_n^+, y_n^- denote the noisy objective function evaluations at the perturbed parameters $\theta_n + \delta_n d_n$ and $\theta_n - \delta_n d_n$ respectively, i.e., $y_n^+ = J(\theta_n + \delta_n d_n) + M_{n+1}^+$ and $y_n^- =$

$J(\theta_n - \delta_n d_n) + M_{n+1}^-$. We assume the noise terms M_n^+, M_n^- are martingale difference noise sequence, $\mathbb{E}[M_{n+1}^+ | \mathcal{F}_n] = \mathbb{E}[M_{n+1}^- | \mathcal{F}_n] = 0$ where $\mathcal{F}_n = \sigma(\theta_m, M_m^+, M_m^-, m \leq n)$ is the information conditioned on the past parameter values and martingale difference terms.

The two-simulation and one-simulation estimates of the gradient $\nabla J(\theta_n)$ based on the observed noisy objective samples for the RDKW algorithm are respectively given by

$$\widehat{\nabla J}(\theta_n) = \left[\frac{(y_n^+ - y_n^-)d_n}{2\delta_n} \right], \quad (14)$$

$$\widehat{\nabla J}(\theta_n) = \left[\frac{(y_n^+)d_n}{\delta_n} \right], \quad (15)$$

respectively. Observe that in the two-sided estimate (14) we use two function samples y_n^+ and y_n^- and the estimate in (15) uses only one function sample y_n^+ .

Now we briefly describe the DSPKW algorithm. Inputs to the DSPKW algorithm are randomly chosen initial point θ_0 , diminishing sequences δ_n and a_n satisfying assumption **A2**. and the matrix of deterministic perturbations Y chosen according to (13). In our algorithms, we iteratively choose the perturbations by cycling through columns of Y with period $p + 1$ and in steps 2-4, we update the parameters along the direction of estimated gradient according to (14) in the DSPKW-2C algorithm and according to (15) in the DSPKW-1C algorithm. Note the choice of gradient estimate (or the algorithm) is dictated by the simulation budget given to us. The algorithms terminate by returning the parameter $\theta_{n_{end}}$ at the end of n_{end} iterations.

IV. CONVERGENCE ANALYSIS

In this section we first provide a few lemmas that assist in computing the proposed deterministic perturbation sequence (see (13) in Section III-B). In the latter part of the section, we prove the almost sure convergence of the iterates for the class of deterministic perturbations characterized in Section III-A.

The following lemma is useful in obtaining the negative square root of C , i.e., $C^{-1/2}$ in a computationally efficient manner. Also note that it takes only $O(p^2)$ operations to compute $C^{-1/2}$ using the lemma and the circulant structure of $C^{-1/2}$. Note that the following lemma could also be utilized in an independent context for efficient computation.

Lemma 1. *Let I be a $p \times p$ identity matrix and $u = [1, 1, \dots, 1]^T$ be a $p \times 1$ column vector of 1s, then*

$$(I + uu^T)^{-1/2} = I - \frac{uu^T}{p} + \frac{uu^T}{p\sqrt{(1+p)}}.$$

Proof. It is enough to show that

$$(I + uu^T) \left[I - \frac{uu^T}{p} + \frac{uu^T}{p\sqrt{(1+p)}} \right]^2 = I.$$

Using $\|u\|^2 = u^T u = p$ in the expansion of $\left[I - \frac{uu^T}{p} + \frac{uu^T}{p\sqrt{(1+p)}} \right]^2$ gives the result. \square

Let C be defined as in (12) and $Y = \sqrt{p+1}[C^{-1/2}, -C^{-1/2}u]$. Let the perturbations d_n be the columns of Y .

Lemma 2. *The perturbations d_n chosen as columns of Y satisfy properties **P1** and **P2**.*

Proof. It easily follows from the discussion in section III-B on the construction of this specific perturbation sequence. \square

In what follows, we prove the almost sure convergence of the iterates in the DSPKW algorithm (Section III-C) under the following assumptions. Note that $\|\cdot\|$ denotes the 2-norm.

