EN.553.763 Stochastic Search and Optimization

Final Report

Ali Geisa

May 16, 2018

Contents

1 Introduction 1

2 Theory 2

3 Analysis 8

4 Numerical Results 9

5 Conclusion and Future Work 13

6 References 14

1 Introduction

Simultaneous Perturbation Stochastic Approximation (SPSA) is a very powerful optimization algorithm [1]. It allows for gradient-free optimization, as it only utilizes loss function evaluations. One key aspect of the SPSA algorithm is the simultaneous perturbation. This is a random perturbation of the loss function from our current best estimate for the parameter vector $\hat{\theta}_k$. The key word here is random. What kind of random? There is an infinite number of probability distributions. Not all of these probability distributions, however, qualify to be used to generate the random perturbations. They have to satisfy certain criteria. Now, not all distributions are created equal. Some distributions are better fit for the task than others. The state of the art status quo so far has been a $-1, 1$ Bernoulli distribution, with $p = 0.5$. Can we do better than that? This report investigates the performance of custom probability distributions of the form

$$\frac{n+1}{2} \cdot x^n$$

n is a positive even integer, and the $\frac{n+1}{2}$ is a normalization term so that we have a proper distribution. Note, these distributions will be termed even monomial distributions from this point going forward. If you want to find out the answer now, you can skip straight to the conclusion.

Why is this important? SPSA is very important in optimization where only loss function values are available and we would like to estimate a gradient. One important application, for example, is model-free control, where SPSA is used to optimize a controller.

2 Theory

The fundamental problem here is that of optimization. We have a real valued function of a real vector, which we simply call a loss function, and denote $L(\cdot)$, which we would like to optimize. Optimize here will be in the sense of minimizing the function, i.e. find $\arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. There are many, many optimization algorithms, all better suited to different types of optimization problems. One subset of optimization problems are those where there is a stochastic element in the optimization process. This stochastic component can take the form of noisy measurements of the loss functions, so that in reality we are measuring $\boldsymbol{y}(\boldsymbol{\theta}) = \boldsymbol{L}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a random variable. The stochastic component can also be intentionally introduced into the problem as part of the optimization algorithm (such as with random search algorithms).

The general paradigm for optimizing these problems in which there is a stochastic component in the measurement process is the Stochastic Approximation (SA) paradigm. In general, the algorithm takes the following form

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k - a_k \mathbf{Y}(\hat{\boldsymbol{\theta}}_k)$$

Where the $\boldsymbol{k}$ subscript indicates which iterate in the iteration (SA is an iterative optimization algorithm), the $\boldsymbol{a_k}$ is a step size that must satisfy certain properties, and $\mathbf{Y}$ is the noisy gradient of the loss function. More concretly, the noisy gradient is $\mathbf{Y}(\hat{\boldsymbol{\theta}}_k) = \mathbf{g}(\hat{\boldsymbol{\theta}}_k) + \varepsilon_k$. $\mathbf{g}(\hat{\boldsymbol{\theta}}_k)$ is the true gradient and $\boldsymbol{\varepsilon_k}$ is an independent noise term.

The subset of optimization problems on which we will focus our efforts are those in which the only available information is in the form of noisy loss function measurements $\boldsymbol{y}(\boldsymbol{\theta})$. We cannot directly measure the gradient for some reason in these cases (simply do not have access, very expensive to measure, etc...). In those cases, when we still wish to optimize with some form of SA, we have to estimate the gradient. The algorithm will then take the following modified form

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k - a_k \hat{\mathbf{g}}(\hat{\boldsymbol{\theta}}_k)$$

Where $\hat{\mathbf{g}}$ is the gradient estimate.

A powerful algorithm to perform the estimation and optimization processes is SPSA. The gradient in SPSA is estimated by the following

$$\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k) = \frac{y(\boldsymbol{\theta}_k + c_k \boldsymbol{\Delta}_k) - y(\boldsymbol{\theta}_k - c_k \boldsymbol{\Delta}_k)}{2 c_k \Delta_{km}}$$

Where m means the m^{th} component of the gradient estimate. $\mathbf{\Delta}_k$ is a random vector and c_k is a step size. There are conditions that need to be satisfied and more to the algorithm that can be checked out in [1]. More pertinently, though, is the choice of $\mathbf{\Delta}_k$. What kind of randomness is it? As mentioned in the introduction, an asymptotically optimal randomness is that it is simply a vector of Bernoulli random variables, taking on the values of 1 or -1 with probability 0.5. Here asymptotically optimal means that it leads to optimal algorithm performance as the number of iterations gets very large [1].

