Analysis of U-Shaped Perturbation Distributions for SPSA

Eren Aldis
ealdis1@jhu.edu

Department of Applied Mathematics and Statistics
The Johns Hopkins University

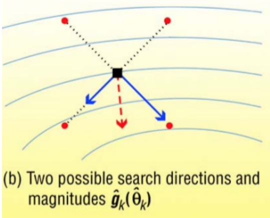
553.763 Stochastic Search and Optimization
Spring 2018

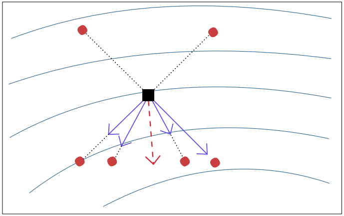
Intuition and Hypotheses

- Bernoulli Perturbation proved to be asymptotically optimal for SPSA (Sadegh, Spall 1998)
- U-shaped perturbation distributions give opportunity to estimate gradient in more search directions

What to investigate?

1. U-shaped distributions should also be able to reach asymptotic optimality.
2. For finite-sample cases U-shaped distributions might work better.





Summary

In this paper, we show

1. The U-shaped perturbation distribution satisfies the conditions for the convergence of $\hat{\theta}_k \to \theta^*$ a.s.
2. Asymptotically, the MSE under the U-shaped perturbation converges to the MSE under the Optimal Bernoulli perturbation
3. For Small-sample cases (k=10), we can derive the conditional MSE as a function of the parameters of the U-shaped distribution. Thus, we can find parameters to minimize the conditional/stepwise MSE.
4. What to pick for the parameters of the U-shaped distribution based on 2, 3
5. Evidence that these results (mostly) hold empirically

Problem Formulation

For a loss function L dependent on the p-dimensional vector $\theta \in \Theta$,

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

which 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 $y(\theta)$ of the loss function, a Kiefer-Wolowitz type SA algorithm can be used in the form of,

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

for nonnegative, decreasing step-size a_k s.t. $\lim_{k \to \infty} a_k = 0$, and gradient estimate $\hat{g}_k$ evaluated at $\hat{\theta}_k$.

SPSA

The noisy loss function evaluated at the k-th iteration is assumed to have the structure,

$$y_k(\hat{\theta}_k) = L(\hat{\theta}_k) + \epsilon, \epsilon \sim N(0, \sigma^2)$$

Then under SPSA, the gradient estimator can be given by,

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

where Δ_{ki} is a random variable symmetric around 0 (mean 0), independent for each i, $1 \leq i \leq p$. We also denote the p-dimensional perturbation vector at the k-th iteration as Δ_k. Also, c_k is another gain sequence.

- Observe that we only need 2 loss function evaluations to compute gradient estimate per iteration, instead of 2p in FD method.

Conditions For Perturbation Distribution

(A1) Δ_{ki} i.i.d and symmetrically distributed around zero ($E[\Delta_{ki}] = 0$)

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

(A3) Finite inverse moments ($2 + 2\tau$): $E\left[\left|\frac{1}{\Delta_{ki}}\right|^{2+2\tau}\right] < \infty$

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

(Refer to the paper for more detailed conditions)

Under these and some other non-perturbation related conditions, we have

- $E[\hat{g}_k(\hat{\theta}_k)|\hat{\theta}_k] \approx g(\hat{\theta}_k)$
- $k^{b/2}(\hat{\theta}_k - \theta^*) \xrightarrow{distr.} Z \sim N(\xi^2 d, \rho^2 D)$

$b > 0$, d, D not dependent on perturbation (See Hill, Fu 1995).

