2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Preliminary Results on Non-Bernoulli Distribution of Perturbations for Simultaneous Perturbation Stochastic Approximation

Xumeng Cao
Department of Applied Mathematics and Statistics, Johns Hopkins University, MD, USA
xcao7@jhu.edu

***Abstract*—Simultaneous perturbation stochastic approximation (SPSA) has proven to be an efficient algorithm for recursive optimization. SPSA uses a centered difference approximation to the gradient based on only two function evaluations regardless of the dimension of the problem. Typically, the Bernoulli ±1 distribution is used for perturbation vectors and theory has been established to prove the asymptotic optimality of this distribution. However, efficiency of the Bernoulli distribution may not be guaranteed for small-samples. In this paper, we investigate the performance of segmented uniform distribution for perturbation vectors. For small-samples, we show that the Bernoulli distribution may not be the best for a certain choice of parameters**.

1. INTRODUCTON

Simultaneous perturbation stochastic approximation (SPSA) has proven to be an efficient stochastic approximation approach (see [7, 8 and 10]). It has wide applications in engineering such as signal processing, system identification and parameter estimation (see www.jhuapl.edu/SPSA and [2, 9]).

Typically, the Bernoulli ±1 distribution is used for perturbation vectors in SPSA. It is easy to implement and has been proven asymptotically optimal (see [5]). But one might be curious if this optimality holds when only small-sample is allowed. This is common situation in practice when it is expensive to evaluate system performances. Thus, we wonder if non-Bernoulli distributions will outperform the Bernoulli ±1 as a distribution for perturbation vectors when the number of function evaluations is small.

Discussion and research on non-Bernoulli perturbation distribution has been found in the literature, see [1, 3]. The application of non-Bernoulli perturbations in small samples has been considered in [4] and [9, Chap. 7].

2. METHODOLOGY

2.1 Problem Formulation

Let $\boldsymbol{\theta} \in \Theta \subseteq R^P$ denote a vector of parameters of interest. Let $L(\boldsymbol{\theta})$ be the loss function which is observed at the presence of noise: $y(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \varepsilon$, where ε is i.i.d noise, with mean zero and variance σ^2. The problem is to

$$\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}). \tag{1}$$

The stochastic optimization algorithm for solving (1) is given by the following iterative scheme:

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) \tag{2}$$

where $\hat{\boldsymbol{\theta}}_k$ is the estimate of $\boldsymbol{\theta}$ at iteration k and $\hat{\boldsymbol{g}}_k(\bullet) \in R^p$ represents an estimate of the gradient of L at iteration k. The step-size sequence $\{a_k\}$ is nonnegative, decreasing, and converging to zero.

2.2 Perturbation Distribution for SPSA

SPSA uses a simultaneous perturbation to estimate the gradient. Let $\boldsymbol{\Delta}_k = [\Delta_{k1}, \ldots, \Delta_{kp}]^T$ be a vector of p independent random variables at iteration k. Let c_k be a sequence of positive scalars. The standard simultaneous perturbation form for the gradient estimator is as follows:

$$\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) = \frac{y(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - y(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k} \times [\Delta_{k1}^{-1}, \ldots, \Delta_{kp}^{-1}]^T. \tag{3}$$

Valid distributions for $\boldsymbol{\Delta}_k$ include Bernoulli ±1 (B), segmented uniform (SU), U-shape distribution and many others (see [9], page 185). In the discussion below, we compare SU with B. The domain of SU is chosen as $(-(19+3\sqrt{13})/20, -(19-3\sqrt{13})/20) \cup ((19-3\sqrt{13})/20, (19+3\sqrt{13})/20)$ such that the mean and variance are the same as those of B. In our analysis, the sequences $\{a_k\}$ and $\{c_k\}$ take standard forms: $a_k = a/(k+2)^{0.602}$, $c_k = c/(k+1)^{0.101}$, where a and c are predetermined constants.