The above result, however, is an asymptotic result. Can we do better in the finite number of iterates case? Specifically, if we generated our random vector from a distribution of the form $\frac{n+1}{2} \cdot x^n$, can we somehow show better performance if we only get 100 loss function evaluations?

Let us first show that $\frac{n+1}{2} \cdot x^n$ is a valid distribution to use for SPSA. One key condition that any SPSA distribution needs to satisfy is the possession of finite second absolute moment. From [1], this translates into requiring the following

$$\int_{-\infty}^{\infty} |\frac{1}{x}|^{2+2\tau} f(x) dx < C$$

Where $\tau > 0$, $f(x)$ is the density of the random variable X, and C is a finite positive number. Let us check this condition

Theorem 1 Distributions of the form $X \sim \frac{n+1}{2} \cdot x^n$ have finite second inverse absolute moment, which by the conditions in [1], means that $\int_{-\infty}^{\infty} |\frac{1}{x}|^{2+2\tau} f(x) dx < C$ where $\tau > 0$ and $C < \infty$.

$$X \sim \frac{n+1}{2} x^n \qquad n \in \mathbb{N} \qquad n \geq 2 \qquad 2|n \qquad x \in [-1, 1]$$
$$\int_{-\infty}^{\infty} |\frac{1}{x}|^{2+2\tau} f(x) dx =$$

$$\frac{n+1}{2}\int_{-1}^{1}|\frac{1}{x}|^{2+2\tau}x^n dx =$$

$$\frac{n+1}{2}\Big(\int_{-1}^{0}(-\frac{1}{x})^{2+2\tau}x^n dx + \int_{0}^{1}\frac{1}{x^{2+2\tau}}x^n dx\Big)$$

$$\int_{-1}^{0}(-\frac{1}{x})^{2+2\tau}x^n dx = (-1)^{2+2\tau}\int_{-1}^{0}\frac{1}{x^{2+2\tau}}x^n dx =$$

$$(-1)^{2+2\tau}\int_{-1}^{0}x^{n-2-2\tau}dx = (-1)^{2+2\tau}\frac{x^{n-1-2\tau}}{n-1-2\tau}\Big|_{-1}^{0} = \frac{1}{n-1-2\tau}$$

$$\int_{0}^{1}x^{n-2-2\tau}dx = \frac{1}{n-1-2\tau}$$

$$\int_{-\infty}^{\infty}|\frac{1}{x}|^{2+2\tau}f(x)dx = \frac{2}{n-1-2\tau}$$

Thus for $n \in \{2,4,6,\cdots\} = 2\mathbb{N}$, we have a U-shaped distribution with finite second absolute inverse moment that fulfills the criteria in [1].

Now for the main question, can these sorts of distributions perform better than the Bernoulli for generating perturbations? Let us examine the mean squared error (MSE). Let the optimal parameter be denoted $\boldsymbol{\theta}^*$. Furthermore, condition on all the previous iterates, so that we have the following

$$\mathbb{E}\big[||\hat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*||^2|\hat{\boldsymbol{\theta}}_k\big]$$

For notational convenience, we will suppress the conditioning on $\hat{\boldsymbol{\theta}}_k$, but all the expectations will be conditioned on it.

Theorem 2 The mean squared error of $\hat{\boldsymbol{\theta}}_{k+1}$, conditioning on all the previous estimates, $\mathbb{E}\big[||\hat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*||^2\big]$, is approximately the following expression

$$\sum_{m=1}^{p}\Bigg(\hat{\theta}_{km}^2 - 2a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\hat{\theta}_{km} + a_k^2(L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + \sum_{i\neq m}^{p}\big(L_i^{'}(\hat{\boldsymbol{\theta}}_k)\big)^2\mathbb{E}\Big[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\Big] +$$
$$\frac{\sigma^2}{2c_k^2}\cdot\mathbb{E}\Big[\frac{1}{\Delta_{km}^2}\Big] - 2\theta_m^*\Big(\hat{\theta}_{km} - a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\Big) + (\theta_m^*)^2\Bigg)$$

Now let us prove that this is the case,

$$\mathbb{E}\big[||\hat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*||^2\big] = \sum_{m=1}^{p} \mathbb{E}\big[(\hat{\theta}_{(k+1)m} - \theta_m^*)^2\big]$$

