On the Choice of Random Directions for Stochastic Approximation Algorithms

James Theiler and Jarod Alper

Abstract

Stochastic approximation provides a simple and effective approach for finding roots and minima of functions whose evaluations are contaminated with noise. We investigate variants of the Random Direction Stochastic Approximation (RDSA) algorithm for optimizing noisy loss functions in high-dimensional spaces. The most popular variant is random selection from a Bernoulli distribution, which also goes by the name Simultaneous Perturbation Stochastic Approximation (SPSA). Viable alternatives include an axis-aligned distribution, normal distribution, and uniform distribution on a spherical shell. Although there are special cases where the Bernoulli distribution is optimal, we identify other cases where it performs worse than the alternatives. We show that performance depends on the orientation of the loss function with respect to its coordinate axes, and consider averages over all orientations. We find that the average asymptotic performance depends only on the radial fourth moment of the distribution of random directions, and is identical for the Bernoulli, axis-aligned, and spherical shell distributions. Of these variants, the spherical shell is optimal in the sense of minimum variance over random orientations of the loss function with respect to the coordinate axes. We also show that for unaligned loss functions, the performance of the Kiefer-Wolfowitz-Blum Finite Difference Stochastic Approximation (FDSA) is asymptotically equivalent to the RDSA algorithms, and we observe numerically that the pre-asymptotic performance of FDSA is often superior.

Index Terms

stochastic approximation, random direction, finite difference, simultaneous perturbation

This work was supported by the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory.

J. Theiler is a technical staff member in the Space and Remote Sensing Sciences Group, MS-B244, Los Alamos National Laboratory, Los Alamos, NM 87545. Email: jt@lanl.gov (Phone: 505-665-5682; fax: 505-665-4414)

J. Alper graduated from Brown University with a degree in Computer Science; he was graduate research assistant at Los Alamos when this work was done. Email: Jarod_Alper@alumni.brown.edu.

I. INTRODUCTION

Consider a smooth[1] p-dimensional loss function $L : \mathbb{R}^p \to \mathbb{R}$, with gradient $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^p$. Assume that L has a unique minimum $\mathbf{x}^* \in \mathbb{R}^p$; that is, $L(\mathbf{x}^*) \leq L(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$, and $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ iff $\mathbf{x} = \mathbf{x}^*$. If $\hat{\mathbf{g}}(\mathbf{x})$ is an unbiased estimator of the gradient, – that is, if $E\{\,\hat{\mathbf{g}}(\mathbf{x}) \mid \mathbf{x}\,\} = \mathbf{g}(\mathbf{x})$, – then the Robbins-Monro [1] algorithm estimates a root of $\mathbf{g}(\mathbf{x})$ with the following recursion:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \hat{\mathbf{g}}(\mathbf{x}_k). \tag{1}$$

If $a_k = a_o k^{-\alpha}$ with $0 < \alpha \leq 1$, then $\mathbf{x}_k$ approaches $\mathbf{x}^*$ with a $k^{-\alpha/2}$ scaling rate. Here, $\alpha = 1$ gives the fastest asymptotic convergence.

For the case when a direct unbiased estimator of gradient is unavailable, Kiefer and Wolfowitz [2], and Blum [3], introduced the finite difference stochastic approximation (FDSA) algorithm, which uses an estimator for the gradient whose ith component is

$$\hat{g}_i(\mathbf{x}) = \frac{\hat{L}(\mathbf{x} + c\mathbf{e}_i) - \hat{L}(\mathbf{x} - c\mathbf{e}_i)}{2c}, \tag{2}$$

where $\mathbf{e}_i$ is the unit vector along the ith axis, and $\hat{L}$ is a noisy measurement of the loss function. For $c > 0$, Eq. (2) is in general a *biased* estimator of the gradient. Convergence is achieved with a decreasing sequence $c_k \to 0$ so that the bias is eventually eliminated. The usual, and in most cases optimal, choice is given by $c_k = c_o k^{-1/6}$ produces a convergence rate for $\mathbf{x} \to \mathbf{x}^*$ with a $k^{-1/3}$ scaling rate.

In FDSA, separate estimates are computed for each component of the gradient, so a p-dimensional problem requires $2p$ evaluations of the loss function per iteration. By contrast, random direction stochastic approximation (RDSA) estimates the gradient with just two loss function evaluations. Let $\boldsymbol{\xi} \in \mathbb{R}^p$ be a direction vector normalized so that $|\boldsymbol{\xi}|^2 = p$. Then

$$\hat{\mathbf{g}}(\mathbf{x}) = \left[\frac{\hat{L}(\mathbf{x} + c\boldsymbol{\xi}) - \hat{L}(\mathbf{x} - c\boldsymbol{\xi})}{2c}\right] \boldsymbol{\xi} \tag{3}$$

and with this approximation, the RDSA iteration becomes

$$\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \left[\frac{\hat{L}(\mathbf{x}_k + c_k\boldsymbol{\xi}_k) - \hat{L}(\mathbf{x}_k - c_k\boldsymbol{\xi}_k)}{2c_k}\right] \boldsymbol{\xi}_k. \tag{4}$$

[1]To simplify exposition, we take L to be infinitely differentiable, but remark that many of the results only require that L be s-times differentiable, where s depends on the particular result.

Several choices are available for choosing the random distributions. The *Axis* distribution takes $\boldsymbol{\xi} = \pm\sqrt{p}\mathbf{e}_i$ with coordinate i chosen at random from $\{1,\ldots,p\}$. The *Normal* distribution takes each component ξ_i to be distributed as $N(0,1)$; it is fully isotropic, but it's radius is not fixed. The average squared radius, however, is the same as for the other distributions: $E\{\,|\boldsymbol{\xi}|^2\,\} = p$. The *Shell* distribution, like the Normal, is fully isotropic, but with fixed radius. This is achieved by taking a vector from the Normal distribution and rescaling so $|\boldsymbol{\xi}|^2 = p$. The *Bernoulli* distribution takes each component ξ_i at random from $\{-1,1\}$.