- A1.** The map $J : \mathbb{R}^p \rightarrow \mathbb{R}$ is Lipschitz continuous and is differentiable with bounded second order derivatives. Further, the map $L : \mathbb{R}^p \rightarrow \mathbb{R}^p$ defined as $L(\theta) = -\nabla J(\theta)$ is Lipschitz continuous.
- A2.** The step-size sequences $a_n, \delta_n > 0, \forall n$ satisfy

$$a_n, \delta_n \rightarrow 0, \sum_n a_n = \infty, \sum_n \left(\frac{a_n}{\delta_n}\right)^2 < \infty.$$

Further, $\frac{a_j}{\delta_n} \rightarrow 1$ as $n \rightarrow \infty$, for all $j \in \{n, n+1, n+2, \dots, n+M\}$ for any given $M > 0$ and $b_n = \frac{a_n}{\delta_n}$ is such that $\frac{b_j}{b_n} \rightarrow 1$ as $n \rightarrow \infty$, for all $j \in \{n, n+1, n+2, \dots, n+M\}$.

- A3.** $\max_n \|d_n\| = K_0, \max_n \|D_n\| = K_1$.
- A4.** The iterates θ_n remain uniformly bounded almost surely, i.e., $\sup_n \|\theta_n\| < \infty$, a.s.
- A5.** The ODE $\dot{\theta}(t) = -\nabla J(\theta(t))$ has a compact set $G \subset \mathbb{R}^p$ as its set of asymptotically stable equilibria (i.e., the set of local minima of J is compact).
- A6.** The sequences $(M_n^+, \mathcal{F}_n), (M_n^-, \mathcal{F}_n), n \geq 0$ form martingale difference sequences. Further, $(M_n^+, M_n^-, n \geq 0)$ are square integrable random variables satisfying

$$\mathbb{E}[\|M_{n+1}^\pm\|^2 | \mathcal{F}_n] \leq K(1 + \|\theta_n\|^2) \text{ a.s., } \forall n \geq 0,$$

for a given constant $K > 0$.

Remark 3. *Assumptions **A1**, **A2** and **A5** are standard stochastic approximation conditions. Assumption **A3** trivially follows from Remark 1. Assumption **A4** is the stability condition on the iterates and holds in many applications [7] (see the discussion in pp 40-41 of [3]). This condition can also be enforced by projecting the iterates into a compact set, however, the iterates converge to a limiting set that contains all possible limit points (see pp.191 in [3]). Assumption **A6** gives the condition on the maximum strength of the martingale difference noise under which convergence of the iterates could be ensured and in many stochastic optimization settings this condition could be easily verified using Jensen's inequality and Lipschitz continuity of ∇J .*

The following two lemmas aid in the proof of almost sure convergence of the iterates in the DSPKW algorithm.

Lemma 3. *Given any fixed integer $P > 0$, $\|\theta_{m+k} - \theta_m\| \rightarrow 0$ w.p.1, as $m \rightarrow \infty$, for all $k \in \{1, \dots, P\}$.*

Proof. Fix a $k \in \{1, \dots, P\}$. Now

$$\begin{aligned} \theta_{n+k} &= \theta_n - \sum_{j=n}^{n+k-1} a_j \left(\frac{J(\theta_j + \delta_j d_j) - J(\theta_j - \delta_j d_j)}{2\delta_j} \right) d_j \\ &\quad - \sum_{j=n}^{n+k-1} a_j M_{j+1}, \end{aligned}$$

where $M_{j+1} = \frac{(M_{j+1}^+ - M_{j+1}^-)d_j}{2\delta_j}$. Thus,

$$\begin{aligned} \|\theta_{n+k} - \theta_n\| &\leq \sum_{j=n}^{n+k-1} a_j \left| \frac{J(\theta_j + \delta_j d_j) - J(\theta_j - \delta_j d_j)}{2\delta_j} \right| \|d_j\| \\ &\quad + \sum_{j=n}^{n+k-1} a_j \|M_{j+1}\|. \end{aligned}$$

