Proceedings of the IASTED International Conference
APPLIED MODELLING AND SIMULATION
November 4-6, 2002, Cambridge, MA, USA

On an Efficient Distribution of Perturbations for Simulation Optimization using Simultaneous Perturbation Stochastic Approximation

David W. Hutchison
Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, MD 21218
David.Hutchison@jhu.edu

Abstract

Stochastic approximation as a method of simulation optimization is well-studied and numerous practical applications exist. One approach, simultaneous perturbation stochastic approximation (SPSA), has proven to be an efficient algorithm for such purposes. SPSA uses a centered difference approximation to the gradient based on two function evaluations regardless of the dimension of the problem. It accomplishes this task by randomizing the directions in which the differences are calculated in each dimension. Typically Bernoulli variables mapped to {–1, 1} are used in the randomization and this distribution is known to be asymptotically most efficient, but the question of best distribution remains open for small-sample approximations. As part of a general theory of small-sample stochastic approximation, the author has studied alternative distributions for the perturbations used to compute the SPSA estimate of the gradient. This paper presents results from that investigation, as well as some insights to parameter selection for the SPSA algorithm.

Key Words: simulation optimization, SPSA, stochastic approximation, iterative algorithms, stochastic gradient.

1. Introduction

Consider the problem of optimizing some performance measure of a stochastic system [1]. If the decision variables are continuous and the solution space may be assumed closed and convex, the problem lends itself to solution with a gradient-based optimization method, that is, to find the zero of the gradient of the performance measure. Even in cases where these conditions do not hold, gradient methods may prove useful.

When the system dynamics are unknown, the usual methods to compute this gradient, such as perturbation analysis or likelihood ratio estimation, are not available [2]. Stochastic approximation techniques overcome these difficulties by estimating the gradient of the performance measure of interest using (perhaps noisy) measurements of the performance measure itself [3]. See Jacobson and Schruber [4] for a general overview of techniques for simulation optimization.

2. Problem Formulation

Let $\theta \in \Theta \subseteq \mathbb{R}^p$ denote a vector of input parameters. Let $Q(\theta, \omega)$ denote the observed performance as a function of θ and the stochastic effects ω. The function of interest is $L(\theta) = E[Q(\theta, \omega)]$, the expected system performance at θ. The problem is then

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

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

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

where $\hat{g}_k(\hat{\theta}_k) \in \mathbb{R}^p$ represents an estimate of the gradient of L at $\hat{\theta}_k$. The step-size sequence $\{a_k\}$ is nonnegative, decreasing, and converges to zero. If $\Theta \neq \mathbb{R}^p$ then the problem is constrained and projection or penalty methods may be applied. The generic iterative form of (2) is analogous to the familiar steepest descent algorithm for deterministic problems.

2.1 Stochastic Approximation

First introduced by Robbins and Monro [5] (measurements of $g(\theta)$) and Kiefer and Wolfowitz [6] (measurements of $L(\theta)$), the method has been the subject of considerable research, expanding its applicability and relevance (see, e.g., [1], [7], [8]).

It is most efficient to use direct estimates of the gradient, but in many cases this may not be feasible. In this case gradient estimates based on (noisy) measurements of the performance measure itself must be made. In this paper we consider the case where the form of $L(\theta)$ and $g(\theta)$ are unknown, and only measurements of $L(\theta)$ are available. The estimate $\hat{\theta}_k$ converges to the optimal value θ^*, under suitable conditions on the loss function and gradient (see, e.g., [9] and [10]).

A common though computationally inefficient method to estimate the gradient from observed measurements is by

finite-differences. Symmetric (centered) finite-differences require $2p$ function evaluations, where p is the dimension of the solution space. If p is large and the function evaluations difficult or time-consuming, the computational effort could be substantial since the estimate must be computed at each iteration in (2).

2.2 Simultaneous Perturbation

Simultaneous perturbation stochastic approximation (SPSA) uses a method of simultaneous perturbation to estimate the gradient [11], [12]. The efficiency of this method is that it requires only two function evaluations to estimate the gradient at each iteration, regardless of the dimension of θ. Thus the major advantage of SPSA is the reduction in computations required to achieve an optimal solution by reducing the number of required simulation experiments. The theoretical basis for SPSA was developed by Spall [14] and [11] and expanded in subsequent work (see [2], [15], [10], and references therein).

