# Analysis of U-Shaped Perturbation Distributions for Simultaneous Perturbation Stochastic Approximation

Eren Aldis

May 14, 2018

## Abstract

Simultaneous Perturbation Stochastic Approximation (SPSA) is known to be an efficient optimization algorithm, especially in high-dimensional settings. SPSA's main power comes from using only two (noisy) loss function evaluations per iteration to estimate the gradient using a symmetric difference approximation. An important part of this gradient approximation is the random perturbation vector, for which sampling from the Bernoulli -1,1 distribution has been proven to be the asymptotically optimal selection. However, this does not imply that the Bernoulli perturbation selection will be optimal for certain finite-sample cases. With regards to this, we will evaluate the performance of valid U-shaped perturbation distributions as a better alternative to Bernoulli in certain optimization settings. We will show that, even in finite-sample cases, the U-shaped distribution will not be able to outperform Bernoulli when used as a perturbation distribution for SPSA.

# 1 Introduction

Consider the problem of finding the input values that minimize the output values of some system. Finding these optimal input values is not always simple. In particular, for high-dimensional inputs, finding these values gets especially difficult. When the true gradient of the system is unknown due to inherent noise in the system, and when the true nature of the system (i.e. the Loss function) is unknown (i.e. no gradient information available), the Stochastic Approximation (SA) algorithm has been proven to achieve the optimal input values through an iterative process. However, in the SA algorithm, at each iteration a gradient approximation is necessary to pick the direction to move in, to hopefully get closer to the optimal input that minimizes the system.

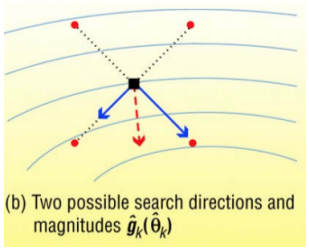
SPSA has been shown to be particularly efficient in approximating the gradient and thus effectively estimating the optimal input. While, it's best alternative, the Finite Difference (FD) method, for a p-dimensional input, uses 2p evaluations of the function or system trying to be minimized at each gradient approximation, the Simultaneous Perturbation method only uses 2. That is, when function evaluations at a given input is expensive or the input is high-dimensional, the SPSA algorithm proves to be significantly more efficient (Spall, 2003).

SPSA approximates the gradient using a symmetric difference by adding and subtracting a random perturbation vector from the value of the input at the current iterate and evaluating the function at for the perturbed values. It has been shown that such an approximation of the gradient is approximately unbiased up to a small bias term, under certain conditions on the perturbation distribution.

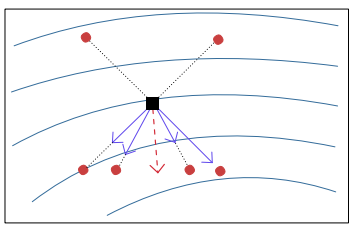
For the simultaneous perturbation, Bernoulli $\{-1, 1\}$ has been shown to be the asymptotically optimal choice for the distribution (Sadegh, Spall 1998). That is, it has been

shown that no other distribution can get a smaller Mean Squared Error in the optimal input estimate asymptotically (i.e. as the number of iterations for SA grow to infinity). However, this does not imply that (1) the Bernoulli perturbation is uniquely asymptotically optimal, and (2) the Bernoulli perturbation is optimal for small sample cases, although it has been observed to perform relatively well overall (Hutchison, 2002). Nonetheless, the superiority of other Non-Bernoulli perturbation distributions has also been investigated in the literature (Hutchison 2002, Cao 2011). Segmented Uniform and Symmetric Double Triangular distributions are among the Non-Bernoulli perturbations that has been evaluated in the literature. However, to our knowledge there has been no significant investigation of U-shaped perturbation distributions in the past.

The main argument for the possible small-sample superiority of these Non-Bernoulli perturbation distributions, has been that while Bernoulli perturbation can only update the input estimates in discrete search directions, the continuous Non-Bernoulli distributions can explore further search directions as shown in Figure 1.