In RDSA, both the adjustment to $\mathbf{x}_k$ and the estimated gradient are in the same direction $\boldsymbol{\xi}$. By contrast, the general form of SPSA [4] employs two different directions, $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$, and estimates

$$\hat{\mathbf{g}}(\mathbf{x}) = \left[\frac{\hat{L}(\mathbf{x}+c\boldsymbol{\xi}) - \hat{L}(\mathbf{x}-c\boldsymbol{\xi})}{2c}\right]\boldsymbol{\zeta}, \tag{5}$$

where $\boldsymbol{\xi}$ is chosen from a distribution that has to satisfy some particular constraints, and the components of $\boldsymbol{\zeta}$ are given by $\zeta_i = 1/\xi_i$. Sadegh and Spall [5] have shown that under very general conditions, the Bernoulli distribution is the best choice among all the SPSA distributions; in this case, we have $\boldsymbol{\xi} = \boldsymbol{\zeta}$ and that SPSA is equivalent to RDSA with a Bernoulli distribution.

For RDSA, on the other hand, it is not clear which distribution is optimal. Chin [6] provides theoretical arguments and a numerical experiment to support the claim that Bernoulli is optimal, but the theory requires (and the experiment satisfies) a condition that the loss function have third cross-derivatives that are identically zero; that is, $\partial^3 L/\partial x_i \partial x_j \partial x_k = 0$ unless $i = j = k$. Chin [6] also compares RDSA (using a non-Bernoulli distribution) with FDSA and finds that FDSA requires p times as much computation to achieve the same result. This result has been widely reported; for instance, Spall writes [7]:

> Under reasonably general conditions, SPSA and Kiefer-Wolfowitz finite-difference-based SA (FDSA) achieve the same level of statistical accuracy for a given number of iterations even though SPSA uses p times fewer evaluations than FDSA (since each gradient approximation uses only $1/p$ the number of function evaluations).

These "reasonably general" conditions include the requirement that the third cross-derivatives vanish. This requirement applies not only to the loss function, but to the choice of coordinate system in which the loss function is expressed. For loss functions that don't satisfy this requirement – and we will provide specific as well as more generic examples – the optimal distribution is not necessarily Bernoulli.

II. Convergence Rates for Stochastic Approximation Algorithms

To achieve optimal convergence, there is a trade-off between variance and bias. We write the gradient estimator as a sum of three terms: true gradient $\mathbf{g}$, noise $\boldsymbol{\eta}$, and bias $\mathbf{b}$:

$$\hat{\mathbf{g}}(\mathbf{x}, \boldsymbol{\xi}, \epsilon_\pm) = \mathbf{g}(\mathbf{x}) + \boldsymbol{\eta}(\mathbf{x}, \boldsymbol{\xi}, \epsilon_\pm) + \mathbf{b}(\mathbf{x}), \tag{6}$$

where

$$\begin{aligned} \boldsymbol{\eta}(\mathbf{x}, \boldsymbol{\xi}, \epsilon_\pm) &= \hat{\mathbf{g}}(\mathbf{x}, \boldsymbol{\xi}, \epsilon_\pm) - E\{\, \hat{\mathbf{g}}(\mathbf{x}, \boldsymbol{\xi}, \epsilon_\pm) \mid \mathbf{x} \,\}, &(7)\\ \mathbf{b}(\mathbf{x}) &= E\{\, \hat{\mathbf{g}}(\mathbf{x}, \boldsymbol{\xi}, \epsilon_\pm) \mid \mathbf{x} \,\} - \mathbf{g}(\mathbf{x}), &(8) \end{aligned}$$

and $\epsilon_\pm = \hat{L}(\mathbf{x} \pm c\boldsymbol{\xi}) - L(\mathbf{x} \pm c\boldsymbol{\xi})$ is the noise on the estimate of the loss function L; it has mean zero and variance σ^2, and we will henceforth assume that it independent of $\mathbf{x}$ and $\boldsymbol{\xi}$.

In general, the more accurately the gradient $\mathbf{g}(\mathbf{x})$ is estimated, the more rapidly the iteration in Eq. (1) converges to the solution. In the following subsections, we derive variance and bias for different estimators of the gradient, and then show how these relate to the convergence of Eq. (1).

A. Variance

For FDSA, from Eqs. (2,7), we have $\eta_i = (\epsilon_+ - \epsilon_-)/2c$ for the ith component of the noise; since there is a different ϵ_+ and ϵ_- for each component i, the covariance of this noise is $E\left\{\, \boldsymbol{\eta}\boldsymbol{\eta}^T \,\right\} = \frac{\sigma^2}{2c^2} I$, where I is the $p \times p$ identity matrix.

For RDSA, we have two contributions to $\boldsymbol{\eta}$. Like FDSA, the main contribution comes from $\epsilon_\pm$. But for RDSA, there is a second contribution, due to fluctuations induced by randomness in the choice of direction ξ. This second contribution, however, is proportional to the gradient $\mathbf{g} = \nabla L$ which approaches zero in the asymptotic limit. Thus, we can write

$$\boldsymbol{\eta} = \frac{\epsilon_+ - \epsilon_-}{2c} \boldsymbol{\xi}, \tag{9}$$

in the asymptotic limit. The covariance is given by

$$E\left\{\, \boldsymbol{\eta}\boldsymbol{\eta}^T \,\right\} = \frac{\sigma^2}{2c^2} E\left\{\, \boldsymbol{\xi}\boldsymbol{\xi}^T \,\right\} = \frac{\sigma^2}{2c^2} I, \tag{10}$$

which is the same as the FDSA case, and is the same for all RDSA distributions.

B. Bias

Although the variance in the estimator of the gradient does not depend on choice of RDSA distribution, its *bias* does. The bias is defined in Eq. (8), and for RDSA,

$$\begin{aligned}\mathbf{b}(\mathbf{x}) &= E\left\{ \frac{\hat{L}(\mathbf{x}+c\boldsymbol{\xi})-\hat{L}(\mathbf{x}-c\boldsymbol{\xi})}{2c}\,\boldsymbol{\xi} \right\} - \mathbf{g}(\mathbf{x}) \;=\; E\left\{ \frac{L(\mathbf{x}+c\boldsymbol{\xi})-L(\mathbf{x}-c\boldsymbol{\xi})}{2c}\,\boldsymbol{\xi} \right\} - \mathbf{g}(\mathbf{x}) \\ &= E\left\{ L_\xi^{(1)}(\mathbf{x})\,\boldsymbol{\xi} \right\} + \frac{1}{6}\,c^2 E\left\{ L_\xi^{(3)}(\mathbf{x})\,\boldsymbol{\xi} \right\} + O(c^4) - \mathbf{g}(\mathbf{x}), \qquad (11)\end{aligned}$$