3. THEORETICAL ANALYSIS

In this section, we provide conditions under which SU outperforms B. To specifically analyze the development of the algorithm, we consider the extreme version of small sample where only one iteration takes place in SPSA, that is, $k = 1$. For larger k, the analysis is too complicated and is not included in this paper. Due to page limit, we present all results without providing detailed analysis. For interested readers, please contact the author for more information.

As a criterion to compare the performance of SU and B, mean squared error (MSE) between $\hat{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}^*$ is used. If we assume that the loss function L has continuous third derivatives, the difference in MSE $E\left(\left\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\right\|^2\right)$ under two distributions is computed as follows:

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

$$= \sum_{i=1}^{p}\left(a_{0S}^2\left(L_i^2 + \sum_{j\neq i}\left(100L_j^2/61\right)\right) - 2(a_{0S} - a_{0B})L_i(\hat{\theta}_{0i} - \theta_i^*)\right)$$

$$- p\left(a_{0B}^2\left(\sum_{i=1}^{p} L_i^2 + 0.5\sigma^2/c_{0B}^2\right) - 50a_{0S}^2\sigma^2/61c_{0S}^2\right) + O(c_0^2), \tag{4}$$

where L_i denotes the first derivative of L with respect to the ith component of $\boldsymbol{\theta}$. All derivatives are evaluated at $\hat{\boldsymbol{\theta}}_0$, same in the analysis below. Subscripts S, B denote SU and the Bernoulli distribution, respectively, same as in the context to follow; $\hat{\theta}_{0i}$ and θ_i^* denote the ith component of $\hat{\boldsymbol{\theta}}_0$ and $\boldsymbol{\theta}^*$,

respectively. The $O(c_0^2)$ term is due to the higher order Taylor expansion.

Furthermore, if we assume $\left|L_{ijk}(\bullet)\right| \le M$ for all i, j, k, where M is a constant and L_{ijk} denotes third derivatives, an upper bound U for the $O(c_0^2)$ term in (4) is:

$$U = (4a_{0S}c_{0S}^2 + a_{0B}c_{0B}^2)M\left(\sum_{i=1}^{p}\left|\hat{\theta}_{0i} - \theta_i^*\right|\right) \times (p-1)^2 + \frac{1}{20}a_{0S}^2 c_{0S}^4 M^2 p^7 + \frac{1}{3}(a_{0S}^2 c_{0S}^3 + a_{0B}^2 c_{0B}^3)Mp^5 \max_i L_i(\hat{\boldsymbol{\theta}}_0). \quad (5)$$

We now represent a theorem and its corollary. The proofs are immediate based on the expressions above.

Theorem 1

Consider loss function with continuous third derivatives. For one iteration of SPSA, the SU distribution produces a smaller MSE between $\hat{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}^*$ than B if the starting point and the relevant coefficients (a_0, c_0, σ^2) are chosen in such a way that the right hand side of (4) is negative.

If in addition, magnitude of third derivatives of L has upper bound M, a sufficient condition for the superiority of SU is that the upper bound of the expression in (4), which could be derived by (5), is negative.

If L is quadratic, the higher order term in (4) vanishes. Moreover, if $p = 2$, expression in (4) can be simplified.

Corollary 1

For a quadratic loss function with $p = 2$, SU produces a smaller MSE between $\hat{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}^*$ than B if the following expression is negative:

$$\begin{aligned} &E_S\left(\left\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\right\|^2\right) - E_B\left(\left\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\right\|^2\right) \\ &= a_{0S}^2(L_1^2 + L_2^2 + 100L_1^2/61 + 100L_2^2/61) - 2(a_{0S} - a_{0B})L_1(\hat{\theta}_{01} - \theta_1^*) \\ &\quad - 2(a_{0S} - a_{0B})L_2(\hat{\theta}_{02} - \theta_2^*) - 2a_{0B}^2(L_1^2 + L_2^2 + 0.5\sigma^2/c_{0B}^2) \\ &\quad + 100a_{0S}^2\sigma^2/61c_{0S}^2, \end{aligned} \quad (6)$$

4. NUMERICAL EXAMPLE

Consider the quadratic loss function $L(\boldsymbol{\theta}) = t_1^2 - t_1 t_2 + t_2^2$, where $\boldsymbol{\theta} = [t_1, t_2]^T$, $\sigma^2 = 1, \hat{\theta}_0 = [0.3, 0.3]^T$, $a_{0S} = 0.0011$, $a_{0B} = 0.01252$, $c_{0S} = c_{0B} = 0.1$. The parameters are chosen according to standard tuning process (see [9, Section 7.5]). The right hand side of (6) is calculated as −0.0114, which satisfies the condition of Corollary 1, indicating SU is superior to B for $k = 1$. This is consistent with our numerical simulation summarized in Table 1.