(a) Bernoulli Perturbation Search Directions



(b) Non-Bernoulli Perturbation Search Directions

Figure 1: Search directions for different perturbation distributions

For this end, we will first evaluate the performance of SPSA under U-shaped distribu-

tions from a theoretical standpoint, by analyzing both the asymptotic and small-sample performance of the algorithm. Then, we will study the empirical performance of the algorithm under the U-shaped perturbation distribution by evaluating the algorithm on various noisy loss functions for small and large samples. Finally, we will offer conclusions based on our findings from both the theoretical and empirical analysis.

# 2 Problem Formulation

## 2.1 Minimization Problem

Consider the minimization problem for a loss function $L : \mathbb{R}^p \to \mathbb{R}$ fully dependent on the p-dimensional vector $\theta \in \Theta$, where $\Theta$ is the set of values $\theta$ is allowed to attain. That is, our problem is,

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

This, is equivalent to the root finding problem for the minimizer $\theta^*$,

$$g(\theta) \equiv \frac{dL(\theta)}{d\theta} = 0$$

When we only have access to noisy measurements of $L(\theta)$ in terms of,

$$y(\theta) = L(\theta) + \epsilon$$

for some noise term typically assumed to be distributed $\epsilon \sim N(0, \sigma^2)$, an iterative Kiefer-Wolfowitz type SA algorithm can be used to find the minimizer $\theta^*$ in the form of,

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

for nonnegative, and decreasing step sizes $a_k$ such that $\lim_{k\to\infty} a_k = 0$ and gradient estimate $\hat{g}_k(\cdot)$ evaluated $\hat{\theta}_k$.

## 2.2 SPSA

For the gradient approximation $\hat{g}_k(\hat{\theta}_k)$, we can use a symmetric difference method with a simultaneous perturbation approximation. We can define the simultaneous perturbation approximation as follows. Let $\Delta_k$ denote p-dimensional perturbation vector at the k-th iteration, with $\Delta_{ki}$ to denote the i-th component of the perturbation vector. Then, for a gain sequence $\{c_k\}$ with the same properties the gradient estimator can be given by,

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} \frac{y(\hat{\theta}_k + c_k\Delta_k) - y(\hat{\theta}_k - c_k\Delta_k)}{2c_k\Delta_{k1}} \\ \vdots \\ \frac{y(\hat{\theta}_k + c_k\Delta_k) - y(\hat{\theta}_k - c_k\Delta_k)}{2c_k\Delta_{kp}} \end{bmatrix} \tag{2}$$

Observe here that we only need the same loss function evaluations $y(\hat{\theta}_k \pm c_k\Delta_k)$ to approximate the gradient for each component. That is, we need two loss function evaluations to approximate the full gradient vector, regardless of how large p is.

It has been shown that under certain conditions on the gain sequence, the error terms of the loss function and the perturbation distribution, we have

1. $E[\hat{g}_k(\hat{\theta}_k)|\hat{\theta}_k] \approx g(\hat{\theta}_k)$ up to a bias term of order $O(c_k^3)$
2. as $k \to \infty, \hat{\theta}_k \to \theta^*$ a.s
3. $k^{\frac{b}{2}}(\hat{\theta}_k - \theta^*) \to Z \sim N(\xi^2 d, \rho^2 D)$ for $b > 0$ and $d, D$ independent of the perturbation vector $\Delta_k$. Thus, only $\xi^2, \rho^2$ are dependent on $\Delta_k$.

For the sake of brevity, we will just assume that the conditions on the gain sequence, and

error terms hold for the rest of the paper. The reader can find the detailed conditions in (Spall, 2003). However, we are interested in the conditions pertaining to the perturbation distribution. These conditions along with the definitions of candidate distributions is given in the next section.

## 2.3 Perturbation Distribution Selections for SPSA

The conditions on the perturbation distribution can be given as follows.

(A1) Each component of the perturbation vector $\Delta_{ki}$ must be i.i.d and symmetrically distributed around zero ($E[\Delta_{ki}] = 0$)

(A2) Uniformly finite in magnitude: $|\Delta_{ki}| < \infty$ a.s.