Now clearly, $N_n = \sum_{j=0}^{n-1} a_j M_{j+1}, n \geq 1$, forms a martingale sequence with respect to the filtration $\{\mathcal{F}_n\}$. Further, from the assumption (A6) we have,

$$\begin{aligned} \sum_{m=0}^n \mathbb{E}[\|N_{m+1} - N_m\|^2 | \mathcal{F}_m] &= \sum_{m=0}^n \mathbb{E}[a_m^2 \|M_{m+1}\|^2 | \mathcal{F}_m] \\ &\leq \sum_{m=0}^n a_m^2 K(1 + \|\theta_m\|^2). \end{aligned}$$

From the assumption (A4), the quadratic variation process of $N_n, n \geq 0$ converges almost surely. Hence by the martingale convergence theorem, it follows that $N_n, n \geq 0$ converges almost surely. Hence $\left\| \sum_{j=n}^{n+k-1} a_j M_{j+1} \right\| \rightarrow 0$ almost surely as $n \rightarrow \infty$. Moreover

$$\begin{aligned} &\left\| \left(J(\theta_j + \delta_j d_j) - J(\theta_j - \delta_j d_j) \right) d_j \right\| \\ &\leq \left| \left(J(\theta_j + \delta_j d_j) - J(\theta_j - \delta_j d_j) \right) \right| \|d_j\| \\ &\leq K_0 \left(|J(\theta_j + \delta_j d_j)| + |J(\theta_j - \delta_j d_j)| \right), \end{aligned}$$

since $\|d_j\| \leq K_0, \forall j \geq 0$. Note that

$$\begin{aligned} |J(\theta_j + \delta_j d_j)| - |J(0)| &\leq |J(\theta_j + \delta_j d_j) - J(0)| \\ &\leq \hat{B} \|\theta_j + \delta_j d_j\|, \end{aligned}$$

where \hat{B} is the Lipschitz constant of the function J . Hence,

$$|J(\theta_j + \delta_j d_j)| \leq \tilde{B}(1 + \|\theta_j + \delta_j d_j\|),$$

for $\tilde{B} = \max(|J(0)|, \hat{B})$. Similarly,

$$|J(\theta_j - \delta_j d_j)| \leq \tilde{B}(1 + \|\theta_j - \delta_j d_j\|).$$

From assumption (A1), it follows that

$$\sup_j \left\| \left(J(\theta_j + \delta_j d_j) - J(\theta_j - \delta_j d_j) \right) d_j \right\| \leq \tilde{K} < \infty,$$

for some $\tilde{K} > 0$. Thus,

$$\begin{aligned} \|\theta_{n+k} - \theta_n\| &\leq \tilde{K} \sum_{j=n}^{n+k-1} \frac{a_j}{2\delta_j} + \left\| \sum_{j=n}^{n+k-1} a_j M_{j+1} \right\| \\ &\rightarrow 0 \text{ a.s. with } n \rightarrow \infty, \text{ proving the lemma. } \square \end{aligned}$$

Lemma 4. For any $m \geq 0$, $\left\| \sum_{n=m}^{m+P-1} \frac{a_n}{a_m} D_n \nabla J(\theta_n) \right\|$ and $\left\| \sum_{n=m}^{m+P-1} \frac{b_n}{b_m} d_n J(\theta_n) \right\| \rightarrow 0$, almost surely, as $m \rightarrow \infty$.

Proof. From Lemma 3, it can be seen that $\|\theta_{m+s} - \theta_m\| \rightarrow 0$ as $m \rightarrow \infty$, for all $s = 1, \dots, P$. Also, from assumption (A1), we have $\|\nabla J(\theta_{m+s}) - \nabla J(\theta_m)\| \rightarrow 0$ as $m \rightarrow \infty$, for all $s = 1, \dots, P$. Now from Lemma 2, $\sum_{n=m}^{m+P-1} D_n = 0$

