# Evaluate the use of non-Bernoulli distributions for the perturbations in SPSA

Keni Xue

February 11, 2025

**Abstract**

Currently, Simultaneous Perturbation Stochastic Approximation distributions can solve many optimization problems, because it only can solve problems where we only know the objective function. Therefore, it is widely applied in many fields, such as astronomy, traffic problems, etc. Generally, Bernoulli distribution is used as the perturbation vector in SPSA. However, non-Bernoulli distribution works better than Bernoulli distribution on some condition when sample size is small in SPSA algorithm. In this report, author will find non-Bernoulli distributions including Segmented Uniform distribution and U-shaped distribution, and provide the condition that non-Bernoulli distributions have better performance than Bernoulli as perturbations in SPSA.

# Introduction

Nowadays, people face many optimization problems in many fields, such as Simultaneous Perturbation Stochastic Approximation (SPSA). Simultaneous Perturbation Stochastic Approximation (SPSA) is an algorithm used for optimizing systems with multiple unknown parameters. SPSA is a powerful tool to solving the multivariate optimization problems, since one of its importance feature is its underlying gradient approximation only requiring two loss function measurements, rather than the number of parameters being optimized[7]. Compared with other optimization algorithms, such as FDSA, which requires direct but often difficult or impossible to obtain measurements of the gradient of the objective function, SPSA algorithm only use one objective function measurements[1]. Thus, SPSA algorithm is widely applied in many fields, such as bioprocess control, neural network training, and human-machine interaction[4]. Moreover, SPSA is good to applied in high-dimensional problems, so this could applied in many practical problem where have the large number of term could be optimized by SPSA[1].

SPSA algorithm uses a p-dimensional random perturbation vector $\Delta_k$ to obtain estimate the gradient. Typically, SPSA algorithm uses Bernoulli distribution with probability of $\frac{1}{2}$ for each $\pm 1$ outcome as perturbation vector, because it is effective and theoretically valid. Cao[3] showed it is easy to implement Bernoulli distribution as perturbation vectors in SPSA algorithm and is proves that Bernoulli distribution is asymptotically optimal. However, Bernoulli distribution may not have optimality when stochastic approximation has small sample size. When running SPSA, it is too physically or computationally expensive to to evaluate system performances when using Bernoulli distribution as perturbation in small sample size, such as experiments on a complicated control system[5]. Therefore, there may exist some non-Bernoulli distribution acting as perturbations in SPSA and perform better than the performance with Bernoulli perturbation when the sample size is small.

This report is to examined a wide range of candidate non-Bernoulli distributions as optimal distribution for simultaneous perturbations, which will involve both mathematical analysis and numerical analysis, and compared algorithm performance. The performance of such alternative distributions is split uniform distribution with different parameters, and a symmetric u-shaped distribution. The objectives considered here to minimize the mean square error of the estimate to compare the performance between those perturbations. The rest of the report would include methodology of SPSA in second section, Theoretical Analysis in third section, numerical results in 4th section, and conclusion in final section.

# Methodology

## 0.1 Problem Formulation

Denote $\theta \in \Theta \subset R^p$ as a p-dimensional vector of input parameters. Let $y(\theta)$ as the function of observed $\theta$ with stochastic effects $\epsilon$: $y(\theta) = L(\theta) + \epsilon$, where $L(\theta)$ is loss function and $\epsilon$ is i.i.d noise with mean zero and variance $\sigma^2$. Thus, the expected system performance at $\theta$ is $L(\theta) = E(y(\theta))$. The objective is to

$$\min_{\theta \in \Theta} L(\theta)$$

The the basic unconstrained SPSA algorithm is in the following iterative SA scheme:

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

where $\hat{\boldsymbol{\theta}}_k$ is the estimate of $\theta$ at iteration $\hat{\boldsymbol{g}}_k(\bullet) \in R^p$ is the simultaneous perturbation estimate of the gradient $g(\theta) = \partial L / \partial \theta$ at iteration $\hat{\boldsymbol{\theta}}_k$. The scalar gain coefficient $\{a_k\}$ is non-negative, decreasing and converging to zero.

Generally, it is most effective to use direct estimates of the gradient in optimization proble, while in many practical problem it is impossible to apply[2]. Thus, we assume there is no direct form of $g(\theta)$ and $L(\theta)$are available, and there only exists measurements of $L(\theta)$

.Also, the estimate $\hat{\theta}_k$ will converge to the optimal value $\theta^*$, under suitable conditions on the loss function and gradient.

