Constrained Optimization via Stochastic Approximation with a Simultaneous Perturbation Gradient Approximation

Payman Sadegh

Dept. of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark.

Abstract. The paper deals with a projection algorithm for stochastic approximation using simultaneous perturbation gradient approximation for optimization under inequality constraints where no direct gradient of the loss function is available and the inequality constraints are given as explicit functions of the optimization parameters. It is shown that under application of the projection algorithm, the parameter iterate converges almost surely to a Kuhn-Tucker point. The procedure is illustrated by a numerical example.

Key Words. Optimization, Stochastic approximation, SPSA, Constraints, Kuhn-Tucker point.

1 INTRODUCTION

The simultaneous perturbation stochastic approximation (SPSA) algorithm has recently attracted considerable attention for multivariate optimization problems where only noisy measurements of the loss function are available (i.e., no gradient information is directly available), see e.g., Rezayat (1995), Maeda et al. (1995), Cauwenberghs (1994), Chin (1994), and Parisini & Alessandri (1995).

SPSA was introduced in Spall (1987) and more thoroughly analyzed in Spall (1992). The algorithm is a variant of the stochastic approximation (SA) in a Kiefer-Wolfowitz setting (Kushner & Clark, 1978) where only noisy measurements of the loss function are available (used for gradient approximations). The essential feature of SPSA is its highly efficient gradient approximation that requires only *two* loss function measurements regardless of the number of optimization parameters. The gradient approximation is generated by simultaneous (random) perturbation relative to the current estimate of the parameter θ. Note the contrast of two function measurements with the $2p$ measurements required in the classical finite-difference based Kiefer-Wolfowitz SA algorithm, where p is the number of parameters. Under reasonably general conditions, it was shown in Spall (1992) that the p-fold savings in function measurements per gradient approximation translates directly into a p-fold savings in the total number of measurements needed to achieve a given level of accuracy in the optimization process.

The original SPSA algorithm as presented in Spall (1992) is an unconstrained algorithm. Constraints, on the other hand, are essential parts of almost all real world optimization applications. The present work may be regarded as the extension of the convergence result of Spall (1992) to constrained optimization problems. This paper presents a projection SPSA algorithm that can handle *inequality* constraints. A similar approach is pursued in L'Ecuyer & Glynn (1994) for optimization of queuing systems using stochastic approximation. This paper considers SPSA and treats more general constraints. However, attention is restricted to the constraints that are given as explicit functions of the optimization parameter. Common to the Kiefer-Wolfowitz stochastic approximations, function evaluations often mean *real measurements* on the system. The problem of interest in this paper concerns situations where the constraints are determined by the feasible operating conditions of the system. Hence, it is assumed that function evaluations at the points where the constraints are violated are not feasible. This is stronger than the requirement of restricting the solution to the feasible domain (as in constrained versions of Robbins-Monro type SA algorithms, see Kushner & Clark (1978)). In this regard, the projection algorithm is advanta-

geous relative to other constrained SA optimization techniques such as the Lagrangian method Kushner & Clark (1978) where the parameter iterate only asymptotically lies in the feasible set. The paper establishes almost sure convergence of the parameter iterate to a Kuhn-Tucker point under application of the projection algorithm.
The organization of the rest of the paper is as follows. Section 2 studies the projection SPSA algorithm and the convergence result. Section 3 presents a numerical example where the procedure is illustrated and tested using (finite sample) numerical experimentations. Finally, Section 4 offers concluding remarks.