$\forall m \geq 0$. Hence $D_m = -\sum_{n=m+1}^{m+P-1} D_n$. Consider first

$$\begin{aligned} & \left\| \sum_{n=m}^{m+P-1} \frac{a_n}{a_m} D_n \nabla J(\theta_n) \right\| \\ &= \left\| \sum_{n=m+1}^{m+P-1} \frac{a_n}{a_m} D_n \nabla J(\theta_n) + D_m \nabla J(\theta_m) \right\| \\ &= \left\| \sum_{n=m+1}^{m+P-1} \frac{a_n}{a_m} D_n \nabla J(\theta_n) - \sum_{n=m+1}^{m+P-1} D_n \nabla J(\theta_m) \right\| \\ &= \left\| \sum_{n=m+1}^{m+P-1} D_n \left(\frac{a_n}{a_m} \nabla J(\theta_n) - \nabla J(\theta_m) \right) \right\| \\ &\leq \sum_{n=m+1}^{m+P-1} \|D_n\| \left\| \left(\frac{a_n}{a_m} \nabla J(\theta_n) - \nabla J(\theta_m) \right) \right\| \\ &\leq K_1 \sum_{n=m+1}^{m+P-1} \left\| \left(\frac{a_n}{a_m} - 1 \right) \nabla J(\theta_n) \right\| + \left\| \nabla J(\theta_n) - \nabla J(\theta_m) \right\| \end{aligned}$$

$\rightarrow 0$ a.s. with $n \rightarrow \infty$, from assumptions (A1) and (A2). Now observe that $\|J(\theta_{m+k}) - J(\theta_m)\| \rightarrow 0$ as $m \rightarrow \infty$, for all $k \in \{1, \dots, P\}$ as a consequence of (A1) and Lemma 3.

Moreover from $d_m = -\sum_{n=m+1}^{m+P-1} d_n$ we have

$$\begin{aligned} & \left\| \sum_{n=m}^{m+P-1} \frac{b_n}{b_m} d_n J(\theta_n) \right\| \\ &= \left\| \sum_{n=m+1}^{m+P-1} \frac{b_n}{b_m} d_n J(\theta_n) + d_m J(\theta_m) \right\| \\ &= \left\| \sum_{n=m+1}^{m+P-1} \frac{b_n}{b_m} d_n J(\theta_n) - \sum_{n=m+1}^{m+P-1} d_n J(\theta_m) \right\| \\ &= \left\| \sum_{n=m+1}^{m+P-1} d_n \left(\frac{b_n}{b_m} J(\theta_n) - J(\theta_m) \right) \right\| \\ &\leq \sum_{n=m+1}^{m+P-1} \|d_n\| \left\| \left(\frac{b_n}{b_m} J(\theta_n) - J(\theta_m) \right) \right\| \\ &\leq K_0 \sum_{n=m+1}^{m+P-1} \left\| \left(\frac{b_n}{b_m} - 1 \right) J(\theta_n) \right\| + \left\| \left(J(\theta_n) - J(\theta_m) \right) \right\| \end{aligned}$$

The claim now follows as a consequence of assumptions (A1) and (A2). \square

Finally, using the following theorems, we conclude the analysis by proving the almost sure convergence of the

iterates to the set of local minima G of the function J .

Theorem 5. $\theta_n, n \geq 0$ obtained from DSPKW-2C satisfy $\theta_n \rightarrow G$ almost surely.

Proof. Note that

$$\theta_{n+P} = \theta_n - \sum_{l=n}^{n+P-1} a_l \left[\frac{J(\theta_l + \delta_l d_l) - J(\theta_l - \delta_l d_l)}{2\delta_l} d_l + M_{l+1} \right].$$

It follows that

$$\begin{aligned} \theta_{n+P} &= \theta_n - \sum_{l=n}^{n+P-1} a_l \nabla J(\theta_l) - \sum_{l=n}^{n+P-1} a_l o(\delta_l) \\ &\quad - \sum_{l=n}^{n+P-1} a_l (d_l d_l^T - I) \nabla J(\theta_l) - \sum_{l=n}^{n+P-1} a_l M_{l+1}. \end{aligned}$$

Now the fourth term on the RHS above can be written as

$$a_n \sum_{l=n}^{n+P-1} \frac{a_l}{a_n} D_l \nabla J(\theta_l) = a_n \xi_n,$$

where $\xi_n = o(1)$ from Lemma 4. Thus, the algorithm is asymptotically analogous to

$$\theta_{n+1} = \theta_n - a_n (\nabla J(\theta_n) + o(\delta) + M_{n+1}).$$

Hence, from Theorem 2 in chapter 2 of [?], it follows that $\theta_n, n \geq 0$ converge to a local minima of the function J . \square

Theorem 6. $\theta_n, n \geq 0$ obtained from DSPKW-1C satisfy $\theta_n \rightarrow G$ almost surely.