(A3) Finite inverse moments of order $2 + 2\tau$: $\exists\ \tau$ s.t. $E\left[\left|\frac{1}{\Delta_{ki}}\right|^{2+2\tau}\right] < \infty$ a.s.

(A4) There exists $\rho^2, \xi^2$ such that as $k \to \infty$, $E\left[\frac{1}{\Delta_{ki}^2}\right] \to \rho^2$, $E[\Delta_{ki}^2] \to \xi^2$

We can see that condition (A3) eliminates many typical distributions centered around zero (e.g. Normal and Uniform distributions symmetric around 0) from candidacy for the perturbation vector. It has been noted in the past that this is intuitively due to large probability masses around zero (Cao 2011)

A common choice for the perturbation distribution that satisfies all of these conditions has been the discrete Bernoulli that attains {-1, 1} each with probability $p = \frac{1}{2}$. Moreover, it has been shown to yield asymptotically optimal results (Sadegh, Spall 1998). We will refer to the perturbation distributed Bernoulli with parameter $p = 1/2$ as $\tilde{\Delta}_{ki}$.

Another candidate for the perturbation vector, given conditions (A1-4) is the U-shaped distribution. The density function for the U-shaped distribution has the general

form,

$$p_{\Delta_{ki}}(\delta) = \alpha\delta^{2+2\tau}\mathbb{1}_{\{-\beta\leq\delta\leq\beta\}} \tag{3}$$

for $\alpha, \beta > 0$, $\tau \in \mathbb{Z}^+, \tau < \infty$. Along with conditions (A1-4) the density must follow,

(u1) $\int_{-\beta}^{\beta} p_{\Delta_{ki}}(\delta)d\delta = 1$

(u2) $Var(\Delta_{ki}) = Var(\tilde{\Delta}_{ki}) = 1$

The first condition is the typical density condition, that is the area under the curve must sum to one. The second condition is a particular condition, placed to produce a fair comparison between the two distributions. Consequently, from these two conditions we can select the parameters of the U-shaped density. Through integration and simple computation, we get

$$\alpha = \frac{3+2\tau}{2} \cdot \frac{1}{\beta^{3+2\tau}},\ \beta = \left(\frac{5+2\tau}{3+2\tau}\right)^{1/2},\ \tau \text{ positive-integer valued and finite} \tag{4}$$

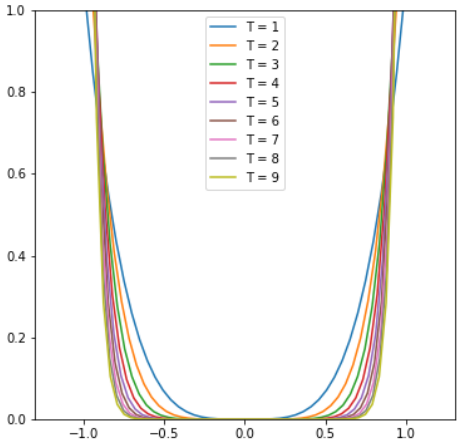
This implies that the values of the density function can be fully determined just by the parameter $\tau$. With regards to this, we will denote $\delta_{ki} \sim U(\tau)$ to say the perturbation is distributed U-shaped with the parameter $\tau$.

Given (3) and (4), it is simple to show that conditions (A1-3) hold. For condition (A4), we can compute the second moment and the inverse moment to be,

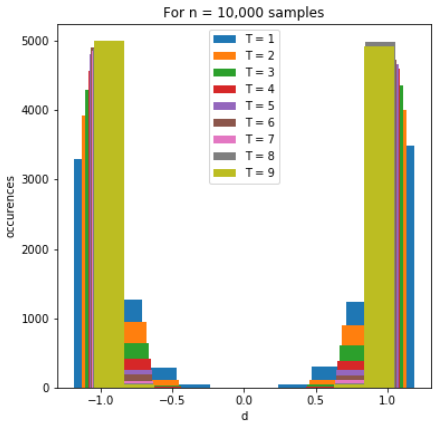
$$E\left[\Delta_{ki}^2\right] = 1 = \xi^2 \text{ (by construction)} \tag{5}$$

$$E\left[\frac{1}{\Delta_{ki}^2}\right] = \frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)} = \rho^2 \tag{6}$$

for all $k,\ i$. We will use these in the next section for asymptotic analysis of SPSA under the U-shaped distribution.




