Yu. Ermoliev R. J-B Wets (Eds.)

Numerical Techniques for Stochastic Optimization

With 62 Figures

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo

[20] L. Schmetterer, "Stochastic approximation", *Proc. Fourth Berkeley Symp. Univ. California Press*, **Vol. I**, (1961), 587–609.

[21] R.L. Sielken, "Some Stopping Times for Stochastic Approximation Procedures", *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **27**(1973), 79–86.

[22] H. Walk, "An Invariance Principle for the Robbins-Monro Process in a Hilbert Space", *Z. Warscheinlichkeitstheorie verw. Gebiete* **39**(1977), 135–150.

CHAPTER 18

ADAPTIVE STOCHASTIC QUASIGRADIENT PROCEDURES*

S. Urasiev

18.1 Introduction

In this chapter we deal with iterative algorithms for solving stochastic optimization problem

$$\min Ef(x,\omega) \tag{18.1}$$

subject to constraints

$$x \in X \subset R^n$$

where x are variables to be chosen which take values in Euclidean space and ω are random parameters which belong to some probability space. Our main concern is the improvement of performance features of the stochastic quasigradient (SQG) method

$$x^{s+1} = \pi_X(x^s - \rho_s \xi^s) \tag{18.2}$$

where π_X is the projection operator on the set X, x^s-current approximation to solution, ρ_s is the stepsize and ξ^s is step direction, which roughly speaking, in average points to the direction of gradient of the function $Ef(x,\omega)$. Reader can find survey of such methods and further references in Chapter 6 (see also [1]). One of the main challenges which arise before implementor of SQG methods is appropriate selection of the stepsize ρ_s. Theory gives only very general guidelines:

$$\rho_s \longrightarrow 0, \sum_{s=0}^{\infty} \rho_s = \infty, \sum_{s=0}^{\infty} \rho_s^2 < \infty.$$

In papers written earlier on stochastic approximation [2], stepsize was chosen in advance to satisfy these conditions. For instance, $\rho_s = c/s$. In what follows, such choices which depend only on iteration number will be called programmed or off-lined rules. Unfortunately they lead to very slow convergence, although they assure in some sense optimal asymptotic rate. However, in practical computations SQG methods can be used to reach reasonable neighborhood of solution, not exact value of solution. For such purposes, asymptotic results are not relevant as well as programmed rules of choosing stepsize. In this chapter, adaptive or on-line rules for computing ρ_s are studied which exhibit much

* **This chapter is based on the report presented at the International Conference on Stochastic Optimization, Kiev, 1984.**

more satisfactory behavior. Such methods utilize information gathered during optimization process to make decision about current value of stepsize ρ_s. More specifically, ρ_s may depend on observations of random function $f(x^k, u^k)$ or stochastic quasigradient ξ^k in some or all preceding iterations $k \le s$. Some on-line rules were proposed in [3]–[7]. This chapter is based on [5]–[7] and describes one particular adaptive SQG method in which stepsize increases or decreases depending on whether subsequent quasigradients point to the same or to the opposite directions.

This chapter consists of 5 sections. In Section 18.2 the adaptive SQG method is described, its convergence is analyzed in Section 18.3. Implementation details are discussed in Section 18.4, and the chapter ends in Section 18.5 with a description of some particular problems solved by algorithm together with results of numerical experiments.

18.2 Algorithm Description

In what follows we shall consider algorithm of type (18.2) for problem (18.1). It will be assumed that the process takes place in probability space (Ω, A, P) where A is σ-field and P- probability measure. Vector ξ^s from (18.2) is stochastic quasigradient, i.e.

$$E(\xi^s/B^s) = F_x(x^s) + b^s$$

where $F_x(x^s)$ is gradient of the function $F(x) = E_\omega f(x,\omega)$, conditions on b^s will be imposed later and B^s is σ-field defined by the process history, i.e., random variables $\{x^o, \dots, x^s\}$. We shall keep in mind that x^s depends on random parameters from Ω, but will not specify this dependence explicitly.