where the scalar $L_\xi^{(n)}(\mathbf{x})$ is the nth derivative of L in the direction $\boldsymbol{\xi}$; it is defined by $L_\xi^{(n)}(\mathbf{x}) = \partial^n L(\mathbf{x}+\boldsymbol{\xi}t)/\partial t^n$. In particular, $L_\xi^{(1)}(\mathbf{x}) = \sum_i g_i(\mathbf{x})\xi_i$, and $L_\xi^{(3)}(\mathbf{x}) = \sum_{ijk} L_{ijk}(\mathbf{x})\xi_i\xi_j\xi_k$, where L_{ijk} is third derivative along the axes $\mathbf{e}_i$, $\mathbf{e}_j$ and $\mathbf{e}_k$. In the asymptotic regime, c is small, and we can neglect the $O(c^4)$ terms. Then, the ℓth component of the bias $\mathbf{b}$ is given by

$$\begin{aligned}b_\ell &= E\left\{ (\sum_i g_i\xi_i)\,\xi_\ell \right\} - g_\ell + \frac{1}{6}\,c^2 \sum_{ijk} L_{ijk} E\{\,\xi_i\xi_j\xi_k\xi_\ell\,\} \\ &= \sum_i g_i E\{\,\xi_i\xi_\ell\,\} - g_\ell + \frac{1}{6}\,c^2 \sum_{ijk} L_{ijk} E\{\,\xi_i\xi_j\xi_k\xi_\ell\,\}\,. \qquad (12)\end{aligned}$$

For nearly all random direction schemes that might be considered, and certainly for the schemes considered here, the only nonzero values for $E\{\,\xi_i\xi_\ell\,\}$ and $E\{\,\xi_i\xi_j\xi_k\xi_\ell\,\}$ involve even moments. Thus,

$$b_\ell = g_\ell(E\left\{\,\xi_\ell^2\,\right\} - 1) + \frac{1}{6}\,c^2 \left[L_{\ell\ell\ell} E\left\{\,\xi_\ell^4\,\right\} + 3 \sum_{i=1, i\neq\ell}^{p} L_{ii\ell} E\left\{\,\xi_\ell^2\xi_i^2\,\right\} \right], \qquad (13)$$

where the factor of three comes from the equality of $L_{ii\ell} = L_{i\ell i} = L_{\ell ii}$ for smooth functions. Furthermore, $E\{\,\xi_\ell^2\,\} = 1$ (since $E\{\,|\boldsymbol{\xi}|^2\,\} = E\{\,\sum_i \xi_i^2\,\} = p$), so the first order gradient term vanishes. We will define

$$\tau \equiv E\left\{\,\xi_\ell^4\,\right\}, \qquad (14)$$

$$\nu \equiv E\left\{\,\xi_\ell^2\xi_i^2\,\right\}, \qquad (15)$$

so then the bias can be expressed as

$$b_\ell = \frac{1}{6}\,c^2 \left[\tau L_{\ell\ell\ell} + 3\nu \sum_{i=1, i\neq\ell}^{p} L_{ii\ell} \right]. \qquad (16)$$

Note that Eq. (2.4) in Chin [6] gives the incorrect form $b_\ell = \frac{1}{6}\,c^2\tau \left[L_{\ell\ell\ell} + 3\sum_{i=1,i\neq\ell}^{p} L_{ii\ell} \right]$.

Table I shows the values of τ and ν for different random direction distributions. The computation is straightforward for Bernoulli, Axis, and Normal; the derivation for Shell is given in the Appendix.

TABLE I

STATISTICAL MOMENTS OF RANDOM DIRECTION DISTRIBUTIONS

	$\tau = E\{\, \xi_\ell^4 \,\}$	$\nu = E\{\, \xi_\ell^2 \xi_m^2 \,\}$	$p\tau + p(p-1)\nu$	$[\tau - 3\nu]^2$
Distribution				
Bernoulli	1	1	p^2	4
Axis	p	0	p^2	p^2
Normal	3	1	$p(p+2)$	0
Shell	$3p/(p+2)$	$p/(p+2)$	p^2	0

For FDSA, the bias is given by

$$b_\ell = \frac{1}{6}\, c^2 L_{\ell\ell\ell}. \tag{17}$$

We remark that these gradient estimators exhibit a bias that scales quadratically with c, as seen in Eq. (16) and Eq. (17). We define a scaled bias that factors out this c dependence:

$$\tilde{\mathbf{b}} = \lim_{k\to\infty} \mathbf{b}(\mathbf{x}_k) c_k^{-2} = \lim_{c\to 0} \mathbf{b}(\mathbf{x}^*) c^{-2}. \tag{18}$$

C. Convergence of stochastic approximation

From the variance and bias of the estimators $\hat{\mathbf{g}}(\mathbf{x})$ of the gradient, we compute convergence rates for different RDSA distributions. For simplicity, we assume that the data are scaled so that the Hessian is $H(\mathbf{x}^*) = 2I$, which is equivalent to assuming that the loss function is $L(\mathbf{x}) = |\mathbf{x} - \mathbf{x}^*|^2 +$ higher order terms. With this Hessian, and for the optimal convergence schedules, given by $a_k = a_o k^{-1}$, and $c_k = c_o k^{-1/6}$, we employ Fabian's result [8] to obtain[2]

$$\lim_{k\to\infty} E\Big\{\, k^{2/3} |\mathbf{x}_k - \mathbf{x}^*|^2 \,\Big\} = \frac{a_o^2 c_o^{-2} p \sigma^2}{4(2a_o - 1/3)} + \frac{a_o^2 c_o^4 |\tilde{\mathbf{b}}|^2}{(2a_o - 1/3)^2}. \tag{19}$$

We can choose a_o and c_o to optimize this expression by setting derivatives with respect to a_o and c_o to zero. We get $a_o = 1/2$ and $c_o = \left(p\sigma^2/12|\tilde{\mathbf{b}}|^2\right)^{1/6}$. Putting this into Eq. (19) leads to[3]

$$|\mathbf{x}_k - \mathbf{x}^*|^2 \sim \text{const}\, |\tilde{\mathbf{b}}|^{2/3} p^{2/3} \sigma^{4/3}\, k^{-2/3}. \tag{20}$$

[2]What Fabian [8] shows is that $k^{\beta/2}(\mathbf{x}_k - \mathbf{x})$, with $\beta = 2/3$ in this case, approaches a gaussian distribution with a certain mean $\boldsymbol{\mu}$ and variance V; the right hand side of Eq. (19) is the scalar value $\mathrm{tr}(V) + \boldsymbol{\mu}^T\boldsymbol{\mu}$. We follow Spall [4, p. 337] and assume that the first and second moments of the asymptotic normal distribution correspond to the first and second asymptotic moments of $k^{\beta/2}(\mathbf{x}_k - \mathbf{x})$; see Ref. [9] for discussion of conditions under which convergence can be guaranteed.