Proof. Note that

$$\theta_{n+P} = \theta_n - \sum_{l=n}^{n+P-1} a_l \left(\frac{J(\theta_l + \delta_l d_l)}{2\delta_l} \right) d_l - \sum_{l=n}^{n+P-1} a_l M_{l+1}.$$

It follows that

$$\begin{aligned} \theta_{n+P} &= \theta_n - \sum_{l=n}^{n+P-1} a_l \nabla J(\theta_l) - \sum_{l=n}^{n+P-1} a_l \frac{J(\theta_l)}{\delta_l} d_l \\ &\quad - \sum_{l=n}^{n+P-1} a_l (d_l d_l^T - I) \nabla J(\theta_l) - \sum_{l=n}^{n+P-1} a_l o(\delta_l) \\ &\quad - \sum_{l=n}^{n+P-1} a_l M_{l+1}. \end{aligned}$$

Now we observe that the third term on the RHS above is

$$\begin{aligned} & \sum_{l=n}^{n+P-1} a_l \frac{J(\theta_l)}{\delta_l} d_l = \sum_{l=n}^{n+P-1} b_l J(\theta_l) d_l \\ &= b_n \sum_{l=n}^{n+P-1} \frac{b_l}{b_n} \frac{J(\theta_l)}{\delta_l} d_l = b_n \xi_n^1, \end{aligned}$$

where $\xi_n^1 = o(1)$ by Lemma 4. Similarly

$$\sum_{l=n}^{n+P-1} a_l (d_l d_l^T - I) \nabla J(\theta_l) = a_n \xi_n^2,$$

with $\xi_n^2 = o(1)$ by Lemma 4. The rest follows as in Theorem 5. \square

V. SIMULATION EXPERIMENTS

In this section, we compare the numerical performance of our DSPKW-2C algorithm against the RDKW algorithm that uses random Bernoulli perturbations and another variant of the RDKW algorithm that uses Hadamard matrix based deterministic perturbations. We refer them by the acronyms RDKW-2R and RDKW-2H respectively. In a similar manner, we also compare DSPKW-1C algorithm against the one-simulation variants RDKW-1R and RDKW-1H. Note that 2 or 1 in the acronyms of these algorithms denote the number of simulations utilized per iteration.¹

A. Experimental setup

For the empirical performance evaluation, we consider the following two loss functions:

a) *Quadratic loss*:

$$J(\theta) = \theta^T A \theta + b^T \theta. \quad (16)$$

b) *Fourth-order loss*:

$$J(\theta) = \theta^T A^T A \theta + 0.1 \sum_{j=1}^N (A\theta)_j^3 + 0.01 \sum_{j=1}^N (A\theta)_j^4. \quad (17)$$

In the loss functions considered above, we set the dimension $p = 10$. We choose A such that pA is an upper triangular matrix with each nonzero entry equal to one and b is a p -dimensional vector of ones. In our experiments, we follow the same noise assumptions considered in [16], i.e., for any θ , the additive noise in the objective is given by $[\theta^T, 1]z$ where $z \sim \mathcal{N}(0, \sigma^2 I_{p+1 \times p+1})$. In all algorithms, we set the step-size schedule as $\delta_n = c/(n+1)^\gamma$ and $a_n = 1/(n+B+1)^\alpha$ with $\alpha = 0.602$ and $\gamma = 0.101$. Note that the chosen values for α and γ have demonstrated good finite-sample performance empirically, while satisfying the theoretical requirements needed for asymptotic convergence (see [16]). We set the same initial point θ_0 for all the algorithms.