2 PROJECTION SPSA

In this section, a projection SPSA algorithm is presented for minimization under constraints, i.e. the problem of

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

where similar to the regularity conditions for the unconstrained case (Spall, 1992), the loss function $L(\theta)$ is continuously differentiable on an open set containing G. The reader is referred to Spall (1992) for a detailed treatment of the (unconstrained) SPSA algorithm. We deal with *inequality* constraints and introduce

A 1: The set $G = \{\theta : q_i(\theta) \leq 0, i = 1, ..., s\}$ is non-empty, bounded, and the functions $q_i(\theta)$, $i = 1, ..., s$, are continuously differentiable. At each $\theta \in \partial G$, where ∂ denotes boundary, the gradients of the active constraints are linearly independent. Furthermore, there exists an $\epsilon < 0$ such that the set $G^- = \{\theta : q_i(\theta) \leq r, i = 1, ..., s\}$ is non-empty for $\epsilon \leq r < 0$ (i.e., the set G has non-empty interior).

The proof of convergence to a Kuhn-Tucker point (see Proposition 1 below and Sadegh (1997)) is based on Theorem 5.3.1 of Kushner & Clark (1978) where the assumption on G (Kushner & Clark (1978), page 190, A5.3.1) states that G is the closure of its interior rather than the non-emptiness of G^- in Assumption 1. It is easy to see that because of the continuity of the $q_i(\theta)$, the set $\{\theta : q_i(\theta) < 0, i = 1, ..., s\}$ is open and indeed equal to $intG$ where int denotes interior. This together with the non-emptiness of G^- yields $G = \overline{intG}$. Assumption 1 is formulated with the goal of easing later presentation.
Another type of constrained problems involves constraint functions that can only be observed in the presence of noise, see e.g. Ljung et al. (1992). Such constraints will not be examined here.
Let $\hat{\theta}_k$ denote the estimate for θ at the kth iteration, and for all $\theta \in \mathbb{R}^p$, let $P(\theta)$ be the nearest point to θ on G where the norm is defined as the usual Euclidean norm. The projection algorithm has the general form

$$\hat{\theta}_{k+1} = P(\hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)) \tag{2.1}$$

where the gain sequence $\{a_k\}$ shall satisfy certain conditions (as follows) and $\hat{g}_k(\hat{\theta}_k)$ is an approximation to the gradient at $\hat{\theta}_k$. The simultaneous perturbation estimate for the gradient at θ, $g(\theta)$, is defined as follows. Let $\Delta_k \in \mathbb{R}^p$ be a vector of p mutually independent mean-zero random variables $\{\Delta_{k1}, \Delta_{k2}, ..., \Delta_{kp}\}$ satisfying certain conditions (Spall, 1992). A condition on random perturbations is norm boundedness, i.e. $||\Delta_k|| \leq \alpha_0$ for some $\alpha_0 > 0$. In Spall (1992), the boundedness condition is $||\Delta_k|| \leq \alpha_0$ a.s. Noting that the user has full control over random perturbations, for simplicity the strict boundedness assumption is introduced. Consistent with the usual framework of stochastic approximations, noisy measurements of the loss function are available. In particular, at the kth iteration

$$\begin{aligned} y_k^{(+)} &= L(\theta + c_k \Delta_k) + \epsilon_k^{(+)}, \\ y_k^{(-)} &= L(\theta - c_k \Delta_k) + \epsilon_k^{(-)} \end{aligned}$$

where $\{c_k\}$ is a gain sequence and $\epsilon_k^{(+)}$ and $\epsilon_k^{(-)}$ represent measurement noise terms that satisfy $E\{\epsilon_k^{(+)} - \epsilon_k^{(-)} | \theta, \Delta_k\} = 0$. The gain sequences $\{a_k\}$ and $\{c_k\}$ are positive for all k and tend to zero as $k \to \infty$. Moreover, $\sum_{k=0}^{\infty} a_k = \infty$, and $\sum_{k=0}^{\infty} (a_k/c_k)^2 < \infty$. For convenience, take $c_k = c/k^\gamma$, $\gamma \geq 0$.
The basic simultaneous perturbation (SP) form for the estimate of $g(\theta)$ at iteration k is defined by

$$g_k^{SP}(\theta) = \begin{bmatrix} \frac{y_k^{(+)} - y_k^{(-)}}{2c_k \Delta_{k1}} \\ \vdots \\ \frac{y_k^{(+)} - y_k^{(-)}}{2c_k \Delta_{kp}} \end{bmatrix}.$$