(a) Density of $\Delta_{ki}$ for different $\tau$




(b) $\Delta_{ki}$ sampled from $U(\tau)$ using ICDF

Figure 2: Distribution of $\Delta_{ki} \sim U(\tau)$ for $\tau = \{1, 2, \ldots, 9\}$

For completeness sake we can also give the Cumulative Distribution Function (CDF) from (3), (4) as,

$$P(\Delta_{ki} \leq x) = F(x) = \frac{\alpha}{3+2\tau}\left(x^{3+2\tau} + \beta^{3+2\tau}\right) \tag{7}$$

From (7), we can derive the Inverse Cumulative Distribution Function (ICDF), to be able to sample from the U-shaped distribution, as,

$$F^{-1}(p) = \left(\frac{(3+2\tau)}{\alpha}p - \beta^{3+2\tau}\right)^{\frac{1}{(3+2\tau)}} \tag{8}$$

Using (6), with $p \sim \text{Uniform}(0, 1)$ we can sample from the U-shaped distribution given some $\tau$.

To be able to compare SPSA under the Bernoulli and U-shaped perturbation, as defined above, we still need to pick some $\tau$. However, it is not quite clear how exactly we should pick this $\tau$. Since our goal is to show that SPSA under some U-shaped perturbation can match or surpass the performance of SPSA under Bernoulli, we want to select the $\tau$ that makes this possible in certain cases. In the next section, we will discuss how to pick

$\tau$ with regards to this, from a theoretical standpoint.

# 3 Theoretical Analysis

## 3.1 Asymptotic Analysis

In Section 2.2, we claimed that $k^{b/2}(\hat{\theta}_k - \theta^*)$ asymptotically (as $k \to \infty$) follows $Z \sim N(\xi^2 d, \rho^2 D)$. In order to compare the performance of SPSA under the U-shaped and Bernoulli perturbations, we need to define a metric for comparison. The most obvious option seems to be to compare the Mean Squared Error (MSE) of the estimates of $\hat{\theta}$ under both distributions. However, this is tricky to do in the asymptotic case, as we know from section 2.2 that, since both distributions satisfy the conditions pertaining to the perturbation, as $k \to \infty$, $\hat{\theta} \to \theta^*$. Hence, it does not make sense to directly compare the MSEs asymptotically. For this reason, we instead examine the two perturbation distributions under the asymptotic MSE criterion given in Sadegh, Spall 1998, as follows.

$$MSE = E[\text{tr}\left(ZZ^T\right)] = \rho^2 \text{tr}(D) + \xi^4 d^T d \tag{9}$$

We will denote the asymptotic MSE under the Bernoulli perturbation by $MSE_B$ and MSE under the U-shaped perturbation by $MSE_U$. We have derived $\xi$ and $\rho$ for the U-shaped perturbation on section 2.3 (5), (6). We can also give the $\xi$ and $\rho$ for SPSA under the Bernoulli perturbation as follows, for $\tilde{\Delta}_{ki} \sim \text{Bernoulli}(-1, 1)$

$$E[\tilde{\Delta}_{ki}^2] = 1 \tag{10}$$

$$E\left[\frac{1}{\tilde{\Delta}_{ki}^2}\right] = 1 \tag{11}$$

Hence, from (5), (6) and (10), (11) we can compute the MSE's under both distributions as,

$$\text{MSE}_U = \frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)}\text{tr}(D) + d^T d \tag{12}$$

$$\text{MSE}_B = \text{tr}(D) + d^T d \tag{13}$$

Since it has been shown that the Bernoulli perturbation is asymptotically optimal, our only goal can be is for $\text{MSE}_U$ to match $\text{MSE}_B$. Thus, we should pick the $\tau$ that minimizes $\text{MSE}_U - \text{MSE}_B$. Equivalently, we seek

$$\arg\min_{\tau \in \mathbb{Z}^+} \frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)} - 1 \tag{14}$$

We know that $\forall \tau < \infty$, this value is positive.