## 0.2 simultaneous perturbation stochastic approximation

In SPSA algorithm, simultaneous perturbation means, all elements of $\hat{\boldsymbol{\theta}}_k$ are randomly perturbed together, so it could obtain two loss measurements $y(\cdot)$ at same time. It is more efficient than finite-difference method, because it only requires two function evaluations at each iteration[5].

Let $\mathcal{G}_k(\theta)$ denote the simultaneous perturbation estimate of $g(\theta)$ and let $\hat{\theta}_k$ denote the estimate for $\theta^*$ at iteration $k$. Let $\Delta_k$ be a vector of $p$ independent random variables at iteration $k$:

$$\Delta_k = \begin{bmatrix} \Delta_{k_1} & \Delta_{k_2} & \cdots & \Delta_{k_p} \end{bmatrix}^{\mathrm{T}}$$

The components of $\Delta_k$ should be independent, such as Bernoulli variables mapped to $\{-1, 1\}$. Let $c_k$ be a sequence of positive scalars. For each iteration, we could dtake measurements of $L$ at $\hat{\theta}_k \pm c_k\Delta_k$:

$$y\left(\hat{\theta}_k + c_k\Delta_k\right) = L\left(\hat{\theta}_k + c_k\Delta_k\right) + \epsilon_k^+$$
$$y\left(\hat{\theta}_k - c_k\Delta_k\right) = L\left(\hat{\theta}_k + c_k\Delta_k\right) + \epsilon_k^-$$

where $\epsilon_k^{\pm}$ are random error terms.

Thus, two-sided SP-gradient approximation is:

$$\hat{\boldsymbol{g}}_k\left(\hat{\boldsymbol{\theta}}_k\right) = \begin{bmatrix} \frac{y\left(\hat{\boldsymbol{\theta}}_k + c_k\Delta_k\right) - y\left(\hat{\boldsymbol{\theta}}_k - c_k\Delta_k\right)}{2c_k\Delta_{k1}} \\ \vdots \\ \frac{y\left(\hat{\boldsymbol{\theta}}_k + c_k\Delta_k\right) - y\left(\hat{\boldsymbol{\theta}}_k - c_k\Delta_k\right)}{2c_k\Delta_{kp}} \end{bmatrix}$$
$$= \frac{y\left(\hat{\boldsymbol{\theta}}_k + c_k\Delta_k\right) - y\left(\hat{\boldsymbol{\theta}}_k - c_k\Delta_k\right)}{2c_k}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right]^T$$

**SPSA procedure show below:**

**1.Initialize and select coefficients** Set counter index $k = 0$. Pick initial guess $\hat{\boldsymbol{\theta}}_0$; coefficients $a, c, A, \alpha$, and $\gamma$ in the SPSA should be nonnegative,where $\alpha = 0.602$ and $\gamma = 0.101$ is practically effective and valid. Then the gain sequences are

$$a_k = a/(k+1+A)^{\alpha} \text{ and } c_k = c/(k+1)^{\gamma}$$

.

**2. Generalize the simultaneous perturbation vector** In our report, we will focus this part for different simultaneous perturbation vector.

**3. Loss function evaluations**: two measurements of the loss function could be obtained by the simultaneous perturbations around the current $\hat{\boldsymbol{\theta}}_k$

**4.Gradient approximation**: Generalize the simultaneous perturbation approximation to the unknown gradient $\boldsymbol{g}\left(\hat{\boldsymbol{\theta}}_k\right)$.

**5.Iteration or termination**: Return to step 1 with $(k+1)$ replacing$(k)$. we could terminate the algorithm when SPSA algorithm has reached the maximum allowable number of iterations.

## 0.3 Perturbation Distribution for SPSA

Typically, Bernoulli distribution $\{-1, +1\}$ is used as perturbation in gradient estimate because of its efficiency and theoretical validation. Actually, if the convergence of the algorithm is satisfied, then the perturbation distribution could be applied in this algorithm. Thus, the assumptions on $\Delta_k$ should be satisfied[3]:

1. $\{\Delta_{ki}\}$ are independent for all $k, i$, identically distributed for all $i$ at each $k$.

2. $\{\Delta_i\}$ are symmetrically distributed about zero and uniformly bounded in magnitude for all $k, i$.