The applicability of SPSA to simulation optimization has been shown for nonlinear control problems using neural networks [12], single-server queueing discrete-event systems [2], air traffic control [13], and many other problems.

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

$$\Delta_k = \begin{bmatrix} \Delta_{k_1} & \Delta_{k_2} & \cdots & \Delta_{k_p} \end{bmatrix}^{\mathrm{T}} \tag{3}$$

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

$$\begin{aligned} y(\hat{\theta}_k + c_k\Delta_k) &= L(\hat{\theta}_k + c_k\Delta_k) + \epsilon_k^+ \\ y(\hat{\theta}_k - c_k\Delta_k) &= L(\hat{\theta}_k + c_k\Delta_k) + \epsilon_k^- \end{aligned} \tag{4}$$

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

The standard simultaneous perturbation form for the gradient estimator is shown in (5).

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} \dfrac{y(\hat{\theta}_k + c_k\Delta_{k_1}) - y(\hat{\theta}_k - c_k\Delta_{k_1})}{2c_k\Delta_{k_1}} \\ \vdots \\ \dfrac{y(\hat{\theta}_k + c_k\Delta_{k_p}) - y(\hat{\theta}_k - c_k\Delta_{k_p})}{2c_k\Delta_{k_p}} \end{bmatrix} \tag{5}$$

Note that $\hat{g}_k(\hat{\theta}_k)$ requires only two measurements of $L(\theta)$ (those in equation (4)) and is independent of the dimension of the solution space, p. Under appropriate regularity conditions we have estimator $\hat{g}_k(\hat{\theta}_k)$ nearly unbiased with

$$E\left[\hat{g}_k(\hat{\theta}_k)\middle|\hat{\theta}_k\right] = g(\hat{\theta}_k) + O(c_k^2) \quad \text{a.s.} \tag{6}$$

where $c_k \to 0$. The iteration converges $\hat{\theta}_k \to \theta^*$ almost surely as $k \to \infty$. See [11] for details.

Common selections for the step size sequences a_k and c_k are shown below.

$$a_k = ak^{-\alpha} \qquad c_k = ck^{-\gamma} \tag{7}$$

where a and c are positive constants and the exponents satisfy $\frac{1}{2} < \alpha \leq 1$ and $1/12 < \gamma \leq 1/6$, with the upper bounds theoretically optimal. This leaves the choice of suitable constants a and c to regulate algorithm performance.

2.3 Perturbation Distribution for SPSA

As discussed above, the perturbations Δ_k in the gradient estimate (5) are based on Bernoulli random variables on $\{-1, 1\}$. In fact, the requirements are merely that the Δ_{ki} must be independent and symmetrically distributed about zero with finite absolute inverse moments $E[|\Delta_{ki}|^{-1}]$ for all k, i. The Bernoulli is just one distribution for Δ_{ki} that satisfies these conditions.

It has been shown that one cannot do better than this distribution in the asymptotic case [15], but less is known about the best distribution for small-sample approximations. Some numerical results seem to show better performance on some problems with non-Bernoulli distributions.

The author examined a wide range of candidate non-Bernoulli distributions and compared algorithm performance. The performance of three such alternative distributions is reported here: a split uniform distribution, an inverse split uniform distribution, and a symmetric double triangular distribution (referred to as candidate distributions in the following).

The $\{-1, 1\}$-Bernoulli distribution has variance and absolute first moment (mean magnitude) both equal to one. It is the only qualified distribution with these qualities. We conjecture that these characteristics are necessary conditions for optimal performance of the SPSA algorithm, given optimal step size parameters. Variations in mean magnitude can be addressed by scaling the gradient step size (c), so for comparisons, candidate distributions should have the same variance as the $\{-1, 1\}$-Bernoulli. Then differences in performance could be attributed to differences in the nature of variability in that distribution.