For $\tau < \infty$, $MSE_U > MSE_B$. Thus, the Bernoulli perturbation is likely to outperform the U-shaped perturbation asymptotically. We should note that, this result is not too surprising as for large $\tau$ the U-shaped density looks exactly like a Bernoulli distribution (Figure 2). Knowing that the Bernoulli perturbation is asymptotically optimal, as $\tau \to \infty$ since $U(\tau) \to$ Bernoulli in d.r., the U-shaped distribution is only asymptotically optimal then. Therefore, if one is better off using the Bernoulli perturbation in very large

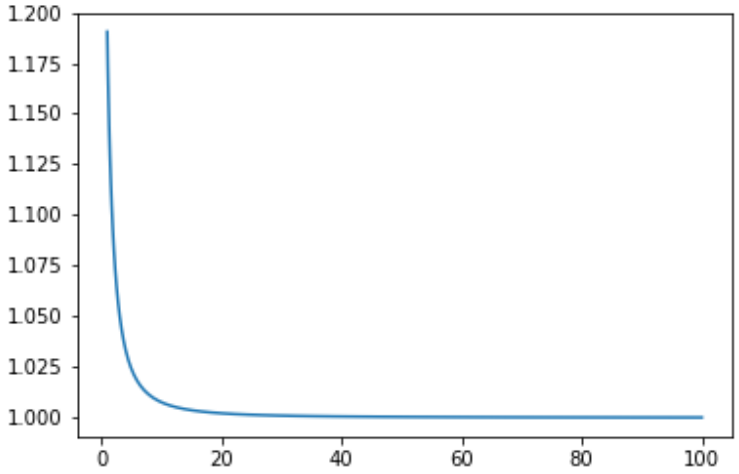



Figure 3: Only for $\tau \to \infty$, $MSE_U$ and $MSE_b$ match

sample cases. However, if one is keen to use the U-shaped perturbation, we could only recommend a choice of a large enough $\tau$ for the parameter such that the algorithm does not go unstable.

Another important result is that $MSE_U$ is significantly larger than $MSE_B$ for smaller values of $\tau$, while as $\tau$ increases, the $MSE_U$ quickly approaches $MSE_B$. The reason is that, the probability mass around zero is significantly larger for small values of $\tau$, in the U-shaped perturbation. Nonetheless, our conjecture is that we should not give up small $\tau$ from candidacy, for the optimal perturbation parameter. In particular, for small samples one would expect the extra variability in the search directions that are gained from a larger probability mass around zero, to yield a better SPSA performance for certain loss functions (Figure 1). With regards to this, in the section forthcoming, we will discuss how to select the parameter $\tau$ in the small sample case.

## 3.2 Small-Sample Analysis

In this section we will derive the MSE of both perturbations for a single-step update. That is we will find the $E\left[(\hat{\theta}_{k+1} - \theta^*)^2 | \hat{\theta}_k\right]$. Although, this comparison does not yield a definitive answer to which perturbation is better for e.g. $k = 10$, the expression will be useful to determine which perturbation has a better single-update performance, and we will be able to determine which perturbation is better at $k = 1$. Of course, determining the better perturbation distribution for $k = 1$ is not sufficient to claim that the perturbation distribution will still outperform the other at other small k, it is surely a starting point and will give us good intuition on the general behavior of SPSA under both distributions for small k. Another reason we compare the single-update MSEs is that they are easy to compute and the unconditional $E\left[(\hat{\theta}_{k+1} - \theta^*)^2\right]$ is unavailable (to our knowledge) in closed form.