[3]We write $u_k \sim k^{-\beta}$ as an informal shorthand for the statement that $\lim_{k\to\infty} E\big\{\, k^\beta u_k \,\big\}$ is $O(1)$.

The value of "const" in the above expression is $3^{7/3}2^{-16/3} \approx 0.322$, but we will treat it as $O(1)$ for simplicity. For RDSA, there are $N = 2k$ function evaluations in k iterations, so the mean squared error scales as

$$\text{MSE}_{\text{RDSA}} \sim \left(|\tilde{\mathbf{b}}_{\text{RDSA}}|\, p \right)^{2/3} \sigma^{4/3} N^{-2/3}, \tag{21}$$

while for FDSA, there are $N = 2pk$ evaluations by the kth iteration, so

$$\text{MSE}_{\text{FDSA}} \sim \left(|\tilde{\mathbf{b}}_{\text{FDSA}}|\, p^2 \right)^{2/3} \sigma^{4/3} N^{-2/3}. \tag{22}$$

This shows how the performance of RDSA and FDSA scale with the bias $|\tilde{\mathbf{b}}|$ and dimension p.

III. Generically rotated loss functions

We introduce the notion of a "generically rotated loss function" to investigate how the performance of RDSA algorithms depend on the alignment of a loss function with its coordinate axes. Fig. 1 provides a visualization of this concept. Here, a forty-five degree rotation of the coordinate axes would not change the fundamental nature of the loss function, but it would effectively swap the Axis directions with the Bernoulli directions. If the Bernoulli directions were best in the original coordinate system, then the Axis directions would be optimal in the rotated axes. We ask: if, as in Fig. 1(c), the loss function has no particular alignment with respect to its coordinate system, then which stochastic approximation scheme should be used?

A. *Expected bias for generically rotated loss functions*

We produce loss functions that have "generic" orientations with respect to their coordinate axes by choosing a rotation matrix U and applying it to each random direction $\boldsymbol{\xi}$, so that $\boldsymbol{\xi}' = U\boldsymbol{\xi}$ becomes the random direction in the rotated axes. Thus, if $P(\boldsymbol{\xi})$ is the distribution (*e.g.,* Axis or Bernoulli) in the initially aligned axes, then we consider a new distribution $P_U(\boldsymbol{\xi}) = |U|P(U\boldsymbol{\xi})$ which is the distribution $P(\boldsymbol{\xi})$ rotated by U. Here $|U| = 1$ is the determinant of U.

We compute τ_U, the fourth moment of ξ'_ℓ where $\boldsymbol{\xi}'$ is taken from the distribution P_U.

$$\begin{aligned}
\tau_U &= E\Big\{ (\xi'_\ell)^4 \Big\} = E\Big\{ (U\xi)^4_\ell \Big\} \\
&= E\left\{ \sum_{ijkm} u_{i\ell}u_{j\ell}u_{k\ell}u_{m\ell}\xi_i\xi_j\xi_k\xi_m \right\} = \sum_{ijkm} u_{i\ell}u_{j\ell}u_{k\ell}u_{m\ell} E\{\, \xi_i\xi_j\xi_k\xi_m \,\}.
\end{aligned} \tag{23}$$