Let us examine the expectation of each term.

$$\mathbb{E}\big[(\hat{\theta}_{(k+1)m} - \theta_m^*)^2\big] = \mathbb{E}\big[(\hat{\theta}_{(k+1)m})^2\big] - 2\theta_m^* \mathbb{E}\big[\hat{\theta}_{(k+1)m}\big] + (\theta_m^*)^2$$

$$\mathbb{E}\big[\hat{\theta}_{(k+1)m}\big] = \mathbb{E}\big[\hat{\theta}_{km} - a_k \hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big] = \hat{\theta}_{km} - a_k \mathbb{E}\big[\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big]$$

$\mathbb{E}\big[\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big] \approx g_m(\hat{\boldsymbol{\theta}}_k)$ by equation (7.4) from [1]

$$\mathbb{E}\big[\hat{\theta}_{(k+1)m}\big] \approx \hat{\theta}_{km} - a_k g_m(\hat{\boldsymbol{\theta}}_k)$$

$$\mathbb{E}\big[(\hat{\theta}_{(k+1)m})^2\big] = \mathbb{E}\Big[\big(\hat{\theta}_{km} - a_k \hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big)^2\Big] =$$

$$\mathbb{E}\Big[\hat{\theta}_{km}^2 - 2a_k \hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\hat{\theta}_{km} + a_k^2\big(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big)^2\Big] \approx \hat{\theta}_{km}^2 - 2a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\hat{\theta}_{km} + a_k^2 \mathbb{E}\Big[\big(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big)^2\Big]$$

$$\mathbb{E}\Big[\big(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big)^2\Big] = \mathbb{E}\Big[\big(\frac{y(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - y(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k\Delta_{km}}\big)^2\Big] \approx$$

$\mathbb{E}\Big[\big(\frac{\mathbf{g}_k(\hat{\boldsymbol{\theta}}_k)^T\boldsymbol{\Delta}_k}{\Delta_{km}} + \frac{\varepsilon_{2k}}{2c_k\Delta_{km}}\big)^2\Big]$ By the derivation of (7.3) from [1] without taking any expectation

$\varepsilon_{2k} = \varepsilon_{k+} - \varepsilon_{k-}$ where each noise term is the noise from each perturbation

$$\mathbb{E}\Big[\big(\frac{\mathbf{g}_k(\hat{\boldsymbol{\theta}}_k)^T\boldsymbol{\Delta}_k}{\Delta_{km}} + \frac{\varepsilon_{2k}}{2c_k\Delta_{km}}\big)^2\Big] = \mathbb{E}\Big[\frac{\big(\mathbf{g}_k(\hat{\boldsymbol{\theta}}_k)^T\boldsymbol{\Delta}_k\big)^2}{\Delta_{km}^2} + \frac{\varepsilon_{2k}\mathbf{g}_k(\hat{\boldsymbol{\theta}}_k)^T\boldsymbol{\Delta}_k}{c_k\Delta_{km}^2} + \frac{\varepsilon_{2k}^2}{4c_k^2\Delta_{km}^2}\Big]$$

Assuming i.i.d. noise with mean 0, variance σ^2, and independent from the perturbations,

$$\mathbb{E}\Big[\frac{\big(\mathbf{g}_k(\hat{\boldsymbol{\theta}}_k)^T\boldsymbol{\Delta}_k\big)^2}{\Delta_{km}^2} + \frac{\varepsilon_{2k}\mathbf{g}_k(\hat{\boldsymbol{\theta}}_k)^T\boldsymbol{\Delta}_k}{c_k\Delta_{km}^2} + \frac{\varepsilon_{2k}^2}{4c_k^2\Delta_{km}^2}\Big] = \mathbb{E}\Big[\frac{\big(\mathbf{g}_k(\hat{\boldsymbol{\theta}}_k)^T\boldsymbol{\Delta}_k\big)^2}{\Delta_{km}^2}\Big] + \frac{\sigma^2}{2c_k^2} \cdot \mathbb{E}\big[\frac{1}{\Delta_{km}^2}\big]$$

$$\mathbb{E}\Big[\frac{\big(\mathbf{g}_k(\hat{\boldsymbol{\theta}}_k)^T\boldsymbol{\Delta}_k\big)^2}{\Delta_{km}^2}\Big] = \mathbb{E}\Big[\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} L_i^{'}(\hat{\boldsymbol{\theta}}_k)L_j^{'}(\hat{\boldsymbol{\theta}}_k)\Delta_{ki}\Delta_{kj}}{\Delta_{km}^2}\Big] =$$