The single-update MSE under each distribution will be denoted by $MSE_U^{(S)}$ and $MSE_B^{(S)}$ for finite-sample SPSA performance under U-shaped and Bernoulli perturbations, respectively. Moreover, since the estimates for $\hat{\theta}_k$ are different under the two perturbation distributions, we will denote each by the subscript $B$ or $U$ for Bernoulli and U-shaped, respectively. Similarly, since we can tune the gain sequences using the semi-automatic tuning method in (Spall, 2003 Ch.7), we will denote relevant gain sequences with the same subscripts as for the $\theta$ estimates. Then, for arbitrary $k < \infty$ we can give the finite-sample conditional MSEs as follows. Suppressing the subscript $k$ for both,

$$MSE_U^{(S)}\left(\hat{\theta}_{k+1}|\hat{\theta}_k\right) = \sum_{i=1}^{p}\left(\hat{\theta}_U^{(i)} - \theta^{*(i)}\right)^2 - 2a_U\sum_{i=1}^{p}L_i^{'}(\hat{\theta}_U)\left[\hat{\theta}_U^{(i)} - \theta^{*(i)}\right] + a_U^2\sum_{i=1}^{p}L_i^{'}(\hat{\theta}_U)^2 + \boxed{\frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)}}a_U^2(p-1)\sum_{i=1}^{p}L_i^{'}(\hat{\theta}_U)^2 \quad (15)$$

$$MSE_B^{(S)}\left(\hat{\theta}_{k+1}|\hat{\theta}_k\right) = \sum_{i=1}^{p}\left(\hat{\theta}_B^{(i)} - \theta^{*(i)}\right)^2 - 2a_B\sum_{i=1}^{p}L_i^{'}(\hat{\theta}_B)\left[\hat{\theta}_B^{(i)} - \theta^{*(i)}\right] + a_B^2\sum_{i=1}^{p}L_i^{'}(\hat{\theta}_B)^2 + \boxed{1}a_B^2(p-1)\sum_{i=1}^{p}L_i^{'}(\hat{\theta}_B)^2 \quad (16)$$

where the $\theta^{(i)}$ denotes the i-th component of the vector $\theta$, $\hat{\theta}^{(i)}_{\cdot}$ denotes the i-th component of the vector at the k-th iteration for the U-shaped (U) or Bernoulli (B) distribution, and $L_i'$ denotes the first derivative of the deterministic (non-noisy) loss function with respect to the i-th component.

For SPSA under the U-shaped perturbation to outperform the Bernoulli perturbation at every step, we would want the $\tau$ and $L$ that makes the following criterion true,

$$MSE_U^{(S)} - MSE_B^{(S)} < 0 \quad (17)$$

For, the first update $\hat{\theta}_1$, given some common $\hat{\theta}_0 = \hat{\theta}_U = \hat{\theta}_B$ and equal step-sizes $a_U = a_B$ due to lack of further information, from (15) and (16), we observe that criterion (17) is equivalent to,

$$\frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)} - 1 < 0 \quad (18)$$

We saw this same exact expression in section 3.1. More importantly, we also saw that

$\forall\ \tau > 0$, expression (18) is false. This implies that, even for a single iteration of SPSA, when we assume the step-sizes are equal, SPSA under the U-shaped perturbation will not outperform SPSA under Bernoulli. Only for very large $\tau$ we will see similarity in performance. This implies that our conjecture at the end of Section 3.1 was wrong, and that the extra variability we gain from smaller $\tau$ does not yield better performance even in the single-step update case.

Therefore, our theoretical results imply that both under the asymptotic and the finite-sample case, the Bernoulli perturbation is superior to the U-shaped perturbation. Furthermore, as $\tau \to \infty$ the performances should get closer. We will evaluate this through numerical studies of SPSA ran on particular noisy-loss functions.

However, it is also the case that after $\tau$ at the elbow on Figure 3, we do not achieve significantly bigger gains. That is, selecting the $\tau$ for which we do not see a significant further decrease in MSE, is almost equivalent to selecting a very large $\tau$. (Zhu, Ghodsi 2006) give us the method for which to select this $\tau$ at the elbow.

# 4 Numerical Study

## 4.1 Asymptotic Performance

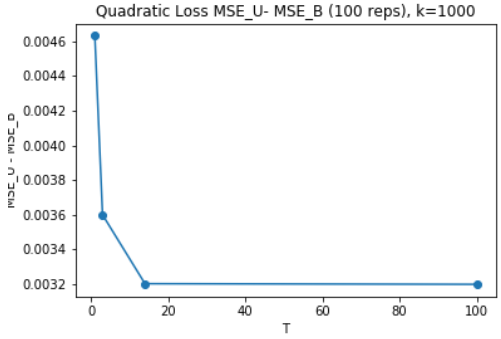
In this section we will evaluate the performance of SPSA under U-shaped and Bernoulli perturbation distributions for two noisy loss functions. We will evaluate the performances by comparing the respective MSEs. To examine the asymptotic case we will use a large sample (k=1000) of iterations, with a $\hat{\theta}_0$ close to $\theta^*$ with the other parameters for gain sequences tuned to emulate asymptotic effects.