Note that at each iteration, only *two* measurements are needed to form the estimate. The main features of our proposed solution relative to the unconstrained SPSA algorithm are as follows. Firstly, the projection $P(\cdot)$ always restricts the iterates $\hat{\theta}_k$ within G which is obviously not needed for the unconstrained case. The projection

is indeed an essential feature of the constrained algorithm: eliminating $P(.)$, the iterates may vary anywhere in $\mathbb{R}^p$ as a result of noisy observations, no matter how the gain or random perturbation sequences of the algorithm are selected. Secondly, for the unconstrained algorithm, it holds that $\hat{g}_k(\theta_k) = g_k^{SP}(\hat{\theta}_k)$. Such approximation cannot be directly used here since it may occur that $\hat{\theta}_k \in G$ but $\theta_k \pm c_k\Delta_k \notin G$. Especially, in case $\theta_k \in \partial G$, there is always a (random) direction Δ_k such that $\hat{\theta}_k \pm c_k\Delta_k \notin G$, no matter how small the gain c_k is selected. Notice that the case $\hat{\theta}_k \in \partial G$ is expected to occur frequently for the very relevant situation that the true optimum belongs to the boundary of the feasible domain. Except for simulation based optimization cases, function evaluations involve real measurements on the system and it is usually not allowed to take measurements outside the feasible domain. To overcome this problem, θ_k is further projected onto a (closed) set G_k contained within G to obtain $P_k(\hat{\theta}_k)$ which shall (only) be used for computing an SP gradient approximation at the kth iteration. If the distance d_k between the nearest points on ∂G and ∂G_k is equal to or larger than $c_k\alpha_0$, then $P_k(\hat{\theta}_k) \pm c_k\Delta_k \in G$, ensuring that the SP approximation to the gradient at $P_k(\theta_k)$ (instead of θ_k) requires no function measurement outside G. The SP gradient approximation at $P_k(\theta_k)$ obviously introduces an (extra) error term relative to the SP gradient approximation at $\hat{\theta}_k$. However, if $G_k \to G$ for $k \to \infty$, then continuous differentiability of $L(\theta)$ yields that the extra error term tends to zero. This line of argument is used for the proof of convergence. But first, let us describe a procedure for selecting the G_k (a simple case of this is given in the illustrative example of the paper). Define $G_k \subset G$ by $G_k = \{\theta : q_i(\theta) \leq r_k < 0, i = 1, ..., s\}$ where $r_k \to 0$ as $k \to \infty$. Assumption 1 states that there exists an $\epsilon < 0$ such that G_k is non-empty for $\epsilon \leq r_k < 0$, $k = 1, 2, ...$ Hence, select $\epsilon \leq r_1 < 0$ and select c such that $d_1 \geq c_1\alpha_0$. Once c is selected, choose $r_k \to 0$ such that $d_k \geq c_k\alpha_0$ (note that $c_k \to 0$ as $k \to \infty$).

REMARK 1: It follows from above that the bottom-line in computing an SP gradient approximation at $P_k(\hat{\theta}_k)$ is to ensure the feasibility of function evaluations. There may therefore exist different methods to obtain a point $\theta'_k \in G$ for the SP gradient approximation at iteration k such that $\theta'_k \pm c_k\Delta_k \in G$, and in some sense, the magnitude of $\theta'_k - \hat{\theta}_k$ is small for all k (to avoid large error terms on the gradient approximations) and becomes infinitesimally small as $k \to \infty$. The proposed solution of the paper provides a suitable technique which can be generically applied to all types of constrained problems where Assumption 1 holds.

Finally, it should be noted that projections in general are unfortunately not very easy to compute unless linear approximations to $q_i(\theta)$ at the current iterate are obtained first. Such approximations can often be justified in practice, since $a_k \to 0$.