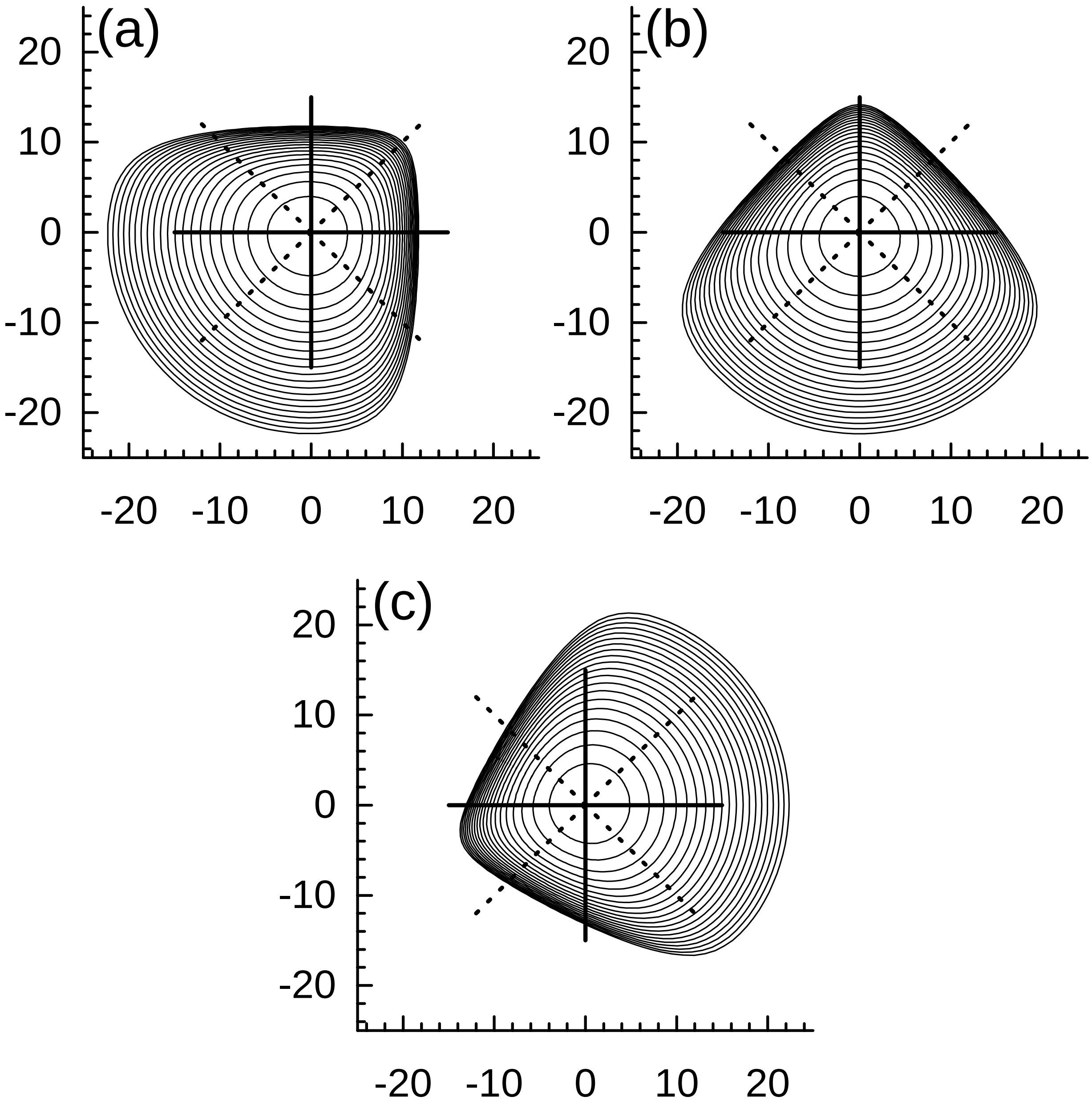

Fig. 1. **(a)** Contours of the loss function in Eq. (38) for $p = 2$ dimensions. Contours are shown for every $\Delta L = 25$ up to $L = 500$. The horizontal and vertical solid lines correspond to the Axis directions; the diagonal dotted lines correspond to the Bernoulli directions. In this orientation, the Axis directions produce the largest bias $\mathbf{b}$, and the Bernoulli directions produce the least bias. **(b)** The same loss function, but rotated forty-five degrees. The Axis directions in this coordinate system slice through the loss function in just the same way that the Bernoulli directions sliced through the original loss function. **(c)** Again, the same loss function, this time rotated by a "random" angle of 151^o.

The average over all random directions of $E\{\,\xi_i\xi_j\xi_k\xi_m\,\}$ is given by

$$\begin{aligned} E\{\,\xi_i\xi_j\xi_k\xi_m\,\} &= \tau\delta(i=j=k=m)+\nu\delta((i=j)\neq(k=m)) \\ &+ \nu\delta((i=k)\neq(m=j))+\nu\delta((i=m)\neq(j=k)), \end{aligned} \tag{24}$$

and therefore

$$\tau_U=\tau\sum_i u_{i\ell}^4+3\nu\sum_{i\neq j}u_{i\ell}^2u_{j\ell}^2. \tag{25}$$

The average performance over an ensemble of rotations depends on the "average" τ_U given by[4]

$$\langle\tau_U\rangle=\tau\left\langle\sum_i u_{i\ell}^4\right\rangle+3\nu\left\langle\sum_{i\neq j}u_{i\ell}^2u_{j\ell}^2\right\rangle=\tau p\left\langle u_{i\ell}^4\right\rangle+3\nu p(p-1)\left\langle u_{i\ell}^2u_{j\ell}^2\right\rangle. \tag{26}$$

Since the matrix elements $u_{\bullet\ell}$ form orthogonal vectors of unit length, we can compute moments as described in the Appendix by Eq. (49). In particular, we have

$$\langle\tau_U\rangle=\tau p\,\frac{3}{p(p+2)}+3\nu p(p-1)\,\frac{1}{p(p+2)}=\frac{3}{p(p+2)}\left[p\tau+p(p-1)\nu\right]. \tag{27}$$

A similar argument shows

$$\langle\nu_U\rangle=\frac{1}{p(p+2)}\left[p\tau+p(p-1)\nu\right]. \tag{28}$$

But it is important to note that

$$p\tau+p(p-1)\nu=pE\left\{\,\xi_i^4\,\right\}+p(p-1)E\left\{\,\xi_i^2\xi_j^2\,\right\}=E\left\{\,(\sum_i\xi_i^2)^2\,\right\}=E\left\{\,r^4\,\right\}, \tag{29}$$

where $E\{\,r^4\,\}$ is the *radial* fourth moment of the direction $\boldsymbol{\xi}$. Since the radius is normalized so that $r=\sqrt{p}$, we have $E\{\,r^4\,\}=p^2$ for the Bernoulli, Axis, and Shell distributions; for the Normal distribution, the radius is only normalized on average, and we have $E\{\,r^4\,\}=p(p+2)$.

As $\langle\tau_U\rangle$ and $\langle\nu_U\rangle$ depend only on the radial fourth moment, they are the same for Bernoulli, Axis, and Shell. So for generic orientations of the loss functions with respect to the coordinate axes, all three will on average give the same result. In particular, the rotation-averaged bias is

$$\langle b_\ell\rangle=\frac{1}{6}c^2\left[\frac{p\tau+p(p-1)\nu}{p(p+2)}\right]\left(3L_{\ell\ell\ell}+3\sum_{i=1,i\neq\ell}^{p}L_{ii\ell}\right)=\frac{1}{2}c^2\left[\frac{E\{\,r^4\,\}}{p(p+2)}\right]\sum_{i=1}^{p}L_{ii\ell}. \tag{30}$$

[4]Previously, we have used the notation $E\{\,\ldots\,\}$ to indicate expectation over loss function evaluation and random directions. Here, we are using $\langle\ldots\rangle$ to represent ensemble average over random rotation matrices U.

For the FDSA algorithm, the rotation-averaged bias can be shown to be

$$\langle b_\ell \rangle = \frac{1}{2}c^2 \left[\frac{1}{p+2}\right] \sum_{i=1}^{p} L_{ii\ell}, \tag{31}$$

which is a factor of p smaller than the RDSA bias in Eq. (30). That is, $|\tilde{\mathbf{b}}_{\text{FDSA}}| = |\tilde{\mathbf{b}}_{\text{RDSA}}|/p$, on average. So from Eq. (21) and Eq. (22), we expect the asymptotic efficiency of FDSA and RDSA to be the same for generic loss functions which do not exhibit any special alignment to their coordinate axes.