First we will evaluate SPSA under U-shaped and Bernoulli perturbations against the

quadratic loss function $L(\theta) = t_1^2 + t_1 t_2 + t_2^2$ with $\sigma = 1$ as in the noisy loss function definition in section 2.1. Thus, $\dim(\theta) = p = 2$ with $\hat{\theta}_0 = [.1,\ .1]^T \approx \theta^*$. Lastly, the gain sequences are in the form of $a_k = \frac{a}{(1+k+A)^{0.602}}$ and $c_k = \frac{c}{(k+1)^{0.101}}$ with $a = 0.017, A = 10$ tuned through the semi-automatic tuning method described in Ch. 7 of (Spall, 2003) and $c = 0.05$ for both perturbations. The results are given below.

| T | MSE for Bernoulli | MSE for U-shaped | P-value |
|---|---|---|---|
| T=1 | 0.0142 | 0.0188 | $<10^{-10}$ |
| T=3 | 0.0142 | 0.0178 | $<10^{-10}$ |
| T=14 | 0.0142 | 0.0174 | $<10^{-10}$ |
| T=100 | 0.0142 | 0.0174 | $<10^{-10}$ |

(a) Table 1: Performance of SPSA for Noisy Quadratic Loss Function



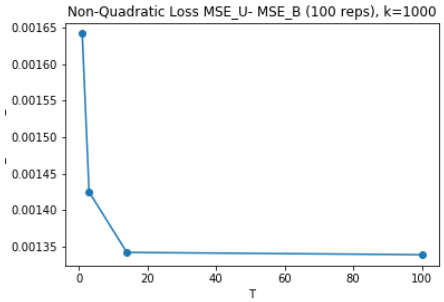

(b) SPSA with U-shaped perturbation approaches the performance of Bernoulli Asymptotically

Figure 4: Asymptotic results for the Quadratic Loss Function for different $\tau$

Now, consider the non-quadratic loss function $L(\theta) = t_1^4 + t_1^2 + t_1 t_2 + t_2^2$ with the same set of parameters $a, c, A, \sigma$ as given above. Results are given below.

| T | MSE for Bernoulli | MSE for U-shaped | P-value |
|---|---|---|---|
| T=1 | 0.0005 | 0.0021 | $<10^{-10}$ |
| T=3 | 0.0005 | 0.0019 | $<10^{-10}$ |
| T=14 | 0.0005 | 0.0018 | $<10^{-10}$ |
| T=100 | 0.0005 | 0.0018 | $<10^{-10}$ |

(a) Table 2: Performance of SPSA for Noisy Non-Quadratic Loss Function




(b) SPSA with U-shaped perturbation approaches the performance of Bernoulli Asymptotically

Figure 5: Asymptotic results for the Non-Quadratic Loss Function for different $\tau$

As seen from Figure 4 and 5, the MSE values under U-shaped perturbations with parameter $\tau$ get closer to SPSA with Bernoulli {-1, 1} perturbation for larger $\tau$. This is as

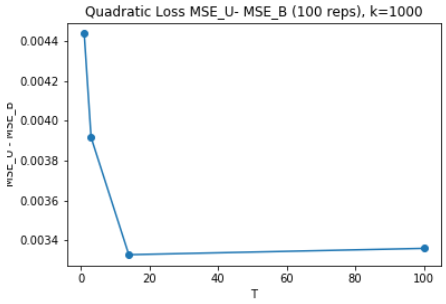
expected, as for larger $\tau$ the U-shaped perturbation converges to a Bernoulli perturbation (in distribution).