$$\mathbb{E}\Big[\big(L_m^{'}(\hat{\boldsymbol{\theta}}_k)\big)^2 + 2\sum_{l\neq m}^{p} L_l^{'}(\hat{\boldsymbol{\theta}}_k)L_m^{'}(\hat{\boldsymbol{\theta}}_k)\frac{\Delta_{kl}}{\Delta_{km}} + \sum_{i\neq 1}^{p}\sum_{j\neq m}^{p} L_i^{'}(\hat{\boldsymbol{\theta}}_k)L_j^{'}(\hat{\boldsymbol{\theta}}_k)\frac{\Delta_{ki}\Delta_{kj}}{\Delta_{km}^2}\Big] =$$

$$(L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + \sum_{i\neq 1}^{p}\sum_{j\neq m}^{p} L_i^{'}(\hat{\boldsymbol{\theta}}_k)L_j^{'}(\hat{\boldsymbol{\theta}}_k)\mathbb{E}\Big[\frac{\Delta_{ki}\Delta_{kj}}{\Delta_{km}^2}\Big]$$

Where the second term has expectation 0 by the last sentence in the first paragraph of page 180 in [1]. Because the perturbations are i.i.d. with mean 0, the sum simplifies further,

$$(L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + \sum_{i\neq m}^{p}\sum_{j\neq m}^{p} L_i^{'}(\hat{\boldsymbol{\theta}}_k)L_j^{'}(\hat{\boldsymbol{\theta}}_k)\mathbb{E}\Big[\frac{\Delta_{ki}\Delta_{kj}}{\Delta_{km}^2}\Big] = (L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + \sum_{i\neq m}^{p}\big(L_i^{'}(\hat{\boldsymbol{\theta}}_k)\big)^2\mathbb{E}\Big[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\Big]$$

Thus we finally get that

$$\mathbb{E}\Big[\big(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big)^2\Big] \approx (L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + \sum_{i\neq m}^{p}\big(L_i^{'}(\hat{\boldsymbol{\theta}}_k)\big)^2\mathbb{E}\Big[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\Big] + \frac{\sigma^2}{2c_k^2}\cdot\mathbb{E}\big[\frac{1}{\Delta_{km}^2}\big]$$

Then putting together everything, we get

$$\mathbb{E}[||\hat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*||^2] \approx \sum_{m=1}^{p}\Bigg(\hat{\theta}_{km}^2 - 2a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\hat{\theta}_{km} + a_k^2(L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + a_k^2\sum_{i\neq m}^{p}\big(L_i^{'}(\hat{\boldsymbol{\theta}}_k)\big)^2\mathbb{E}\Big[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\Big] + \frac{\sigma^2 a_k^2}{2c_k^2}\cdot\mathbb{E}\big[\frac{1}{\Delta_{km}^2}\big] - 2\theta_m^*\Big(\hat{\theta}_{km} - a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\Big) + (\theta_m^*)^2\Bigg)$$

In terms of the bias-variance trade-off, we have the following,

$$\mathbb{E}\big[||\hat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*||^2|\hat{\boldsymbol{\theta}}_k\big] = \sum_{m=1}^{p}\mathbb{E}\big[(\hat{\theta}_{(k+1)m} - \theta_m^*)^2\big] = \sum_{m=1}^{p}\text{Variance}_m + \text{Bias}_m^2 =$$

$$\text{Bias}_m^2 \approx E^2[\hat{\theta}_{(k+1)m} - \theta_m^*] = E^2\Big[\hat{\theta}_{(k+1)m}\Big] - 2\theta_m^* E\Big[\hat{\theta}_{(k+1)m}\Big] + (\theta_m^*)^2 =$$

$$\Big(\hat{\theta}_{km} - a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\Big)^2 - 2\theta_m^*\Big(\hat{\theta}_{km} - a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\Big) + (\theta_m^*)^2$$

$$\text{Variance}_m \approx E^2\Big[\hat{\theta}_{km}\Big] - E^2\Big[\hat{\theta}_{km}\Big] =$$

$$\hat{\theta}_{km}^2 - 2a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\hat{\theta}_{km} + a_k^2\mathbb{E}\Big[\big(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big)^2\Big] - \Big(\hat{\theta}_{km} - a_k g_{km}(\hat{\boldsymbol{\theta}}_k)\Big)^2 =$$

$$\mathbb{E}\Big[\big(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\big)^2\Big] = a_k^2(L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + a_k^2\sum_{i\neq m}^{p}\big(L_i^{'}(\hat{\boldsymbol{\theta}}_k)\big)^2\mathbb{E}\Big[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\Big] + \frac{\sigma^2 a_k^2}{2c_k^2}\cdot\mathbb{E}\big[\frac{1}{\Delta_{km}^2}\big]$$