**3.** For all $k$, $E\left[\left(\frac{y(\hat{\theta}_k+c_k\Delta_k)}{\Delta_{ki}}\right)^2\right]$ is uniformly bounded over $k$ and $i$.

In condition 3 , if the ratio of measurement of perturbation $E\left[\left(\hat{y}\left(\theta_k \pm c_k\dot{\Delta}_k\right)/\Delta_{ki}\right)^2\right]$ is uniformly bounded if there exists a $\tau > 0$ such that $(1+\eta)^{-1} + (1+\tau)^{-1} = 1$ and inverse moments $E\left(\left|\frac{1}{\Delta_{ki}}\right|^{2+2\tau}\right) \leq C \quad$ for some $\mathbf{C} > 0$[1]. From Introduction to Stochastic Search and Optimization[4], we could know valid distributions have the Bernoulli $\pm 1$, segmented uniform, U-shape distribution, and many others, where they have finite inverse moments. Two common distribution, uniformed distribution and Normal distribution $N(0, \sigma^2)$, are symmetrical, have zero mean but their inverse moments are not finite, thus they are not valid distribution for perturbation vector in SPSA. It is because both of these distributions have the amount of probability mass near zero[4]. The picture below is from Introduction to Stochastic Search and Optimization[6].

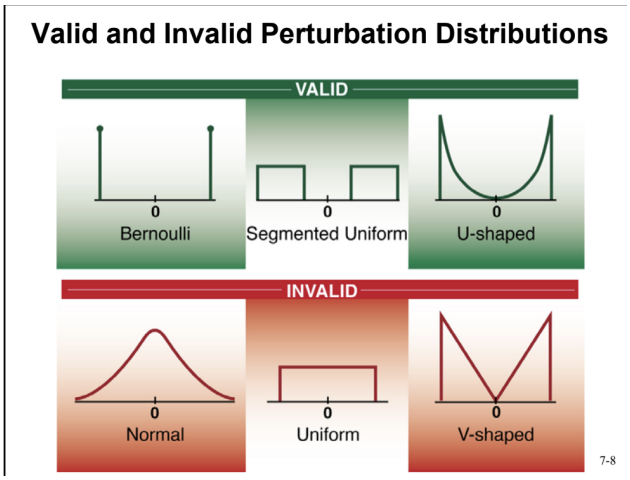


Figure 1: Valid and Invalid Perturbation Distributions

In this report, we will analyze and compare segmented uniform (SU) distributions, U-shaped distribution with the Bernoulli distribution. To ensure a fair comparison, we should make sure that all distributions as peturbation should be normalized so that their variances should be 1 and their mean should be 0. The probability density functions of these distributions are given above.

# THEORETICAL ANALYSIS

Now we provide mathematical analysis that non-Bernoulli distribution works better than Bernoulli distribution, where non-Bernoulli distributions includes Segmented Uniform distribution(SU) and U-shaped distribution.

**Find the valid parameters on Segmented Uniform Distribution.** In order for a fair comparison, Segmented Uniformed distribution should have same mean and variance as Bernoulli distribution. After calculation, we get Bernoulli distribution have mean 0 and variance 1. Also, as we discussed on methodology, SU distribution are symmetrically distributed about zero and uniformly bounded, thus we set $[-b,-a]\cup[a,b]$ where $a,b\in R$. To calculate the variance, we find the parameters of SU must satisfy: $a^2+b^2+ab=3$. Thus, we get its probability denity function

$$f(x)=\begin{cases}\frac{1}{2(b-a)} & [-b,-a]\cup[a,b]\\ 0 & o.w.\end{cases}$$

and its inverse moment:

$$E\left(\frac{1}{x^2}\right)=\int_{-b}^{-a}\frac{1}{2(b-a)}\frac{1}{x^2}dx+\int_{a}^{b}\frac{1}{x(b-a)}\frac{1}{x^2}dx=\frac{1}{2(b-a)}\left(\frac{2}{a}-\frac{2}{b}\right)=\frac{1}{ab}$$

Thus, after calculation, we get three SU uniform that

$$f_{SU}(\mathrm{x};0.5,1.472)$$

$$f_{SU}(\mathrm{x};0.9,1.0986)$$

$$f_{SU}(\mathrm{x};0.4092,1.4908)$$

**Find the valid parameters on U-shaped Distribution.** We choose the U-shaped distribution where mean 0 and variance 1 and uniformly bounded. After calculation based on

basic statistics and simple algebra, we get the probability density fucntion:

$$f_U(\mathrm{x}) = \frac{9\sqrt{15}}{50}x^2, \text{where } x \in (\frac{-\sqrt{15}}{3}, \frac{\sqrt{15}}{3})$$