PROPOSITION 1: Let Assumption 1, and assumptions A1-A5 and conditions of Lemma 1 (for simplicity, replace the a.s. boundedness of Δ_k by strict boundedness) of Spall (1992) hold where all the regularity conditions on $L(\cdot)$ hold on an open set containing G. Then under the projection algorithm (see Eq(2.1)) where $\hat{g}_k(\theta_k) = g_k^{SP}(P_k(\theta_k))$, as $k \to \infty$

$$\hat{\theta}_k \to KT \quad \text{a.s.},$$

where KT is the set of Kuhn-Tucker points (i.e. the set of points θ where there are $\lambda_i \geq 0$ such that $g(\theta) + \sum_{i:q_i(\theta)=0} \lambda_i dq_i(\theta)/d\theta = 0$).

Proof: See Sadegh (1997).

3 ILLUSTRATIVE EXAMPLE

Let us study a simple numerical example concerning optimization of temperature profiles in a tubular reactor for two first-order irreversible consecutive reactions. See Fan (1966) for details. The first-order reactions $A \longrightarrow B \longrightarrow C$ take place in the reactor. The reaction $A \longrightarrow B$ has the specific rate $k_1(t)$ and $B \longrightarrow C$ has the rate $k_2(t)$ at time t. Denoting the concentration of A by $x_1(t)$ and the concentration of B by $x_2(t)$, we arrive at the following state-space equation which describes the dynamics of the reactions (Fan, 1966)

$$\begin{aligned} \dot{x}_1(t) &= -k_1(t)x_1(t) \\ \dot{x}_2(t) &= k_1(t)x_1(t) - k_2(t)x_2(t). \end{aligned} \tag{3.1}$$

The specific rates are given by $k_1(t) = k_{10}e^{-E_1/RT(t)}$ and $k_2(t) = k_{20}e^{-E_2/RT(t)}$ where $T(t)$ is the temperature profile (the control variable) and k_{10}, k_{20}, E_1, E_2, and R are constants. It is desired to find the temperature profile that (starting from time $t_0 = 0$) maximizes $x_2(t_f)$, i.e. the concentration of the product B at $t = t_f$. By selecting a sampling time, the problem becomes a multivariate optimization problem where the temperature values at discrete time points should be determined such that the final concentration of B is maximized. Solutions are given both under no constraints and under the situation that the applied profiles should satisfy $335 \leq T(t) \leq 342$.

Unlike the (unconstrained) solution given in Fan (1966), our solution is in principle based on trials and experimentations on the system. The trials consist of applying temperature profiles to the system and doing measurements on $x_2(t_f)$ for each applied profile. SPSA requires no model for the optimization, neither does it require knowledge of the values of the constants. In this example, the presented model is used for (and only for) simulation, data generation, and testing our procedure.

Let us assume the following numerical values (Fan, 1966): $k_{10} = 0.534 \times 10^{11}$/min, $k_{20} = 0.461 \times 10^{18}$/min, $E_1 = 18000$cal/mole, $E_2 = 30000$cal/mole, $R = 2$cal/mole-K°, $t_f = 8$min, $x_1(0) = 0.8160$mole/liter, $x_2(0) = 0.2260$mole/liter.

Let us further assume that the temperature $T(t)$ is constant for $i-1 \leq t < i$, $i = 1, 2, ..., 8$ (i.e., a piecewise constant input). The ith element of the 8-dimensional optimization parameter θ is equal to $T(t)$ for $i-1 \leq t < i$.