Let us now examine $\mathbb{E}\left[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\right]$ for $i \neq m$ for the Bernoulli and the even monomial distributions.

$$X \sim \text{Bernoulli } {-1}, 1 \text{ with p} = 0.5$$
$$\mathbb{E}[x^2] = 0.5 \cdot (-1)^2 + 0.5 \cdot (1)^2 = 1$$
$$\mathbb{E}[\frac{1}{x^2}] = 0.5 \cdot \frac{1}{(-1)^2} + 0.5 \cdot \frac{1}{(1)^2} = 1$$
$$\mathbb{E}\left[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\right] = \mathbb{E}[\Delta_{ki}^2]\mathbb{E}[\frac{1}{\Delta_{km}^2}] = 1$$
$$\mathbb{E}\left[\left(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\right)^2\right] \approx (L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + \sum_{i \neq m}^{p} \left(L_i^{'}(\hat{\boldsymbol{\theta}}_k)\right)^2 + \frac{\sigma^2}{2c_k^2}$$

And the even monomial distributions,

$$Y \sim \frac{n+1}{2}y^n \quad n \in 2\mathbb{N} \quad y \in [-1, 1]$$
$$\mathbb{E}[y^2] = \frac{n+1}{2}\int_{-1}^{1} y^2 y^n dy = \frac{n+1}{2} \cdot \frac{2}{n+3} = \frac{n+1}{n+3}$$
$$\mathbb{E}[\frac{1}{y^2}] = \frac{n+1}{2}\int_{-1}^{1} y^{-2} y^n dy = \frac{n+1}{2} \cdot \frac{2}{n-1} = \frac{n+1}{n-1}$$
$$\mathbb{E}\left[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\right] = \mathbb{E}[\Delta_{ki}^2]\mathbb{E}[\frac{1}{\Delta_{km}^2}] = \frac{\frac{n+1}{n+3}}{\frac{n+1}{n-1}} = \frac{n-1}{n+3}$$
$$\mathbb{E}\left[\left(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\right)^2\right] \approx (L_m^{'}(\hat{\boldsymbol{\theta}}_k))^2 + \sum_{i \neq m}^{p} \frac{n-1}{n+3}\left(L_i^{'}(\hat{\boldsymbol{\theta}}_k)\right)^2 + \frac{\sigma^2}{2c_k^2}\frac{n+1}{n-1}$$

Let us now analyze these results.

3 Analysis

The only place, then, where any valid perturbation differs from the others is in $\mathbb{E}\left[\left(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\right)^2\right]$, in the variance part of the MSE. We would like to minimize our MSE, and hence the variance of our estimates

More specifically, the second term is $\sum_{i\neq m}^{p}\left(L_i^{'}(\hat{\theta}_k)\right)^2\mathbb{E}\left[\frac{\Delta_{ki}^2}{\Delta_{km}^2}\right]$, essentially a scaled gradient magnitude. The third term is $\frac{\sigma^2}{2c_k^2}\cdot\mathbb{E}\left[\frac{1}{\Delta_{km}^2}\right]$, a scaled noise term. To make $\mathbb{E}\left[\left(\hat{g}_{km}(\hat{\boldsymbol{\theta}}_k)\right)^2\right]$ as small as possible (and hence minimize the variance as much as possible), we would need to have both terms being zero. The objective then, is to reduce these two terms as much as possible.

With the Bernoulli, the multipliers/scale factors are both 1. The even monomial distributions have a $\frac{n-1}{n+3}$ scaling factor on the second term, hence reducing it, and $\frac{n+1}{n-1}$ on the third term, increasing it.

Note that as n gets larger, the estimate of the even monomial distributions approaches that of the Bernoulli. At a low n, then, the even monomial distributions should outperform the Bernoulli in the presence of low noise and high gradient magnitude. The opposite should occur when there is high noise and small gradient magnitude. In other cases, it will be a trade-off, reducing the squared gradient magnitude term at the expense of increasing the noise term. The hope is that the gain in reducing variance by reducing the second term (gradient magnitude) offsets the increase in variance by increasing the third term (noise). In the presence of low noise and small gradient magnitude, the results from both perturbation classes should be similar.