Distribution	Mean	Mean Magnitude	Variance
{-1,1}-Bernoulli	0	1	1
Split Uniform	0	$\frac{1}{2}(a+b)$	$\frac{1}{3}(a^2+ab+b^2)$
Inverse Split Uniform	0	$\frac{ab}{b-a}\log\frac{b}{a}$	ab
Sym. Double Triangular	0	$\frac{1}{3}(a+b+c)$	$\frac{1}{6}(a^2+b^2+c^2+ab+ac+bc)$

Table 1 – Characteristics of the Perturbation Distributions

To ensure consistency in the comparison, we normalized the candidate distributions so that their variances were one and their mean magnitudes were close to one, but not so close that the essential character of the distributions were lost. The probability density functions of these distributions are given at right. The characteristics of each distribution are given in Table 1.

2.4 Testing Procedure

The SPSA algorithm with each distribution for the perturbations was applied to 34 functions from Moré's suite of optimization problems [17]. The initial points recommended in Moré were used for each function. The functions values were obscured with normally distributed errors with mean zero and a variance of one. We then used these noisy function values to calculate a simultaneous perturbation gradient approximation.

For nearly all of the functions, errors of this magnitude are insignificant away from the minimum. However, most functions in the optimization suite have minimums at or near zero, where $N(0, 1)$ errors are quite significant. This situation is further complicated by the fact that many functions are extremely flat near the minimum as well. The result was a demanding examination of the SPSA algorithm offering ample opportunity to test alternative perturbation distributions.

We used the step sequences in (7). The step size parameters of the SPSA algorithm (that is, a and c) were optimized for each distribution and each function by random search. The procedure to optimize the step parameters used 20,000 iterations of a directed random search algorithm. In the directed random search (sometimes called a localized random search, see [10], p. 45), new trial values are generated near the location of the current best value. The algorithm accepts the input parameters as the current optimal values if they produce results that are better than the best yet obtained, otherwise they are rejected.

$$f_{SU}(x;a,b)=\begin{cases}\dfrac{1}{2(b-a)} & -b\le x\le -a \text{ or } a\le x\le b\\ 0 & \text{otherwise}\end{cases}$$

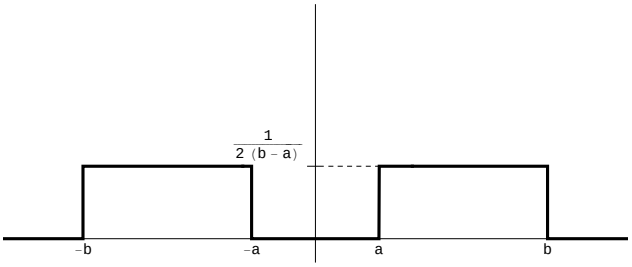

Figure 1 – The Split Uniform Distribution

$$f_{ISU}(x;a,b)=\begin{cases}\dfrac{ab}{2(b-a)x^2} & -b\le x\le -a \text{ or } a\le x\le b\\ 0 & \text{otherwise}\end{cases}$$

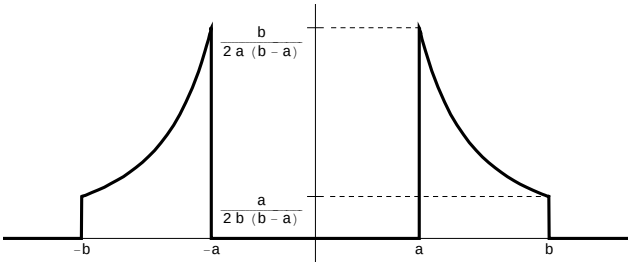

Figure 2 - The Inverse Split Uniform Distribution

$$f_{SDT}(x;a,b,c)=\begin{cases}\dfrac{x+c}{(c-a)(c-b)} & -c\le x\le -b\\ \dfrac{x+a}{(c-a)(a-b)} & -b\le x\le -a\\ \dfrac{x-a}{(c-a)(b-a)} & a\le x\le b\\ \dfrac{x-c}{(c-a)(b-c)} & b\le x\le c\\ 0 & \text{otherwise}\end{cases}$$

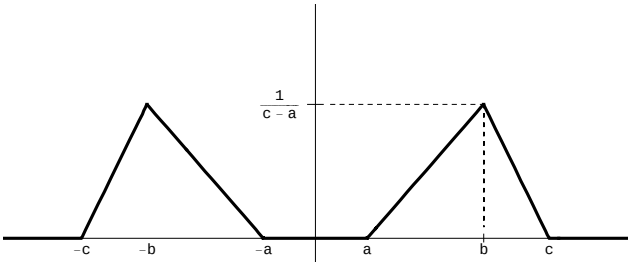

Figure 3 - The Symmetric Double Triangular Distribution

This method is somewhat more sophisticated than simple random search, and generally more computationally efficient in that it uses information from previous iterations. For more information on random search methods, see Solis and Wets [18].