Table 1: Empirical MSE values

	B	SU	*P*- value
k=1	0.1913	0.1798	$<10^{-10}$
k=5	0.2094	0.1796	$<10^{-10}$
k=10	0.1890	0.1786	$<10^{-10}$
k=1000	0.0421	0.1403	$>1-10^{-10}$

In Table 1, each MSE is approximated by averaging over 10^6 independent runs. *P*-values are derived from *t*-tests for comparing the MSEs of B and SU. For $k = 1$, the difference between MSEs under SU and B is −0.0115 (as compared to theoretical value of −0.0114), with corresponding *P*-value being almost 0, showing a strong indication that SU is preferred to B for $k = 1$.

We also notice that the advantage of SU holds for $k = 5$ and $k = 10$ in this example. In fact, the better performance of SU for $k > 1$ has been observed in other examples as well (e.g., [4] and [9, Exercise 7.7]). Even though this paper only provides theoretical foundation for $k = 1$ case, it might be possible to generalize the theory to $k > 1$ provided that k is not too large a number.

5. CONCLUSIONS

We provide conditions under which segmented uniform distribution outperforms the Bernoulli distribution for one iteration of SPSA. Furthermore, numerical examples indicate that we may generalize the above conclusion to other small sample sizes as well, but we have not yet pursued that avenue of research.

Furthermore, advantage of segmented uniform distribution has also been observed in numerical study where constrained optimization problem is considered. A further line of research might be to investigate the superiority of SU in dealing with constrained problems. Non-Bernoulli perturbations provide greater flexibility in the search direction and consequently provide an improved ability to avoid potential entrapments due to constraints. In future work we intend to apply non-Bernoulli SPSA to the constrained optimization problem in Spall and Hill [6].

REFRENCES

[1] Bhatnagar, S., Fu, M.C., Marcus, S.I., and Wang, I.J. (2003), "Two-Timescale Simultaneous Perturbation Stochastic Approximation Using Deterministic Perturbation Sequences," *ACM Transactions on Modeling and Computer Simulation*, vol. 13, pp. 180–209.

[2] Bhatnagar, S. (2011), "Simultaneous Perturbation and Finite Difference Methods," in *Wiley Encyclopedia of Operations Research and Management Science* (J. Cochran, ed.), vol. 7, pp. 4969−4991, Wiley, Hoboken, NJ

[3] 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.

[4] Maeda, Y. and De Figueiredo, R.J. P. (1997), "Learning Rules for Neuro-Controller via Simultaneous Perturbation," *IEEE Transactions on Neural Networks*, vol. 8, pp. 1119–1130.

[5] Sadegh, P., 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).

[6] Spall, J. C. and Hill, S. D. (1990), "Least-Informative Bayesian Prior Distributions for Finite Samples Based on Information Theory," *IEEE Transactions on Automatic Control*, vol. 35, no. 5, pp. 580–583.

[7] Spall, J. C. (1992), "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Transactions on Automatic Control*, vol. 37, No. 3, pp. 332–341.

[8] Spall, J. C. (1998), "An Overview of the Simultaneous Perturbation Method for Efficient Optimization," *Johns Hopkins APL Technical Digest*, vol. 19(4), pp. 482–492.

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

[10] Spall, J.C. (2009), "Feedback and Weighting Mechanisms for Improving Jacobian Estimates in the Adaptive Simultaneous Perturbation Algorithm," *IEEE Transactions on Automatic Control*, vol. 54(6), pp. 1216-1229.