B. Squared bias for generically rotated loss functions

In this section, we compute the mean *squared* bias, averaged over all orientations. First,

$$|\mathbf{b}|^2 = \sum_{\ell=1}^{p} b_\ell^2 = \frac{1}{36}c^4 \sum_{\ell=1}^{p} \left(\tau L_{\ell\ell\ell} + 3\nu \sum_{m\neq\ell} L_{\ell mm}\right)^2 \tag{32}$$

$$= \frac{1}{36}c^4 \Big[\tau^2 p L_{\ell\ell\ell}^2 + 6\tau\nu p(p-1) L_{\ell\ell\ell} L_{\ell mm} + 9\nu^2 p(p-1) L_{\ell mm}^2 + 9\nu^2 p(p-1)(p-2) L_{\ell mm} L_{\ell nn}\Big]. \tag{33}$$

To compute the average value of $|\mathbf{b}|^2$, we need to consider the effect of rotation on the third derivatives. In general,

$$L_{\ell mn} = \sum_{ijk} L'_{ijk} u_{i\ell} u_{jm} u_{kn}, \tag{34}$$

describes the rotated third derivative $L_{\ell mn}$ in terms of the original third derivatives L'_{ijk}. Here, u_{ij} is the ij matrix element of the rotation matrix U.

Forgoing full generality, we consider loss functions whose anisotropy is characterized (like that of the Chin function, described in the next section) by the existence of a coordinate system in which the loss is a sum of separate but identical losses on each of the coordinates; that is, $L(\mathbf{z}) = \sum_i \mathcal{L}(\sum_j u_{ij} z_j)$. It follows that the third derivative is given by

$$L_{\ell mn} = \mathcal{L}''' \sum_i u_{i\ell} u_{im} u_{in}. \tag{35}$$

To compute the expectation value for the various products that appear in Eq. (33), we substitute from Eq. (35). For example,

$$\left\langle L_{\ell\ell\ell}^2 \right\rangle / (\mathcal{L}''')^2 = \sum_{jk} \left\langle u_{j\ell}^3 u_{k\ell}^3 \right\rangle = \sum_j \left\langle u_{j\ell}^6 \right\rangle = p \left\langle u_{j\ell}^6 \right\rangle = \frac{15}{(p+2)(p+4)}, \tag{36}$$

where the last equality follows from Eq. (49) in the Appendix. We can write similar expressions for $\langle L_{\ell mm}^2 \rangle$, $\langle L_{\ell\ell\ell} L_{\ell mm} \rangle$, and $\langle L_{\ell mm} L_{\ell nn} \rangle$, and substituting these into Eq. (33), and performing a little algebraic manipulation, we obtain

$$\left\langle |\mathbf{b}|^2 \right\rangle = p \left(\frac{\mathcal{L}''' c^2}{2p(p+2)} [p\tau + p(p-1)\nu] \right)^2 + \frac{p(p-1)(\mathcal{L}''')^2 c^4}{6(p+2)^2(p+4)} [\tau - 3\nu]^2. \quad (37)$$

Here, the first term is the square of the average bias, computed in Section III-A, and it is the same for the Shell, Axis, and Bernoulli distributions. The second term is a "variance of the bias" and is proportional to an anisotropy factor $[\tau - 3\nu]^2$ which, as Table I shows, vanishes for spherically symmetric distance distributions such as Shell and Normal, but is positive for the anisotropic Axis and Bernoulli distributions. Thus, the average squared bias, which is the sum of these two terms, is smallest for the Shell distribution.

IV. Numerical Studies

A. *Comparison of RDSA variants for two specific loss functions*

In Chin [6], various finite difference stochastic algorithms are compared using a loss function

$$L(\mathbf{x}) = |\mathbf{x}|^2 + \sum_{i=1}^{p} e^{x_i/p}, \quad (38)$$

which can also be written $L(\mathbf{x}) = \sum_{i=1}^p \mathcal{L}(x_i)$, with $\mathcal{L}(x) = x^2 + e^{x/p}$. One particular aspect of this symmetry is that the cross-derivatives are zero, and so Eq. (16) simplifies in this case to $b_\ell = (1/6)c^2 \tau L_{\ell\ell\ell} = (1/6)c^2 \mathcal{L}'''$ and the bias is simply proportional to the fourth moment τ defined in Eq. (14). And as Table I shows, the Bernoulli distribution has the smallest τ.

In our numerical experiments with this loss function, we used parameters from Chin [6]: $p = 15$, $\alpha = 1$, $a_o = 0.5$, and $\gamma = 1/6$. (Chin [6] reported using $\beta = 1/6$, but we presume that was a typo, as it would lead to nonoptimal $\gamma = 5/12$.) Each loss function evaluation added gaussian noise with $\sigma = 1.9365$, corresponding to a variance in the difference of two loss function estimates of $2\sigma^2 = 7.5$. We used $\mathbf{x}_o = [-0.01, \ldots, -0.01]$ for initial conditions, and express squared error relative to this initial value: MSE $= |\mathbf{x} - \mathbf{x}^*|^2 / |\mathbf{x}_o - \mathbf{x}^*|^2$, where $\mathbf{x}^* \approx [-0.03326, \ldots, -0.03326]$ is the minimum of the loss function.

Fig. 2(a) shows that scaling of squared error as $O(N^{-2/3})$ is observed for all the schemes, but the coefficient is best for Bernoulli, not as good for Shell and Normal, and worst for the Axis distribution and for FDSA. And in the asymptotic regime, Bernoulli achieves the same accuracy