Another point to note is the dependence on the dimensionality. The higher the number of dimensions, the higher the second term becomes (more things being summed), and the more variance. Thus it may be that the even order monomials scale better with higher dimensions for SPSA simply because they scale down the second term, and hence there is less variance in the parameter estimate.

4 Numerical Results

The results from the numerical experiments seem to confirm the analysis from above. three different loss functions were to be optimized in the presence of noise. Specifically, the functions were the following

- Skewed quartic loss function. $\mathbf{B}$ is a $p \times p$ matrix and $\boldsymbol{\theta}$ is a p-dimensional vector. The minimum is 0 at $(0, 0, \cdots, 0)$. This is a very steep function.

$$\boldsymbol{\theta}^T \mathbf{B}^T \mathbf{B} \boldsymbol{\theta} + 0.1 \sum_{i=1}^{p} (\mathbf{B}\boldsymbol{\theta})_i^3 + 0.01 \sum_{i=1}^{p} (\mathbf{B}\boldsymbol{\theta})_i^4$$

- Quadratic loss function. $\boldsymbol{\theta}$ is a p-dimensional vector. The minimum is 0 at $(0, 0, \cdots, 0)$.

$$\boldsymbol{\theta}^T \boldsymbol{\theta}$$

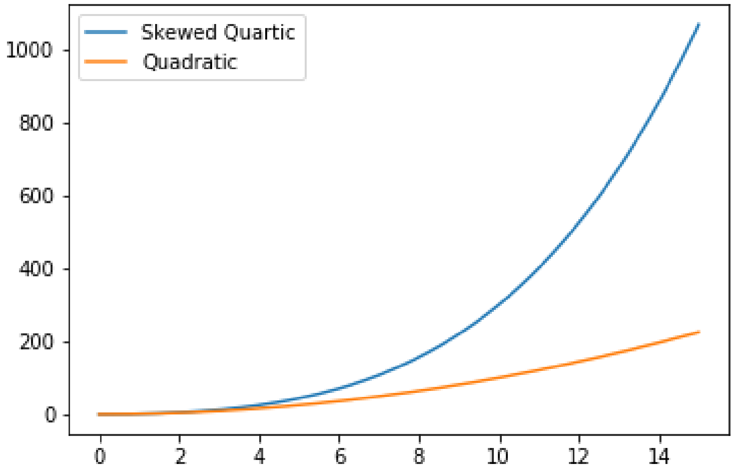
- Ackley loss function. $\boldsymbol{\theta}$ is a p-dimensional vector. The minimum is 0 at $(0, 0, \cdots, 0)$.

$$-20 \cdot \exp\Big(-0.2\sqrt{\frac{1}{p}\sum_{i=1}^{p} \theta_i^2}\Big) - \exp\Big(\frac{1}{p}\sum_{i=1}^{p} cos(2\pi\theta_i)\Big) + 20 + \exp$$

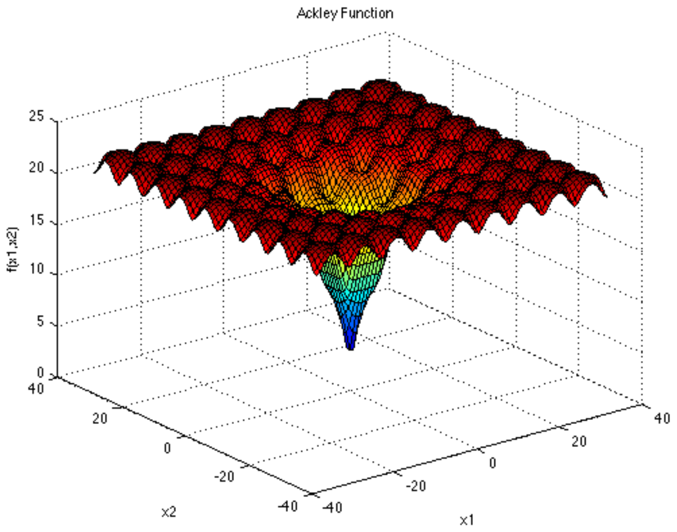
The SPSA algorithm was run with the following parameters $a_k = \frac{0.1}{(k+1+50)^{0.602}}$, $c_k = \frac{1}{(k+1)^{0.101}}$. The noise was normal with mean 0 and standard deviation 0.5. The SPSA algorithm was run with the following perturbation distributions $\frac{3}{2}x^2$, $\frac{9}{2}x^8$, $\frac{101}{2}x^{100}$, and the Bernoulli. For each perturbation and loss, the algorithm had a budget of 100 loss function evaluations and the results were averaged over 50 runs. The parameter was 10 dimensions and the initial vector generated from a uniform random on the interval $[-5, 5]^{10}$. The following are the results.