The following constants are used throughout the example unless otherwise specified. The number of iterations for the SPSA algorithm is 250, the random perturbations are Bernoulli distributed with magnitude one, i.e. $Pr(\Delta_{ki} = \pm 1) = 0.5$ for $i = 1, ..., 8$, and all $k = 1, 2, ...$ and the gain sequences are selected as $a_k = 1000/k^{0.602}$, $c_k = 1/k^{0.101}$. These decay rates for a_k and c_k are empirically found to yield optimal performance for the unconstrained SPSA algorithm in finite sample cases, see Spall (1995). It is moreover assumed throughout the example that the measured values of $x_2(t_f)$ are corrupted with additive i.i.d. Gaussian noise with standard deviation 0.0005, and finally, the initial temperature profile, $T_{in}(t)$, for the optimizations is chosen to be 342K° at $t = 0$ and to drop 1K° per minute.

In order to determine the true optimal profiles, we use standard techniques which unlike SPSA make use of the model given by Eq(3.1) and assume noise free data. Eq(3.1) can be written as $\dot{x}(t) = A(t)x(t)$ where $x(t)$ is the state vector and $A(t)$ is a piecewise constant matrix ($A(t)$ is constant in the interval $i-1 \leq t < i$, $i = 1, ..., 8$). The explicit relation between the objective function and the control variable is obtained using $x(t_f = 8) = \exp\{\sum_{j=0}^{t_f - 1} A(j)\}x(0)$, and standard optimization algorithms can be applied to find the optimal profiles. We use MATLAB® optimization toolbox functions CONSTR and FMINS (Grace, 1994) for the constrained and unconstrained cases, respectively.

Now, let us try both the constrained and unconstrained SPSA algorithms to *estimate* the optimal profiles. Define $G_k = \{\theta : 335 + c_k \leq \theta_i \leq 342 - c_k, i = 1, ..., 8\}$ for the constrained case. For each of the constrained and unconstrained cases, the optimal profile is estimated 500 times (i.e., 500 cross-sections for each algorithm). The obtained estimates are denoted by $\hat{T}_c(t)$ and $\hat{T}_u(t)$, respectively (notice the randomness in the iterates due to measurement noise for SPSA). As expected, all the 500 realizations of $\hat{T}_c(t)$ are restricted within [335.342] (for all $0 \leq t < 8$) while the largest value (among 500 realizations) of $\max_t \hat{T}_u(t)$ is 345.5. For each realization of $\hat{T}_c(t)$ and $\hat{T}_u(t)$, we compute (1) the relative error defined by $\{\int_0^{t_f}[T^*(t) - T_r(t)]^2 dt / \int_0^{t_f}[T^*(t) - T_{in}(t)]^2 dt\}^{\frac{1}{2}}$ where $T_r(t)$ and $T^*(t)$ are the relevant realization and true optimal profile (as computed previously), and (2) the noise free value of $x_2(t_f)$ corresponding to the realization (hence randomness in this computed value is only due to randomness in $\hat{T}_c(t)$ and $\hat{T}_u(t)$). By averaging over these computed values, an average relative error (ARE) and an average final product value (AFP) are obtained for both the constrained and unconstrained cases. The results are summarized in Table 1 where OFP denotes the relevant optimal final product value as given by the true optimal profiles.

In order to investigate the effect of the extra error on the gradient approximation (introduced to make the measurements feasible), the constrained optimal profile is estimated 500 times using the projection SPSA algorithm, but we use $g_k(\theta_k) = g_k^{SP}(\hat{\theta}_k)$. The corresponding ARE and AFP values for this case are 0.1561 and 0.6988 respectively. Comparing the obtained ARE to 0.1819 (see Table 1) indicates improvement, but at the expense of infeasibility of the measurements.

It is also of interest to assess the convergence rate of the constrained algorithm. The optimal profile is estimated 500 times using constrained SPSA with 1000 iterations (same algorithm constants as before) for each cross-section which yields an ARE value of 0.1139. The quantity $-\log(0.1819/0.1139)/\log(250/1000) = 0.338$ is then used as an assessment for the convergence rate. Using Proposition 2 of Spall (1992), the (asymptotic) convergence rate of the unconstrained algorithm for the gain sequences of this example is equal to 0.2 which is considerably less than the computed rate 0.338.