We shall explain at first the idea of adaptive stepsize control informally. Here, for simplicity we shall assume that function $F(x)$ is smooth and $X = R^n$. It is quite naturally to choose step ρ_s to minimize $F(x)$ along direction ξ^s, i.e., such that function $\varphi_s(p)$ has minimum over ρ, where

$$\varphi_s(\rho) = E[F(x^s - \rho\xi^s)/x^s].$$

This is analogue of stepsize rules used extensively in deterministic optimization. It is easy to see that

$$\begin{aligned}\frac{\partial}{\partial\rho}\varphi_s(\rho)|_{\rho_s} &= E[\frac{\partial}{\partial\rho}F(x^s - \rho\xi^s)|_{\rho=\rho_s}/x^s] \\ &= -E[\langle\nabla F(x^s - \rho_s\xi^s), \xi^s\rangle/x^s] \\ &= -E[\langle\nabla F(x^{s+1}), \xi^s\rangle/x^s] \\ &= -E[\langle(\xi^{s+1}, \xi^s > /x^s], \quad s = 0, 1\dots\end{aligned}$$

Thus, $-(\xi^{s+1}, \xi^s)$ is stochastic quasigradient of function $\varphi_s(\rho)$ in point ρ_s on iteration $s+1$. To modify step ρ_s, let us use the following gradient procedure:

$$\rho_{s+1} = \rho_s + \lambda_s\langle\xi^{s+1}, \xi^s\rangle\lambda_s > 0, s = 0, 1\dots \tag{18.3}$$

The value of (ξ^{s+1}, ξ^s) gives some information whether current value ρ^s exceeds minimum of function $\varphi_s(\rho)$ over ρ on the iteration s. If $(\xi^{s+1}, \xi^s) > 0$ then it is probable that minimum of $\varphi_s(\rho)$ is greater than ρ_s and it is necessary to increase stepsize and decrease it when sign is negative. This information is used to modify step ρ_s.

Naturally, decision based on this arguments will be subject to error due to stochastic phenomena. However, this errors will be smoothed out in the course of iterations. It is convenient to rewrite relation (18.3) in the following form

$$\rho_{s+1} = \rho_s a_s(\xi^{s+1}, \xi^s), a_s > 0, \quad s = 0, 1, \dots, \tag{18.4}$$

(18.3) is the special case of (18.4) since for each λ_s, such that $\rho_{s+1} > 0, a_s$ can be selected respectively so that $\rho_s + 1$ computed by formulas (18.3), (18.4) coincide. In order to guarantee fulfillment of the convergence condition for SQG algorithms $\sum_{s=0}^{\infty}\rho_s = \infty$ (see Chapter 6), the value a_s is calculated by formula

$$a_s = a^{\rho_s}, a > 1, \quad s = 0, 1, \dots,$$

Convergence of the algorithm (18.2) with the stepsize rule (18.4) can be established [7] in deterministic case, when $\xi^s = F_x(x^s)$ and $F(x)$ is a strongly convex function. For stochastic case, let us modify formula (18.4) as follows

$$\begin{aligned}\rho_{s+1} &= \rho_s a^{\rho_s\langle\xi^{s+1}, \xi^s\rangle - \delta\rho_s} \\ &= \rho_s a^{\langle\xi^{s+1}, x^s - x^{s+1}\rangle - \delta\rho_s}, \delta > 0, \quad s = 0, 1, \dots\end{aligned} \tag{18.5}$$

Introduction of the term $\delta\rho_s$ guarantees fulfillment of one more convergence condition

$$\rho_s \longrightarrow 0 \text{ a.s. } s \longrightarrow \infty.$$

18.3 Convergence Analysis

Besides convergence of sequence x^s to the solution of problem (18.1), we are also interested in convergence of some convex combinations of this sequence. With sequence x^s generated by algorithm (18.2), (18.5), it will be associated the sequence

$$\overline{x}^s = \sum_{\ell=0}^{s}\rho_\ell x^\ell / \sum_{\ell=0}^{s}\rho_\ell \tag{18.6}$$

and the convergence of $\overline{x}^s$ to the solution will be studied. If such convergence does occur the initial sequence x^s will be called Cesaro convergent. The advantages of dealing with such convergence are the following:

- the sequence $\overline{x}^s$ displays much more regular behavior than original sequence x^s
- $\overline{x}^s$ can be computed with almost no additional effort in iterative way using the sequence $\overline{x}^s$.

– some convergence conditions can be relaxed for Cesaro convergence and in some cases x^s does not converge to the solution in ordinary sense, but is Cesaro convergent.