And its inverse moment is

$$\begin{aligned} E\left(\frac{1}{x^2}\right) &= \int_{-\sqrt{15}/3}^{\sqrt{15}/3} \frac{9\sqrt{15}}{50}x^2 \cdot \frac{1}{x^2}dx \\ &= \frac{9\sqrt{15}}{50}[x]_{-\sqrt{15}/3}^{\sqrt{15}/3} = \frac{9}{5} \end{aligned}$$

which is a bounded. Thus, we summarized moments of all the perturbations which we used in our report on Table, where i and j are the elements of $\Delta_0$ and $i \neq j$.

Table: Moments of perturbations under two distributions

| Expectation | Bernoulli | SU(0.5, 1.472) | SU(0.9, 1.0986) | SU(0.4092, 1.4908) | U-shaped |
|---|---|---|---|---|---|
| $E\left(\Delta_{0i}\right)$ | 0 | 0 | 0 | 0 | 0 |
| $E\left(\Delta_{0i}/\Delta_{0j}\right)$ | 0 | 0 | 0 | 0 | 0 |
| $E\left(\Delta_{0i}^2/\Delta_{0j}^2\right)$ | 1 | 125/92 | 50000/49167 | 100/61 | 9/5 |
| $E\left(1/\Delta_{0i}^2\right)$ | 1 | 125/92 | 50000/49167 | 100/61 | 9/5 |

We used simplest version, i.e. iteration = 1, to analyze SPSA algorithm when the sample size is small, because it is to complicated to analyze when k is large. In this paper, we assume that loss function has third derivatives. We used mean squared error (MSE)

$$E\left(\left\|\hat{\theta}_i - \theta^*\right\|^2\right) = E\left(\left(\hat{\theta}_i - \theta^*\right)^T\left(\hat{\theta}_i - \theta^*\right)\right)$$

to evaluate and compare the performance among those distributions, where $\hat{\theta}_i$ is the updated estimate of $\theta$ after i iteration, and $\theta^*$ is optimal and true value of $\theta$.

**Compare Segmented Uniform distribution with Bernoulli distribution.** We compute the difference of mean squared error (MSE) between SU uniform and Bernoulli distri-

bution as follow:

$$
\begin{aligned}
&E_{SU}\left(\left\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^*\right\|^2\right)-E_B\left(\left\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^*\right\|^2\right)\\
&=\sum_{i=1}^{p}\left(a_{0SU}^2\left(\frac{\partial L_i}{\partial\theta_i}+\sum_{\substack{j=1,\ldots,p\\ j\neq i}}\left(\frac{\partial {L_i}^2}{\partial\theta_i}/ab\right)\right)-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_i}{\partial\theta_i}\left(\hat{\theta}_{0i}-\theta_i^*\right)\right)\\
&\quad-p\left(a_{0B}^2\left(\sum_{i=1}^{p}\frac{\partial {L_i}^2}{\partial\theta_i}+0.5\sigma^2/c_{0B}^2\right)-a_{0SU}^2\sigma^2/2abc_{0SU}^2\right)+O\left(c_0^2\right),
\end{aligned}
$$

, where $\hat{\theta}_{0i}$ is ith component of $\hat{\boldsymbol{\theta}}_0$ and $\theta_i^*$ is ith component $\boldsymbol{\theta}^*$, and

$$
\begin{aligned}
\frac{\partial L_1}{\partial\theta_1}&=3\left(t_1-2\right)^3+2\left(t_1-2t_2\right)\\
\frac{\partial L_2}{\partial\theta_2}&=-4\left(t_1-2t_2\right)
\end{aligned}
$$