U-Shaped vs. Bernoulli Perturbation Distributions

Under Bernoulli $\{-1, 1\}$ Perturbation,

$$\tilde{\Delta}_{ki} = \begin{cases} 1 & \text{with } p = \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

Under U-shaped Perturbation,

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

for $\alpha, \beta > 0$, $\tau \in \mathbb{Z}^+, \tau < \infty$ where α, β picked such that

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

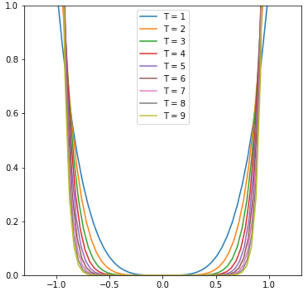
(u2) $\mathrm{Var}(\Delta_{ki}) = \mathrm{Var}(\tilde{\Delta}_{ki}) = 1$ for fair comparison.

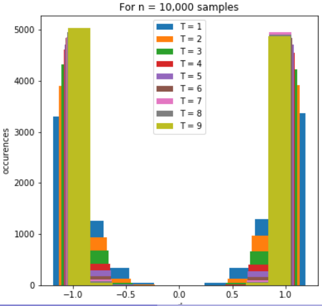
More on U-Shaped Perturbation

From the previous conditions (u1), (u2) we get,

$$\alpha = \frac{3+2\tau}{2} \cdot \frac{1}{\beta^{3+2\tau}}, \beta = \left(\frac{5+2\tau}{3+2\tau}\right)^{1/2}$$

Hence we can denote, $\Delta_{ki} \sim U(\tau)$. We can easily check that Δ_{ki} satisfies (A1-4). Using the inverse CDF, we can sample from the distribution.





Asymptotical Analysis of U-Shaped Perturbation w.r.t τ

Remember that the asymptotic distribution for $k^{b/2}(\hat{\theta}_k - \theta^*)$ follows $Z \sim N(\xi^2 d, \rho^2 D)$. Thus,

$$MSE = E[\mathrm{tr}(ZZ^T)] = \rho^2 \mathrm{tr}(D) + \xi^4 d^T d$$

For $\Delta_{ki} \sim U(\tau)$,

- $E[\Delta_{ki}^2] = 1$ by construction
- $E\left[\frac{1}{\Delta_{ki}^2}\right] = \frac{2\alpha\beta^{1+2\tau}}{1+2\tau} = \frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)}$

For $\tilde{\Delta}_{ki} \sim$ Bernoulli,

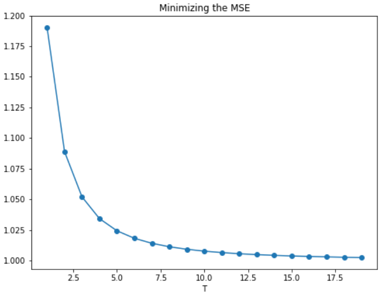
- $E[\Delta_{ki}^2] = 1$
- $E\left[\frac{1}{\Delta_{ki}^2}\right] = 1$

Finding the Asymptotically Optimal τ

$$\mathsf{MSE}_{\Delta_{ki}} = \frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)}\mathrm{tr}(D) + d^T d > \mathrm{tr}(D) + d^T d = \mathsf{MSE}_{\tilde{\Delta}_{ki}}$$

for all $\tau < \infty$.

$$\underset{\tau \in \mathbb{Z}^+}{\arg\min} \frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)}$$



Finite Sample Analysis of MSE

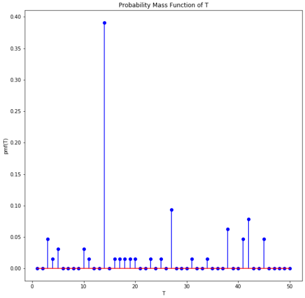
Conditional Mean-Squared Error $\mathrm{MSE}(\hat{\theta}_{k+1}|\hat{\theta}_k)$ can be given as,

$$\sum_{i=1}^{p}\left(\hat{\theta}_k^{(i)}-\theta^{*(i)}\right)^2-2a_k\sum_{i=1}^{p}L_i'(\hat{\theta}_k)\left[\hat{\theta}_k^{(i)}-\theta^{*(i)}\right]+$$
$$a_k^2\sum_{i=1}^{p}L_i'(\hat{\theta}_k)^2+\boxed{\frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)}}a_k^2(p-1)\sum_{i=1}^{p}L_i'(\hat{\theta}_k)^2$$

- We have seen the boxed term before.
- Consider $\mathrm{MSE}(\hat{\theta}_1|\hat{\theta}_0)$. To minimize this, we need to minimize the boxed term.
- Same conclusion as before for the finite sample case.