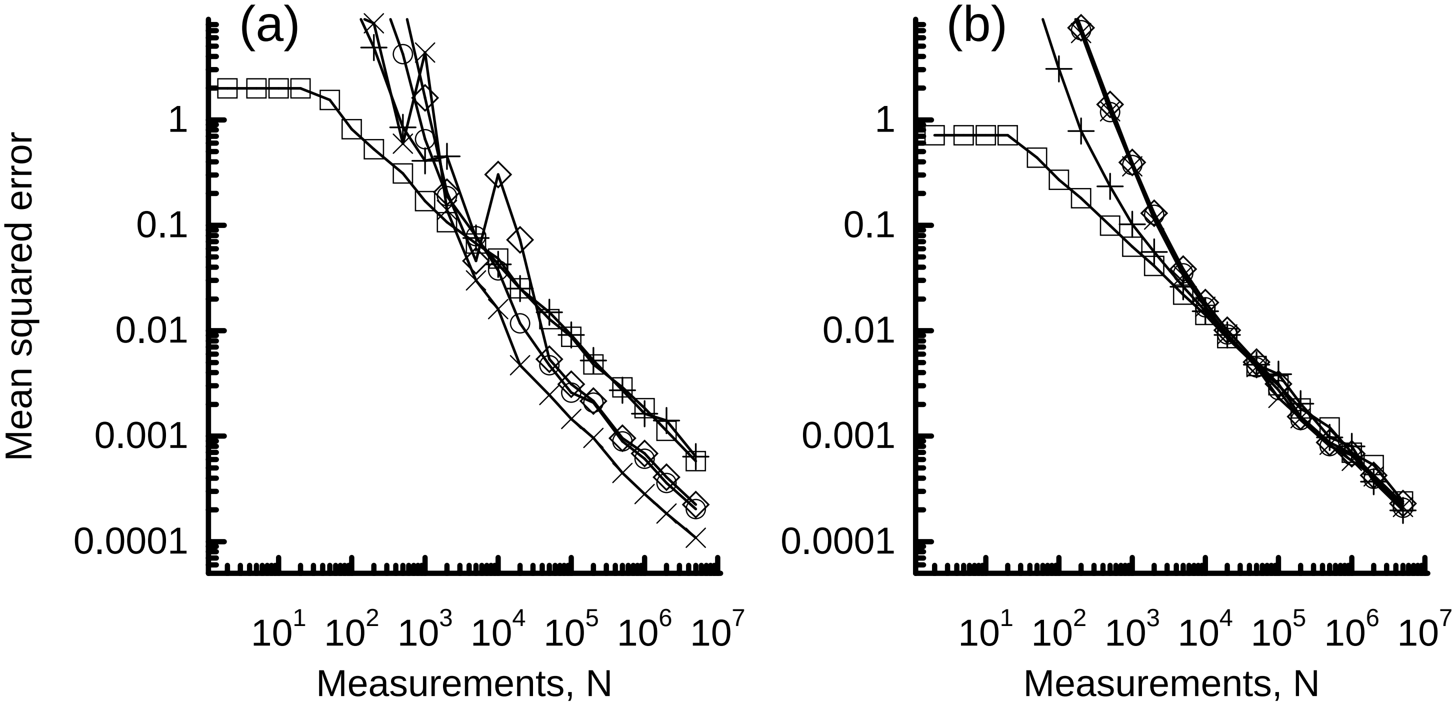

Fig. 2. Mean squared error (MSE), computed as an average over 10 trials, is plotted as a function of the number of measurements N for different choices of random direction distribution Here, plus ($+$) is the Axis distribution, cross ($\times$) is the Bernoulli distribution, circle ($\circ$) is the uniform Shell distribution, and diamond ($\diamond$) is the Normal distribution. We use squares ($\square$) to indicate the performance of the FDSA algorithm. **(a)** Using the Chin loss function in Eq. (38), the best (smallest) asymptotic MSE is given by the Bernoulli distribution ($\times$), the Shell and Normal distributions have larger MSE, and the Axis distribution and FDSA are tied with the largest asymptotic MSE. **(b)** When the Chin function is rotated so that it has no particular alignment with respect to its coordinate axes, the asymptotic performance of the different algorithms is essentially the same.

as FDSA with $p = 15$ times fewer measurements. This confirms the theoretical predictions in Chin [6], and agrees with the scaling in Eq. (21) and Eq. (22). But we also observe that for $N \ll 1000$ measurements, the FDSA algorithm is substantially more accurate.

We also investigated a loss function

$$L(\mathbf{x}) = |\mathbf{x}|^2 + K \left[\sum_i x_i \sum_i x_i^2 - \sum_i x_i^3 \right] / p^2. \tag{39}$$

for which the direct third derivatives $L_{\ell\ell\ell}$ are zero, but the cross-derivatives are nonzero. Where the bias for Chin's function depends on τ only, the bias for this function depends on ν only. Since $\nu = 0$ for the Axis distribution, we expect – and in numerical experiments (not shown), we have observed – substantially better performance using the Axis distribution instead of the Bernoulli distribution.

It is not clear that the loss function in Eq. (39) is any more or less "typical" than the loss function in Eq. (38). But Eq. (39) provides a specific example of a loss function for which Bernoulli is *not* optimal. Since the loss function in Eq. (39) has nonzero third cross-derivatives, it violates the "reasonably general conditions" referred to in Ref. [7], and so the theoretical arguments for the superiority of Bernoulli do not apply to this loss function. The question is: do those arguments apply to typical loss functions?

B. Generically rotated loss function

To investigate the performance of different RDSA distributions on a more generic class of loss function, we altered Chin's function so that it would not be so neatly aligned with the axes. We chose a random rotation matrix U and defined $\tilde{L}(\mathbf{x}) = L(U\mathbf{x}) = |U\mathbf{x}|^2 + \sum_i e^{(U\mathbf{x})_i/p}$, or

$$\tilde{L}(\mathbf{x}) = \sum_{i=1}^{p} \left[x_i^2 + \exp\left(\sum_j U_{ij} x_j / p \right) \right], \tag{40}$$

which uses $|\mathbf{x}|^2 = |U\mathbf{x}|^2$ since U is orthogonal. For this loss function, we did the same experiment as for the unrotated Chin loss function, and in Fig. 2(b), we see that the asymptotic performance for each of the different random direction distributions was virtually the same. The performance of FDSA was also comparable to the RDSA algorithms, and in the pre-asymptotic regime, FDSA was better. Eq. (40) provides another example of a loss function which violates the conditions required for fast convergence of RDSA-Bernoulli; but unlike Eq. (39), this is a more "generic" example, obtained simply by using a different set of coordinates for a system that in its original coordinates, did satisfy these conditions.

V. CONCLUSIONS

The performance of different variants of the RDSA algorithm is driven by the accuracy with which they estimate the gradient of the loss function. In Eq. (16), that bias is expressed in terms of two statistics – τ and ν, defined in Eq. (14) and Eq. (15), respectively – which characterize the random direction distribution, and in terms of the third derivatives of the loss function.

For loss functions whose third cross-derivatives are zero (that is, $\partial^3 L/\partial x_i \partial x_j \partial x_k = 0$ unless $i = j = k$), the bias depends only on τ and is smallest for the Bernoulli distribution. For such loss functions, RDSA with a Bernoulli distribution is p times more efficient than the Kiefer-Wolfowitz-Blum FDSA algorithm. We remark that RDSA-Bernoulli is also p times more efficient than RDSA-Axis; the factor of p has more to do with bias in the gradient estimators in Eqs. (16,17) than the fact that FDSA takes p times as many function evaluations per iteration.