For each iteration of the random search we executed fifty Monte Carlo trials of the SPSA algorithm, and then accepted or rejected the parameter values based on the average of these fifty trials. The theoretically optimal values for α and γ were used.

The SPSA algorithm in the procedure outlined above was run for stopping times of n = 10, 100, and 1000 iterations to determine whether any one distribution outperformed the others over small, moderate, and large iteration domains.

Common random numbers (CRN) were used to minimize variance. With CRN, the sequences of function values generated by the iteration differ only as a result of how the SPSA algorithm processes the random numbers in a different way. In this evaluation, the sequence of CRN were used to generate random perturbations from the appropriate distribution. This method allows the use of matched pairs testing to determine the significance of differences in the minimum values observed. Matched pairs testing generally leads to sharper analysis.

3. Empirical Results

This section describes some of the results obtained from a systematic application of SPSA to the Moré suite of optimization problems. The base algorithm used Bernoulli perturbations. Alternatives included split uniform, inverse split uniform, and symmetric double triangular distributions for the perturbations.

The exact distributions used in this analysis were

$$f_{SU}(x; 0.4092, 1.4908)$$
$$f_{ISU}(x; 0.6667, 1.5000)$$
$$f_{SDT}(x; 0.3333, 1.0781, 1.5000)$$

Parameter values for the distributions were chosen to give a variance of one and a mean magnitude close to one. The mean magnitude for each distribution is shown in Table 2.

Split Uniform	0.95
Inverse Split Uniform	0.973116
Symmetric Double Triangular	0.970484.

Table 2 - Mean Magnitudes of the Candidate Distributions

While many of the functions in the Moré suite do not satisfy the convergence criteria for stochastic approximation, in practice well-chosen parameter values for a and c can yield good results. However, there were four functions in this suite for which we were not able to get SPSA to converge for any distribution.[1] These functions are particularly ill-behaved even in the deterministic case, and the addition of noise to the function was sufficient to obscure any underlying trend information and frustrate the stochastic approximation procedure. The algorithm was very sensitive to the selection of the step size parameters a and c for these functions, and our assumption is that with more effort suitable parameter values could be found.

We highlight only two functions of the Moré optimization suite here. Of the remaining functions, none showed results dramatically different from those presented here. In most cases the mean minimum value found at the end of the procedure was lowest for Bernoulli perturbations. In the cases where it did not return the lowest mean minimum, the difference was not statistically significant.

The results for the two-dimensional Rosenbrock function are shown in Table 3. The Rosenbrock function is difficult for optimization procedures because the surface is flat bottomed and curved. As a result, iterative procedures progress slowly to the unique minimum of zero at (1, 1).

Bernoulli	$n = 10$	$n = 100$	$n = 1000$
a	3.574×10^{-3}	5.237×10^{-3}	7.891×10^{-3}
c	2.044×10^{-1}	1.031×10^{-1}	0.845×10^{-1}
$\bar{L}(\theta_n)$	1.829	1.689	1.155
s^2	1.406	1.210	0.932

SU	$n = 10$	$n = 100$	$n = 1000$
a	7.911×10^{-3}	8.134×10^{-3}	9.560×10^{-3}
c	3.482×10^{-1}	2.997×10^{-1}	1.458×10^{-1}
$\bar{L}(\theta_n)$	2.090	1.801	1.545
s^2	1.406	1.276	1.049

ISU	$n = 10$	$n = 100$	$n = 1000$
a	6.818×10^{-3}	5.112×10^{-3}	6.621×10^{-3}
c	3.455×10^{-1}	1.804×10^{-1}	1.002×10^{-1}
$\bar{L}(\theta_n)$	2.296	1.947	1.572
s^2	1.423	1.406	0.899