We consider two settings in our experiments. In the first noise-free setting, we do not add any noise to the objective function evaluations and in the second setting, we corrupt the function evaluations by adding noise (with variance parameter $\sigma = 0.01$ as described above). We evaluate the performance of these algorithms based on Normalized Mean Square Error (NMSE) metric. NMSE is defined as the ratio $\|\theta_{n_{\text{end}}} - \theta^*\|^2 / \|\theta_0 - \theta^*\|^2$, where $\theta_{n_{\text{end}}}$ is the parameter returned by the algorithm.

B. Discussion of Results

The performance comparisons of all the algorithms based on NMSE values are summarized in Tables I, II, III and IV. In the tables, we have highlighted the algorithm that has the minimum NMSE. We summarize our findings:

- Even in the absence of noise, due to the random directions chosen by RDKW-2R and RDKW-1R algorithms, the standard deviation is significantly high compared to the corresponding deterministic counterparts.

¹The implementation is available at <https://github.com/cs1070166/1RDSA-2Cand1RDSA-1C/>

Noise parameter $\sigma = 0$	
Method	NMSE
RDKW-2R	$5.755 \times 10^{-3} \pm 2.460 \times 10^{-3}$
RDKW-2H	$1.601 \times 10^{-5} \pm 2.724 \times 10^{-20}$
DSPKW-2C	$2.474 \times 10^{-8} \pm 1.995 \times 10^{-23}$
Noise parameter $\sigma = 0.01$	
Method	NMSE
RDKW-2R	$5.762 \times 10^{-3} \pm 2.473 \times 10^{-3}$
RDKW-2H	$4.012 \times 10^{-5} \pm 1.654 \times 10^{-5}$
DSPKW-2C	$2.188 \times 10^{-5} \pm 9.908 \times 10^{-6}$

TABLE I: NMSE values of two-simulation methods for the quadratic objective (16) without and with noise for 2000 simulations: standard deviation of 100 replications shown after \pm symbol

Noise parameter $\sigma = 0$	
Method	NMSE
RDKW-2R	$2.747 \times 10^{-2} \pm 1.413 \times 10^{-2}$
RDKW-2H	$3.901 \times 10^{-3} \pm 4.359 \times 10^{-18}$
DSPKW-2C	$3.535 \times 10^{-3} \pm 1.743 \times 10^{-18}$
Noise parameter $\sigma = 0.01$	
Method	NMSE
RDKW-2R	$2.762 \times 10^{-2} \pm 1.415 \times 10^{-2}$
RDKW-2H	$3.958 \times 10^{-3} \pm 4.227 \times 10^{-4}$
DSPKW-2C	$3.598 \times 10^{-3} \pm 4.158 \times 10^{-4}$

TABLE II: NMSE values of two-simulation methods for the fourth order objective (17) without and with noise for 10000 simulations: standard deviation of 100 replications shown after \pm symbol

Noise parameter $\sigma = 0$	
Method	NMSE
RDKW-1R	$8.584 \times 10^{-2} \pm 3.681 \times 10^{-2}$
RDKW-1H	$2.770 \times 10^{-2} \pm 3.836 \times 10^{-17}$
DSPKW-1C	$8.225 \times 10^{-3} \pm 1.569 \times 10^{-17}$
Noise parameter $\sigma = 0.01$	
Method	NMSE
RDKW-1R	$8.582 \times 10^{-2} \pm 3.691 \times 10^{-2}$
RDKW-1H	$2.774 \times 10^{-2} \pm 2.578 \times 10^{-4}$
DSPKW-1C	$8.225 \times 10^{-3} \pm 5.959 \times 10^{-5}$

TABLE III: NMSE values of one-simulation methods for the quadratic objective (16) without and with noise for 20000 simulations: standard deviation of 100 replications shown after \pm symbol