	Bernoulli	$\frac{3}{2}x^2$	$\frac{9}{2}x^8$	$\frac{101}{2}x^{100}$
Ackley	10.20139475	10.12751731	10.29903367	10.14978409
Quadratic	22.44023186	86.28275554	70.71872389	77.77264939
Skewed Quartic	3.09652132	0.91149493	0.85090612	1.01842016

The results are very interesting. On one hand they seem to confirm that the even monomial distributions perform better with steepr functions. The skewed quartic is a very steep function and with that loss function the Bernoulli performed worse. With the quadratic loss function, the Bernoulli performed better. The difference in steepness is exemplified by the 2-dimensional plots of the quadratic and the skewed quartic functions.



It did not seem to matter what perturbation was used with the Ackley function. This may be due to the fact that it is a very undulating function, with many hills and valleys. This means that we are in a low noise, low magnitude gradient environment, so it makes sense that the performance is similar by all the perturbation distributions. The 3-dimensional plot demonstrates the many undulations.



Another round of simulations were run but this time in the presence of normal noise with standard deviation of 50 and mean 0. The simulations were averaged over 30,000 iterations to account for the high noise and get a stable result. This numerical study was only done with the skewed quartic loss function. This is the function in which the even monomials performed better due to low noise and very high magnitude in the gradient.

The Bernoulli, which is not as hurt by noise, performed best. The low order even monomial (the quadratic) was very hurt by the noise and performed the worst of all. As the order of the monomials grew, the performance improved. This fits in with our theoretical analysis. The low order even monomials are significantly hurt by noise, much more so than higher order even monomials. Furthermore, in very high noise, the Bernoulli should have performed best and it did. As the order of the even monomials grew, and hence their bias terms approached those of a Bernoulli, the performance improvced.

	Bernoulli	$\frac{3}{2}x^2$	$\frac{9}{2}x^8$	$\frac{101}{2}x^{100}$
Skewed Quartic	5.40642709	7.03763302	5.71239234	5.42507183

5 Conclusion and Future Work

Have we managed to beat the Bernoulli distribution? Yes and no. As the famous saying goes, ”there ain’t no free lunch”. It turns out that the hypothesis that even monomial distributions of the form $\frac{n+1}{2} \cdot x^n$ perform better than the Bernoulli distribution for generating perturbations is partially true. There is a trade-off between reducing variance from noise and reducing variance because of a high magnitude gradient. Preliminary numerical results seem to confirm this conclusion. One should be careful, however, in reading too much into these results. They are simply preliminary and need to be followed up with more theoretical analysis and numerical studies.

Another important conclusion to draw is that by the result from the theoretical section, there is a very clear principle on evaluating different perturbation distributions. The different valid distribtutions will have a trade-off between reducing the bias from noise and reducing the bias from the gradient magnitude. Future work should focus on the trade-off between these two factors. It should evaluate different perturbations based on these two factors and explore other possible distributions.

Another potential area of future work is the effect of dimensionality of the parameter on perturbation performance (i.e. how well SPSA performs with that perturbation distribution). There is dependence of the bias on the number of dimensions (i.e. higher dimension means more bias and hence higher mean squared error). This relationship can be explored and delineated further. Furthermore, studies can be done to examine which perturbations scale best with higher dimensions.

In the bias-variance trade-off with SPSA, we can reduce the MSE only through reducing the variance term if we are choosing among different perturbations (there is nothing much we can do about the bias). In our case, it turned out to be a trade-off with reducing variance by minimizing the effect of noise and minimizing the effect of high gradient magnitude. Future work should explore the possibility of perturbations which simultaneously reduce both these terms (and hence reducing variance significantly).

6 References

1. James C. Spall, ”Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control” Wiley, 2003

2. For the Ackley function and picture: https://www.sfu.ca/ ssurjano/ackley.html