However, this behavior is not generic. We have shown how to construct loss functions, such as Eq. (39), for which Bernoulli is *less* efficient than other distributions. For loss functions that are randomly rotated with respect to their coordinate axes, we find that the average bias depends only on the fourth radial moment of the distribution, and is the same for Axis, Shell, and Bernoulli. Furthermore, we find for these generically rotated loss functions that FDSA performs as well as the various RDSA algorithms in the asymptotic limit. And in our numerical experiments, FDSA was actually better than RDSA in the pre-asymptotic regime. Finally, by looking at the squared bias, averaged over random rotations, we argue that the optimal choice of distributions in RDSA is given by a uniform spherical shell.

APPENDIX

In this appendix, we derive an expression for the moments of the isotropic distribution of $(x_1, \ldots, x_p)$ constrained by $x_1^2 + \ldots + x_p^2 = 1$. Define

$$F_{p,\mathbf{m}} \equiv E\Big\{ x_1^{m_1} x_2^{m_2} \cdots x_p^{m_p} \Big\} = \frac{\int_{\mathcal{S}} x_1^{m_1} x_2^{m_2} \cdots x_p^{m_p} \, dS}{\int_{\mathcal{S}} dS}, \tag{41}$$

where $\int_{\mathcal{S}} dS$ corresponds to the integral over the surface of the unit sphere.

Let $\mathbf{z}$ be a p-dimensional vector in which each component is independently distributed as $\mathcal{N}(0,1)$. We will compute the quantity $E\Big\{ z_1^{m_1} \cdots z_p^{m_p} \Big\}$ in two different ways, and by equating the results, we will obtain an expression for $F_{p,\mathbf{m}}$.

For our first computation, we use the fact that the gaussian components are independent, and write $E\Big\{ z_1^{m_1} \cdots z_p^{m_p} \Big\} = E\{ z_1^{m_1} \} \cdots E\Big\{ z_p^{m_p} \Big\}$. We have for each component that

$$E\{ z^m \} = \frac{\int_{-\infty}^{\infty} z^m e^{-z^2/2} \, dz}{\int_{-\infty}^{\infty} e^{-z^2/2} \, dz} = (m-1)!! \tag{42}$$

if m is even (and zero if m is odd). Note that '!!' corresponds to the double factorial: $n!! = n(n-2)(n-4)\cdots 1$. So as long as $m_1, \ldots, m_p$ are even, we have

$$E\Big\{ z_1^{m_1} \cdots z_p^{m_p} \Big\} = (m_1 - 1)!! \cdots (m_p - 1)!! \tag{43}$$

For our second derivation, we use the relation $\mathbf{z} = r\mathbf{x}$, where r is the radius of the vector $\mathbf{z}$, and $\mathbf{x}$ is the unit vector in the direction of $\mathbf{z}$. Thus we have

$$z_1^{m_1} \cdots z_p^{m_p} = r^{m_1 + \cdots + m_p} x_1^{m_1} \cdots x_p^{m_p}. \tag{44}$$

Rather than integrate coordinate-wise, this time we will use polar coordinates:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} dz_1 \, \ldots \, dz_p = \int_0^{\infty} \int_{\mathcal{S}} r^{p-1} \, dr \, dS. \tag{45}$$

Then,

$$\begin{aligned} E\Big\{ z_1^{m_1} \cdots z_p^{m_p} \Big\} &= \frac{\int_0^{\infty} \int_{\mathcal{S}} z_1^{m_1} \cdots z_p^{m_p} \, e^{-r^2/2} \, r^{p-1} \, dr \, dS}{\int_0^{\infty} \int_{\mathcal{S}} e^{-r^2/2} \, r^{p-1} \, dr \, dS} && (46) \\ &= \frac{\left(\int_0^{\infty} e^{-r^2/2} \, r^{m_1+\cdots+m_p} \, r^{p-1} \, dr\right) \left(\int_{\mathcal{S}} x_1^{m_1} \cdots x_p^{m_p} \, dS\right)}{\int_0^{\infty} e^{-r^2/2} \, r^{p-1} \, dr \, \int_{\mathcal{S}} dS} && (47) \\ &= \frac{(p + m_1 + \cdots + m_p - 2)!! \, F_{p,\mathbf{m}}}{(p-2)!!}. && (48) \end{aligned}$$

Equating Eq. (43) and Eq. (48), and solving for $F_{p,\mathbf{m}}$, allows us to write

$$E\Big\{ x_1^{m_1} x_2^{m_2} \cdots x_p^{m_p} \Big\} = \frac{(m_1 - 1)!! \cdots (m_p - 1)!!}{p(p+2) \cdots (p + m_1 + m_2 + \cdots + m_p - 2)}. \tag{49}$$

Two special cases of Eq. (49) provide τ and ν for the spherical shell distribution; these are

$$\begin{aligned} \tau = E\Big\{ x_i^4 \Big\} &= 3/[p(p+2)], && (50) \\ \nu = E\Big\{ x_i^2 x_j^2 \Big\} &= 1/[p(p+2)]. && (51) \end{aligned}$$

REFERENCES

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp. 400–407, 1951.

[2] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Stat.*, vol. 23, pp. 462–466, 1952.

[3] J. R. Blum, "Multidimensional stochastic approximation methods," *Ann. Math. Stat.*, vol. 25, pp. 737–744, 1954.

[4] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Automatic Control*, vol. 37, pp. 332–341, 1992.

[5] P. Sadegh and J. C. Spall, "Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation," *Proc. American Control Conference, 4-6 June 1997, Albuquerque, NM*, pp. 3582–3586, 1997.

[6] D. C. Chin, "Comparative study of stochastic algorithms for system optimization based on gradient approximation," *IEEE Trans. Systems Man and Cybernetics – Part B*, vol. 27, pp. 244–249, 1997.

[7] J. C. Spall, "Implementation of the simultaneous perturbation algorithm for stochastic optimization," *IEEE Trans. Aerospace and Electronic Systems*, vol. 34, pp. 817–823, 1998.

[8] V. Fabian, "On asymptotic normality in stochastic approximation," *Ann. Math. Stat.*, vol. 39, pp. 1327–1332, 1968.

[9] L. Gerencsér, "Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting," *IEEE Trans. Automatic Control*, vol. 44, pp. 894–905, 1999.