Noise parameter $\sigma = 0$	
Method	NMSE
RDKW-1R	$3.192 \times 10^{-1} \pm 1.991 \times 10^{-1}$
RDKW-1H	$8.173 \times 10^{-2} \pm 1.255 \times 10^{-16}$
DSPKW-1C	$4.403 \times 10^{-2} \pm 9.066 \times 10^{-17}$
Noise parameter $\sigma = 0.01$	
Method	NMSE
RDKW-1R	$3.240 \times 10^{-1} \pm 1.836 \times 10^{-1}$
RDKW-1H	$8.916 \times 10^{-2} \pm 1.896 \times 10^{-2}$
DSPKW-1C	$4.972 \times 10^{-2} \pm 9.812 \times 10^{-3}$

TABLE IV: NMSE values of one-simulation methods for the fourth order objective (17) without and with noise for 20000 simulations: standard deviation of of 100 replications shown after \pm symbol

- We would like to emphasize that the quality of the solution (characterized by standard deviation) is significantly better for the case of proposed deterministic perturbations compared to the existing Hadamard based deterministic perturbations and random perturbations. Note however that we do not make comparisons between two-simulation and one-simulation algorithms.
- In the case of two simulation algorithms (see Tables I and II), DSPKW-2C performs marginally better than RDKW-2H, while both of them outperform RDKW-2R significantly.
- In the case of one simulation algorithms (see Tables III and IV), DSPKW-1C performs better than both RDKW-1H and RDKW-1R.

VI. CONCLUSIONS

We have generalized the deterministic perturbation sequences from lexicographical ordering and Hadamard matrix based constructions for the RDKW algorithm and presented a novel construction of deterministic perturbations that has least cycle length within the class of deterministic perturbation sequences. Further, we have proved the almost sure convergence of the iterates for the class of deterministic perturbation sequences. Now that we have a characterization of the class of deterministic perturbation sequences, it would be interesting as future work, to theoretically study and compare the rate of convergence of deterministic perturbation algorithms against their random perturbation counterparts. A challenging future direction would be to study the asymptotic normality or weak convergence of the iterates. It would also be interesting to similarly characterize the class of deterministic perturbation sequences for the SPSA algorithm.

REFERENCES

[1] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 462–466, 09 1952. [Online]. Available: <http://dx.doi.org/10.1214/aoms/1177729392>

[2] 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.

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

[4] Y. M. Ermol'Ev, "On the method of generalized stochastic gradients and quasi-fejér sequences," *Cybernetics*, vol. 5, no. 2, pp. 208–220, 1969.

[5] M. Styblinski and T.-S. Tang, "Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing," *Neural Networks*, vol. 3, no. 4, pp. 467–483, 1990.

[6] L. Prashanth, S. Bhatnagar, M. Fu, and S. Marcus, "Adaptive system optimization using random directions stochastic approximation," *IEEE Transactions on Automatic Control*, vol. 62, no. 5, pp. 2223–2238, 2017.

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

[8] 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.

[9] J. Theiler and J. Alper, "On the choice of random directions for stochastic approximation algorithms," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 476–481, 2006.

[10] J. C. Spall, "A one-measurement form of simultaneous perturbation stochastic approximation," *Automatica*, vol. 33, no. 1, pp. 109–112, 1997.

[11] S. Sandilya and S. Kulkarni, "Deterministic sufficient conditions for convergence of simultaneous perturbation stochastic approximation algorithms," in *Proceedings of the 9th INFORMS Applied Probability Conference*, 1997.

[12] I.-J. Wang and E. K. Chong, "A deterministic analysis of stochastic approximation with randomized directions," *IEEE Transactions on Automatic Control*, vol. 43, no. 12, pp. 1745–1749, 1998.

[13] S. Bhatnagar, M. C. Fu, S. I. Marcus, I. Wang *et al.*, "Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences," *ACM TOMACS*, vol. 13, no. 2, pp. 180–209, 2003.

[14] X. Xiong and I.-J. Wang, "Randomized-direction stochastic approximation algorithms using deterministic sequences," in *Simulation Conference, 2002. Proceedings of the Winter*, vol. 1. IEEE, 2002, pp. 285–291.

[15] A. Hedayat, N. Sloane, and J. Stufken, "Orthogonal arrays: theory and applications," Springer, 1999.

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