## 4.2 Finite-Sample Performance

In this section, we will evaluate the small sample ($k = 10$) performance of SPSA under the two perturbation distributions. We will consider the same loss functions with noise $\sigma = 1$. However, for analysis of small sample performance, we set $A = 1,\ a = 0.07,\ c = 1$ with $\hat{\theta}_0 = [1,\ 1]^T$. The results are given below.

| T | MSE for Bernoulli | MSE for U-shaped | P-value |
|---|---|---|---|
| T=1 | 0.6275 | 1.1658 | $<10^{-10}$ |
| T=3 | 0.6275 | 1.1467 | $<10^{-10}$ |
| T=14 | 0.6275 | 1.1353 | $<10^{-10}$ |
| T=100 | 0.6275 | 1.1458 | $<10^{-10}$ |

(a) Table 3: Performance of SPSA for Noisy Quadratic Loss Function

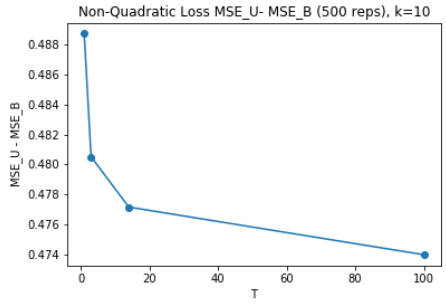



(b) SPSA with U-shaped perturbation approaches the performance of Bernoulli in Small-Samples

Figure 6: Small-Sample results for the Quadratic Loss Function for different $\tau$

| T | MSE for Bernoulli | MSE for U-shaped | P-value |
|---|---|---|---|
| T=1 | 0.8852 | 1.3739 | $<10^{-10}$ |
| T=3 | 0.8852 | 1.3657 | $<10^{-10}$ |
| T=14 | 0.8852 | 1.3623 | $<10^{-10}$ |
| T=100 | 0.8852 | 1.3592 | $<10^{-10}$ |

(a) Table 4: Performance of SPSA for Noisy Non-Quadratic Loss Function




(b) SPSA with U-shaped perturbation approaches the performance of Bernoulli in Small-Samples

Figure 7: Small-Sample results for the Non-Quadratic Loss Function for different $\tau$

We observe that, as expected even in the small-sample cases, the Bernoulli perturbation outperforms the U-shaped perturbation.

# 5 Conclusion

In this paper, we examined the performance of SPSA under the U-shaped perturbation distribution against SPSA under the Bernoulli perturbation. We provided theoretical insight to the inability for the U-shaped perturbation to outperform Bernoulli, both in asymptotic cases and in small-sample cases. We examined how the performance under the U-shaped perturbation improves and gets closer to the performance of Bernoulli perturbations as we increase the $\tau$ that parametrizes the U-distribution. We also gave numerical results that confirm these insights from the theoretical analysis. Thus, we have shown that the U-shaped perturbation regardless of the sample size will be inferior to Bernoulli perturbation, under our assumptions. Nonetheless, it is still possible to construct some function discontinuous only at finite points that has high values for the noisy loss only at points where the Bernoulli perturbation can get to. On the other hand, since the U-shaped perturbation especially for smaller $\tau$ has the ability to search in other directions, it would be able to avoid these discontinuous points with probability one, and thus reach lower loss function values in finite-sample cases.

# 6 References


- Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Wiley, Hoboken, NJ

- Sadegh, P. and Spall, J. C. (1998), "Optimal Random Perturbations for Stochastic Approximation with a Simultaneous Perturbation Gradient Approximation," *IEEE Transactions on Automatic Control*, vol. 43, pp. 1480–1484 (correction to references: vol. 44, 1999, pp. 231–232).

- Hutchison, D. W. (2002), "On an Efficient Distribution of Perturbations for Simulation Optimization using Simultaneous Perturbation Stochastic Approximation," *Proceedings of IASTED International Conference*, 4–6 November 2002, Cambridge, MA, pp. 440–445.

- Cao, X (2011), "Preliminary Results on Non-Bernoulli Distribution of Perturbations for Simultaneous Perturbation Stochastic Approximation," *2011 American Control Conference*, June 29 - July 01, 2011, San Francisco, CA