SDT	$n = 10$	$n = 100$	$n = 1000$
a	3.805×10^{-3}	5.237×10^{-3}	7.921×10^{-3}
c	2.100×10^{-1}	1.031×10^{-1}	0.914×10^{-1}
$\bar{L}(\theta_n)$	2.121	1.791	1.758
s^2	1.806	1.741	1.121

Table 3 - Two-dimensional Rosenbrock Function (#1)

[1] These four were the Brown badly scaled function (#4), Meyer function (#10), Brown and Dennis function (#16), and the trigonometric function (#26).

Despite these difficulties, the SPSA algorithm performed well with each of the perturbation distributions. For this function the {–1, 1}-Bernoulli distribution achieved significantly better minimums than any of the candidate distributions for any stopping time. For example, the *p*-values for the $n = 10$ case are 0.176, 0.243, and 0.219 for SU, ISU, and SDT, respectively. *P*-values for the $n = 100$ and $n = 1000$ cases are similar.

Results for the Kowalik and Osborne function (Table 4) show a more favorable outcome for SU and SDT for the short ($n = 10$) and moderate ($n = 100$) stopping times. However, these results are not significant (*p*-values 0.031 and 0.022, respectively, for $n = 10$; similar for $n = 100$). Moreover, this apparent advantage disappears in the case $n = 1000$, where the {–1, 1}-Bernoulli is significantly better than any of the tested candidate distributions.

Bernoulli	$n = 10$	$n = 100$	$n = 1000$
a	1.326×10^{-3}	2.992×10^{-3}	5.192×10^{-3}
c	1.252×10^{-1}	3.127×10^{-1}	5.731×10^{-1}
$\bar{L}(\theta_n)$	4.817×10^{-3}	4.464×10^{-3}	3.105×10^{-3}
s^2	0.881×10^{-4}	0.524×10^{-4}	0.223×10^{-4}

SU	$n = 10$	$n = 100$	$n = 1000$
a	1.939×10^{-3}	3.444×10^{-3}	6.023×10^{-3}
c	2.056×10^{-1}	4.058×10^{-1}	5.998×10^{-1}
$\bar{L}(\theta_n)$	4.762×10^{-3}	4.376×10^{-3}	3.252×10^{-3}
s^2	0.970×10^{-4}	0.881×10^{-4}	0.314×10^{-4}

ISU	$n = 10$	$n = 100$	$n = 1000$
a	1.359×10^{-3}	2.761×10^{-3}	5.832×10^{-3}
c	1.552×10^{-1}	1.934×10^{-1}	1.819×10^{-1}
$\bar{L}(\theta_n)$	4.837×10^{-3}	4.498×10^{-3}	3.383×10^{-3}
s^2	0.771×10^{-3}	0.703×10^{-3}	0.431×10^{-3}

SDT	$n = 10$	$n = 100$	$n = 1000$
a	1.691×10^{-3}	1.688×10^{-3}	1.742×10^{-3}
c	1.555×10^{-1}	1.681×10^{-1}	2.142×10^{-1}
$\bar{L}(\theta_n)$	4.777×10^{-3}	4.410×10^{-3}	3.224×10^{-3}
s^2	1.127×10^{-3}	0.955×10^{-3}	0.527×10^{-3}

Table 4 - Kowalik and Osborne Function (#15)

Altogether, the Bernoulli-perturbed algorithm performed significantly better than all three of the candidate distributions on 12 of 30 function optimizations. In the remaining 18 cases there was at least one candidate distribution that performed at least as well as the {–1, 1}-Bernoulli. In no case did any candidate distribution perform significantly better than the algorithm with Bernoulli perturbations.

As a general observation, the inverse split uniform distribution gave results closest to those of the {–1, 1}-Bernoulli distribution, followed closely by the symmetric double triangular, and more distantly by the split uniform distribution. One possible explanation (that still requires investigation) is that this outcome is tied to the mean magnitudes of the distributions used. The ISU distribution used for this analysis had the largest mean magnitude of all the candidate distributions, followed by SDT, and SU.