How to pick τ?

- Asymptotically or Large Sample: Pick the largest τ such that the algorithm does not go unstable.
- Finite Sample or Small Sample: Pick the "elbow" (Zhu, Ghodsi 2006) so that we do not lose all variability in search directions, yet it is still optimally efficient. Depending on the number of candidate τ considered, we get $\tau \in \{3, 14, 27, 38\}$

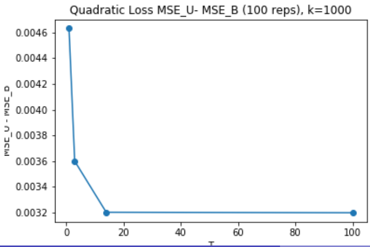


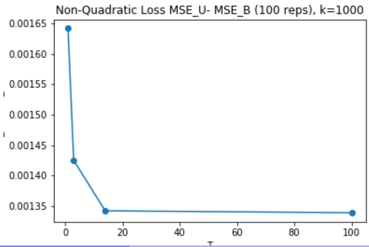
Empirical Analysis: Asymptotic Performance

For the noisy ($\sigma = 1$) Quadratic (Left) and Non-quadratic (Right) Loss functions (p=2), $\hat{\theta}_0 = [.1, .1]^T$, to emulate asymptotic effects $A = 10$, $c = 0.05$, $a = 0.017$ for both, (k=1000)

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	$<10^{-10}$

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	$<10^{-10}$



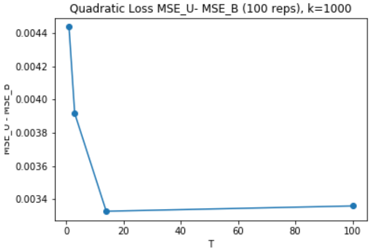


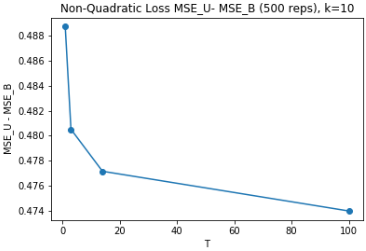
Empirical Analysis: Small Sample Performance

For the noisy ($\sigma = 1$) Quadratic (Left) and Non-quadratic (Right) Loss functions (p=2), A=1, $a \approx 0.07$, k=10, $c = 1$,$\hat{\theta}_0 = [1, 1]^T$

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	$<10^{-10}$

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	$<10^{-10}$





Back to the Finite-Sample Analysis

Under equal gain sequences a_k for both distributions, the MSE under the Bernoulli Distribution can be derived as

$$\sum_{i=1}^{p}\left(\hat{\theta}_k^{(i)} - \theta^{*(i)}\right)^2 - 2a_k \sum_{i=1}^{p} L_i^{'}(\hat{\theta}_k)\left[\hat{\theta}_k^{(i)} - \theta^{*(i)}\right] +$$

$$a_k^2 \sum_{i=1}^{p} L_i^{'}(\hat{\theta}_k)^2 + \boxed{1} a_k^2 (p-1) \sum_{i=1}^{p} L_i^{'}(\hat{\theta}_k)^2$$

Hence, for MSE under the U-shaped distribution to beat the MSE under the Bernoulli distribution, we want

$$\frac{(3+2\tau)^2}{(1+2\tau)(5+2\tau)} < 1$$

But this is not possible, for all $\tau < \infty$ the term on the left-hand side is greater than term on the right due to Schwarz Inequality.

Conclusions

- Asymptotically, the U-shaped distribution attains similar performance as Bernoulli.
- For small sample cases, the U-shaped perturbation does not beat the Bernoulli perturbation.
- Extra variability in the search directions under the U-shaped distribution does not yield better performance.