Finally, the constrained two-sided FDSA algorithm is applied 500 times with the same algorithm constants as for the constrained SPSA, but 32 iterations for each cross-section. Notice that the total number of measurements for the FDSA algorithm with 32 iterations is equal to $32 \times 2 \times 8 = 512$ which is slightly larger than the total number of measurements for the SPSA algorithm with 250 iterations ($250 \times 2 = 500$). The ARE and AFP values become 0.2117 and 0.6988. The ARE value for the constrained FDSA is no-

ticeably larger than 0.1819 obtained for the constrained SPSA algorithm for (almost) the same number of measurements.
It should be noted that a formal treatment of the convergence rate and accuracy of the estimate of the constrained SPSA algorithm is required before one is able to draw any definitive conclusion about the behavior of the algorithm.

TABLE 1 Constrained and unconstrained optimization using 500 cross-sections of SPSA. All the final product values are based on noise free evaluations of $x_2(t_f)$.

Constrained		
ARE	AFP	OFP
0.1819	0.6988	0.6989
Unconstrained		
ARE	AFP	OFP
0.3291	0.6996	0.6999

4 CONCLUDING REMARKS

The paper presents a projection algorithm for constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation where no gradient information is directly available. The algorithm can handle inequality constraints given as explicit functions of the parameter. The constraints should define a set with non-empty interior. We have considered the case where measurements outside the constraint set are not feasible which is stronger than restricting the solution to the feasible domain. The paper establishes almost sure convergence of the iterate to a Kuhn-Tucker point.

Possible directions for future study are the performance of the algorithm, distribution or convergence rate of the iterate, possible error bounds on the estimate, and optimal tuning of the algorithm constants, i.e. optimal selection of gain sequences. Finally, an identical proof of convergence can be applied to a projection FDSA algorithm (see Remark ??). It will be of interest to compare the number of measurements that constrained SPSA and constrained FDSA need to reach a certain level of accuracy (see Section 1 and Spall (1992) for a similar comparison in the unconstrained case).

REFERENCES

Cauwenberghs, G. (1994). *Analog VLSI Autonomous Systems for Learning and Optimization*. Ph.D. thesis, Dept of Electrical Engineering, California Institute of Technology.

Chin, D. C. (1994). A more efficient global optimization based on Styblinski and Tang. *Neural Nets.*, **7**, 573–574.

Fan, L. T. (1966). *The Continuous Maximum Principle*. John Wiley & Sons, New York.

Grace, A. (1994). *Optimization Toolbox for Use with MATLAB®*. The Math Works Inc.

Kushner, H. J. & Clark, D. S. (1978). *Stochastic Approximation for Constrained and Unconstrained Systems*. Springer-Verlag, Berlin.

L'Ecuyer, P. & Glynn, P. W. (1994). Stochastic optimization by simulation: convergence proofs for the GI/G/1 queue in steady-state. *Management Science*, **40**(11), 1562–1578.

Ljung, L., Pflug, G., & Walk, H. (1992). *Stochastic Approximation and Optimization of Random Systems*. Birkhäuser, Berlin.

Maeda, Y., Hirano, H., & Kanata, Y. (1995). A learning rule of neural networks via simultaneous perturbation and its hardware implementation. *Neural Nets.*, **8**, 251–259.

Parisini, T. & Alessandri, A. (1995). Non-linear modeling and state estimation in a real power plant using neural networks and stochastic approximation. In *Proc. American Control Conference*, pp. 1561–1567.

Rezayat, F. (1995). On the use of an SPSA-based model free controller in quality improvement. *Automatica*, **31**, 913–915.

Sadegh, P. (1997). Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation. *Automatica*, in press.

Spall, J. C. (1987). A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proc. American Control Conference*, pp. 1161–1167.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, **37**(3), 332–341.

Spall, J. C. (1995). Implementation of simultaneous perturbation algorithm for stochastic optimization. Submitted to *American Statistician*.