Since we will use loss function which is quadratic function which p=2, so we could produce explicit form and applied in numerical example. For a quadratic loss function with p=2, the MSE of Segmented Uniformed distribution is smaller than the MSE of Bernoulli distribution between if

$$
E_{SU}\left(\left\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^*\right\|^2\right)-E_B\left(\left\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^*\right\|^2\right)<0
$$

where

$$
\begin{aligned}
&E_{SU}\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)-E_B\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)\\
&=a_{0SU}^2\left(\frac{\partial {L_1}^2}{\partial\theta_1}+\frac{\partial {L_2}^2}{\partial\theta_2}+\frac{\partial {L_1}^2}{\partial\theta_1}/ab+\frac{\partial {L_2}^2}{\partial\theta_2}/ab\right)\\
&-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_1}{\partial\theta_1}\left(\hat{\theta}_{01}-\theta_1^*\right)\\
&-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_2}{\partial\theta_2}\left(\hat{\theta}_{02}-\theta_2^*\right)-2a_{0B}^2\left(\frac{\partial {L_1}^2}{\partial\theta_1}+\frac{\partial {L_2}^2}{\partial\theta_2}+0.5\sigma^2/c_{0B}^2\right)\quad+a_{0SU}^2\sigma^2/abc_{0SU}^2,
\end{aligned}
$$

Where the calculation based on the theorem from [3]. Since we have three SU distribu-

tions with different parameters. Thus, in $f_{SU}(\mathrm{x}; 0.5, 1.472)$,

$$
\begin{aligned}
&E_{SU_1}\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)-E_B\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)\\
&=a_{0SU}^2\left(\frac{\partial {L_1}^2}{\partial\theta_1}+\frac{\partial {L_2}^2}{\partial\theta_2}+125\frac{\partial {L_1}^2}{\partial\theta_1}/92+125\frac{\partial {L_2}^2}{\partial\theta_2}/92\right)\\
&-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_1}{\partial\theta_1}\left(\hat{\theta}_{01}-\theta_1^*\right)-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_2}{\partial\theta_2}\left(\hat{\theta}_{02}-\theta_2^*\right)\\
&-2a_{0B}^2\left(\frac{\partial {L_1}^2}{\partial\theta_1}+\frac{\partial {L_2}^2}{\partial\theta_2}+0.5\sigma^2/c_{0B}^2\right)+125a_{0SU}^2\sigma^2/92c_{0SU}^2
\end{aligned}
$$

In $f_{SU}(\mathrm{x}; 0.9, 1.0986)$, $E_{SU_2}$ is

$$
\begin{aligned}
&E_{SU_2}\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)-E_B\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)\\
&=a_{0SU}^2\left(\frac{\partial {L_1}^2}{\partial\theta_1}+\frac{\partial {L_2}^2}{\partial\theta_2}+50000\frac{\partial {L_1}^2}{\partial\theta_1}/49167+50000\frac{\partial {L_2}^2}{\partial\theta_2}/49167\right)\\
&-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_1}{\partial\theta_1}\left(\hat{\theta}_{01}-\theta_1^*\right)-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_2}{\partial\theta_2}\left(\hat{\theta}_{02}-\theta_2^*\right)\\
&-2a_{0B}^2\left(\frac{\partial {L_1}^2}{\partial\theta_1}+\frac{\partial {L_2}^2}{\partial\theta_2}+0.5\sigma^2/c_{0B}^2\right)+50000a_{0SU}^2\sigma^2/49167c_{0SU}^2
\end{aligned}
$$

And in $f_{SU}(\mathrm{x}; 0.4092, 1.4908)$, we get

$$
\begin{aligned}
&E_{SU_3}\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)-E_B\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)\\
&=a_{0SU}^2\left(\frac{\partial {L_1}^2}{\partial\theta_1}+\frac{\partial {L_2}^2}{\partial\theta_2}+100\frac{\partial {L_1}^2}{\partial\theta_1}/61+100\frac{\partial {L_2}^2}{\partial\theta_2}/61\right)\\
&-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_1}{\partial\theta_1}\left(\hat{\theta}_{01}-\theta_1^*\right)-2\left(a_{0SU}-a_{0B}\right)\frac{\partial L_2}{\partial\theta_2}\left(\hat{\theta}_{02}-\theta_2^*\right)\\
&-2a_{0B}^2\left(\frac{\partial {L_1}^2}{\partial\theta_1}+\frac{\partial {L_2}^2}{\partial\theta_2}+0.5\sigma^2/c_{0B}^2\right)+100a_{0SU}^2\sigma^2/61c_{0SU}^2
\end{aligned}
$$