4. Conclusions and Further Study

The decision to optimize the algorithm parameters *a* and *c* was key to this analysis. The algorithm is sensitive to poor choices for these parameters. In every case where a candidate distribution appeared to perform better than the {–1, 1}-Bernoulli distribution, it turned out that the values *a* and *c* were not optimally selected and further tuning resulted in better algorithm performance and established the superiority, or at least equivalence, of the Bernoulli distribution. Based on the preceding results, we offer the following conjecture:

Conjecture. *Given optimal parameter selection, no choice of perturbation distribution provides better performance over the Bernoulli distribution for the simultaneous perturbation stochastic approximation algorithm for any sample size.*

A case has been made that the Bernoulli distribution, already proven asymptotically optimal, is also the best distribution to use in small-sample analysis, given optimal parameter selection. However, empirical results are not proof, and additional work needs to be done to develop a theory of small-sample stochastic approximation that can answer these and other questions.

5. Acknowledgement

This work was supported in part by the Johns Hopkins University Applied Physics Laboratory Independent Research and Development (IRAD) Program.

References

[1] M. C. Fu, A tutorial review of techniques for simulation optimization, *Proceedings of the 1994 Winter Simulation Conference*, 1994, 149-156.

[2] S. D. Hill, M. C. Fu, Optimization of discrete event systems via simultaneous perturbation stochastic approximation, *IIE Transactions*, *29*, 1997, 233-243.

[3] S. Andradottir, Simulation optimization, in *Handbook of simulation*, J. Banks, ed. (John Wiley and Sons, Inc., 1998, 307-333).

[4] S. H. Jacobson L.W. Schruben, A review of techniques for simulation optimization, *Operations Research Letters*, *8*, 1989, 1-9.

[5] H. Robbins, S. Monro, A stochastic approximation method, *Annals of Mathematical Statistics*, *22*(3), 1951, 400-407.

[6] J. Kiefer, J. Wolfowitz, Stochastic estimation of the maximum of a regression function, *Annals of Mathematical Statistics*, *23*(3), 1952, 462-466.
[7] P. L'Ecuyer, N. Giroux, P. W. Glynn, Stochastic optimization by simulation: numerical experiments with the M/M/1 queue in steady-state, *Management Science*, *40*, 1994, 1245-1261.
[8] A. Shapiro, Simulation-based optimization – convergence analysis and statistical inference, *Communications in Statistics – Stochastic Models*, *12*, 1996, 425-453.
[9] H. J. Kushner, D. C. Clark, *Stochastic approximation methods for constrained and unconstrained systems* (New York, NY: Springer-Verlag, 1978).
[10] J. C. Spall, *Introduction to stochastic search and optimization* (New York, NY: Wiley and Sons, 2003).
[11] J. C. Spall, Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, *IEEE Transactions on Automatic Control*, *37*, 1992, 332-341.
[12] J. C. Spall, J. A. Cristion, Nonlinear adaptive control using neural networks; estimation based on a smoothed form of simultaneous perturbation gradient approximation, *Statistica Sinica*, *4*, 1994, 1-27.
[13] D. W. Hutchison, S. D. Hill, Simulation optimization of airline delay with constraints, *Proceedings of the 2001 Winter Simulation Conference*, 2001, 1017-1022.
[14] J. C. Spall, A stochastic approximation technique for generating maximum likelihood parameter estimates, *Proceedings of the American Control Conference*, Minneapolis, MN, 1987 , 1161-1167.
[15] P. Sadegh, J. C. Spall, Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation, *IEEE Transactions on Automatic Control*, *43*, 1998, 1480-1484.
[16] J. C. Spall, Implementation of the simultaneous perturbation algorithm for stochastic optimization, *IEEE Transactions on Aerospace and Electronic Systems*, *34*, 1998, 817-823.
[17] J. J. Moré, B. S. Garbow, K. E. Hillström, Testing unconstrained optimization software, *ACM Transactions on Mathematical Sciences*, *7*(1), 1981, 17–41.
[18] F. J. Solis, R. J.-B. Wets, Minimization by random search techniques, *Mathematics of Operations Research*, *6*, 1981, 19-30.