**Compare U-shaped distribution with Bernoulli distribution.**Similarly, we compute the difference of mean squared error (MSE) between U-shaped distribution and Bernoulli

distribution as follow:

$$
\begin{aligned}
&E_U\left(\left\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^*\right\|^2\right)-E_B\left(\left\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^*\right\|^2\right)\\
&=\sum_{i=1}^{p}\left(a_{0U}^2\left(\frac{\partial L_i}{\partial\theta_i}+\sum_{\substack{j=1,\ldots,p\\ j\neq i}}\left(9\frac{\partial L_i}{\partial\theta_i}^2/5\right)\right)-2\left(a_{0U}-a_{0B}\right)\frac{\partial L_i}{\partial\theta_i}\left(\hat{\theta}_{0i}-\theta_i^*\right)\right)\\
&\quad-p\left(a_{0B}^2\left(\sum_{i=1}^{p}\frac{\partial L_i}{\partial\theta_i}^2+0.5\sigma^2/c_{0B}^2\right)-9a_{0SU}^2\sigma^2/10c_{0SU}^2\right)+O\left(c_0^2\right),
\end{aligned}
$$

If we want to find the condition where U-shaped distribution outperform Bernoulli distribution, we must have

$$
E_U\left(\left\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^*\right\|^2\right)-E_B\left(\left\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^*\right\|^2\right)<0
$$

Especially, when the loss function is quadratic with 2 dimensions, we must keep following equation negative:

$$
\begin{aligned}
&E_U\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)-E_B\left(\left\|\hat{\theta}_1-\theta^*\right\|^2\right)\\
&=a_{0U}^2\left(\frac{\partial L_1}{\partial\theta_1}^2+\frac{\partial L_2}{\partial\theta_2}^2+9\frac{\partial L_1}{\partial\theta_1}^2/5+9\frac{\partial L_2}{\partial\theta_2}^2/5\right)\\
&-2\left(a_{0U}-a_{0B}\right)\frac{\partial L_1}{\partial\theta_1}\left(\hat{\theta}_{01}-\theta_1^*\right)-2\left(a_{0U}-a_{0B}\right)\frac{\partial L_2}{\partial\theta_2}\left(\hat{\theta}_{02}-\theta_2^*\right)\\
&-2a_{0B}^2\left(\frac{\partial L_1}{\partial\theta_1}^2+\frac{\partial L_2}{\partial\theta_2}^2+0.5\sigma^2/c_{0B}^2\right)+9a_{0U}^2\sigma^2/5c_{0U}^2
\end{aligned}
$$

# NUMERICAL EXAMPLE

Let the quadratic loss function $L(\theta)=(t_1-2)^4+(t_1-2t_2)^2$ in Exercise 1.10 in [book], where $\boldsymbol{\theta}=[t_1,t_2]^T$, where noise is $\sigma$ is 0.1, and initial guess $\hat{\theta}_0=[0,3]^T$( So $L(\theta_0)=52$). The parameters is chosen by the gain-selection guidelines in Section 7.5[4]. To the simplest setting, we applying 500 loss measurements and one replication, we get gain sequence parameters as

follow table:

Table: Gain Sequence parameters under different distributions

| Distribution | Bernoulli | SU(0.5, 1.472) | SU(0.9, 1.0986) | SU(0.4092, 1.4908) | U-shaped |
|---|---|---|---|---|---|
| $A$ | 50 | 50 | 50 | 50 | 50 |
| $a_k$ | 0.02343 | 0.02489 | 0.02423 | 0.02043 | 0.02211 |
| $c_k$ | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |

Thus we should verify that whether the difference of MSE between the non-Bernoulli distribution an Bernoulli $\pm 1$ distribution is negative by the formulas in theoretical analysis. Calculate the right side of difference of MSE between distributions based above parameters and equations from **THEORETICAL ANALYSIS**, all results are negative, which means all SU distributions and non-Bernoulli distributions outperform Bernoulli when loss measurements 1000 and replication = 1. Table below show the result of MSE of different distributions when loss measurements = 1000 and replication = 1000:

Table: MSE when replication = 1 and loss measurement =1000

| Distribution | Bernoulli | SU(0.5, 1.472) | SU(0.9, 1.0986) | SU(0.4092, 1.4908) | U-shaped |
|---|---|---|---|---|---|
| $MSE$ | 0.03813 | 0.00927 | 0.02611 | 0.01825 | 0.014048 |

This table shows the mean square error (MSE) of SPSA algorithm with differennt distribution as perturbation vector with one iteration and 1000 loss measurements. From the data, we can see that three different Segmented uniform distributions and U-shaped uniform distribution have smaller MSE than Bernoulli distribution. Among these given distributions, the segmented uniform distribution where a= 0.5 and b=1.472 has the lowest MSE, indicating that the algorithm performs best under this distribution.

When iteration is 10 and loss measurements is 1000, the result will shown on table:

Table: MSE when replication = 10 and loss measurement =1000

| Distribution | Bernoulli | SU(0.5, 1.472) | SU(0.9, 1.0986) | SU(0.4092, 1.4908) | U-shaped |
|---|---|---|---|---|---|
| $MSE$ | 0.01208 | 0.00688 | 0.00747 | 0.00315 | 0.00755 |

After 10 replications and 1000 loss measurement, this table again shows the mean square error (MSE) results of SPSA under same parameters as replication = 1 and loss measurements = 1000, but the repications become 10 under different distributions. All different Segmented uniform distributions and U-shaped uniform distribution have better performace than Bernoulli distributions, which is same as the result on replication = 1. The SU(0.4092, 1.4908) distribution has the lowest MSE in this iteration, indicating that SPSA performs best under SU(0.4092, 1.4908) as perturbation vector condition.

### 0.3.1 Analysis

Compare both table of MSE in different iterations under different distributions as perturbation vectors in SPSA, we can find three segmented uniformed distribution we set and u-shaped distribution has better performance than the Bernoulli distribution under the $c, A$ and $a$ that consists the condition that we mentioned on theoretical analysis. Compared those two tables, we can see the MSE of each distribution decreases significantly after the replication times are increased. This indicates that increasing the number of replications can improve the stability of results and reduce errors, so as to obtain more reliable evaluation results. In particular, the MSE of SU(0.4092, 1.4908) distribution decreased the most, indicating that the distribution may be more sensitive to the increase of replication times. And the reason why SU(0.5, 1.472) distribution has the better performace than SU(0.4092, 1.4908) distribution is because when there is only one replication the noise of the loss function leading this error. In summary, there exist non-Benoulli distributions including different segmented uniform distributions and u-shaped distribution have better performance than the

Bernoulli $\pm 1$ distribution by choosing parameters by the gain-selection guidelines in Section 7.5[4], and consists the condition on **theoretical analysis**.

# Conclusion

In this report, we try to find out the performance of non-Bernoulli distributions, including Segmented Uniform distributions with different parameters and U-shaped distribution, as perturbation vectors in SPSA algorithm. We prove the condition that under some condition that choose some specific parameters of gain sequences, a, c, and A, non-Bernoulli distribution will perform better than Bernoulli distribution as perturbations in SPSA with small sample size. In theoretical analysis section, we provide specific conclusion to show the condition when iteration is 1. Moreover, in the numerical result, we verified the conclusion we made in theoretical analysis section, which means non-Bernoulli distributions outperform Bernoulli distribution when the iterations more than 1. Since, in report, we find there are many different parameters that could satisfied the condition SU distributions works better than Bernoulli distribution. In the future, we could find the specific paremeters of a and b in SU distribution as perturbation vector that perform best in SPSA algorithm.

# Reference

[1]SPSA Algorithm. (n.d.). https://www.jhuapl.edu/SPSA/

[2]Hutchison, D. (2002). On an Efficient Distribution of Perturbations for Simulation Optimization using Simultaneous Perturbation Stochastic Approximation. https://www.semanticscholar.org/p
an-Efficient-Distribution-of-Perturbations-for-Hutchison-Hutchison/bae6b40fe4c910b4e169321768c567a91e

[3]Cao, X. (2014, May 5). Non-Bernoulli Perturbation Distributions for Small Samples in Simultaneous Perturbation Stochastic Approximation. arXiv.org. https://arxiv.org/abs/1405.0769

[4]Introduction to Stochastic Search and Optimization: (n.d.). https://www.jhuapl.edu/ISSO/

[5] Fukumori, I., & Malanotte-Rizzoli, P. (1995). An approximate Kaiman filter for ocean data assimilation: An example with an idealized Gulf Stream model. *Journal of Geophysical Research*, *100*(C4), 6777–6793. https://doi.org/10.1029/94jc03084

[6] *Chapter 7_Handout.pdf*. (n.d.). Google Docs. https://drive.google.com/file/d/1A1bNUxxxZ31mGvK

[7] Sadegh, P., & Spall, J. C. (1998). Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, *43*(10), 1480–1484. https://doi.org/10.1109/9.720513