This type of convergence was used in [8],[9]. The following theorem from [7] gives conditions for Cesaro convergence of the method (18.2). We shall use the abbreviation a.s. for the words "almost sure".

Theorem 1. *Let $F(x)$ be a convex function defined on convex closed bounded set $X \subset R^n$,*

$$\max_{x,y \in X} \|x - y\| = C_1, \tag{18.7}$$

$$E\|\xi^s - F_s(x^s) - b^s\|^2 \le C_2^2 \tag{18.8}$$

$$\overline{\lim_{s \to \infty}} \|b^s\| \le \bar{b}, \tag{18.9}$$

$$\rho_s > 0 \text{ a.s.}, \quad s = 0, 1, \ldots, \tag{18.10}$$

$$E\rho_s^2 < \infty, \quad s = 0, 1, \ldots, \tag{18.11}$$

$$\rho_s \longrightarrow 0 \text{ a.s. with } s \longrightarrow \infty, \tag{18.12}$$

$$\sum_0^\infty \rho_s = \infty \text{ a.s.}, \tag{18.13}$$

and at least one of the two following conditions is satisfied:

(1) Step ρ_s depends only on $(x^0, \ldots, x^s, \xi^0, \ldots, \xi^{s-1})$ (it is measurable with respect to σ-algebra B_s induced by $(x^o, \ldots, x^s, \xi^o, \ldots, \xi^{s-1})$);

(2) $\rho_s \rho_{s-1}^{-1} \longrightarrow 1$ a.s., ρ_s depends only on $(x^0, \ldots, x^s, \xi^0, \ldots, \xi^s)$ (it is measurable with respect to σ-algebra induced by $(x^0, \ldots, x^s, \xi^0, \ldots, \xi^s)$).

Then

$$\overline{\lim_{s \to \infty}} F(\overline{x}^s) - F(x^*) \le \bar{b} C_1 \text{ a.s.}$$

where

$$x^* \in X^* = \{x^* : F(x^*) = \min_{x \in X} F(x)\}.$$

and $\overline{x}^s$ is defined by (18.6).

Corollary. *If $b^s \longrightarrow 0$ a.s. then all accumulating points of the sequence $\overline{x}^s$ are the solutions of problem (18.1).*

The main difference between conditions for Cesaro convergence and convergence in usual sense for SQG methods it that condition $\sum_{s=0}^\infty p_s^2 < \infty$, which is needed for normal convergence a.s., is not needed for Cesaro convergence. This makes verifying convergence conditions for adaptive SQG methods much easier.

Now we are prepared to give convergence results for adaptive SQG method (18.2), (18.5).

Theorem 2. *Let $f(x)$ be a convex (possibly nonsmooth) function defined on some vicinity of convex compact subset $X \subset R^n$. If the following conditions are satisfied*

$$\max \|x - y\| = C_1 \tag{18.14}$$

$$\sup \|\xi^s\| < C_2 \text{ a.s.} \tag{18.15}$$

$$\overline{\lim}_{s \to \infty} \|b^s\| \le \bar{b}, \tag{18.16}$$

$$\delta > C_1 \overline{\lim}_{s \to \infty} \inf_{h \in \partial F(x^s)} \|\xi^s - h\| \text{ a.s.} \tag{18.17}$$

where $\partial F(x)$ is the set of subgradients of $F(x)$ at point x. Then

$$\overline{\lim}_{s \to \infty} (F(\overline{x}^s) - \min_{x \in X} F(x)) \le \bar{b} C_1 \text{ a.s.}$$

i.e. if $\lim_{s \to \infty} b^s = 0$ then

$$F(x^s) - \min x \in X F(x) \longrightarrow 0 \text{ a.s.}$$

and all accumulating points of the sequence $\overline{x}^s$ are solutions of the problem (18.1) a.s.

Proof. Condition (18.10) of Theorem 1 follows directly from (18.5) since $\rho_0 > 0$ and $a > 0$. Here we shall give only an outline of the proof, which consists of checking conditions of theorem 1. We shall check here conditions (18.12)-(18.13) of the theorem 1 and assume $b^s = 0$ (for more details see [5]-[7]).

1. Let us show that condition (18.13) of Theorem 1 is satisfied, i.e. $\sum_{s=0}^\infty \rho_s = \infty$ a.s.. Assume the opposite, i.e. exists such constant K that probability of the event

$$A = \{\omega : \sum_{s=0}^\infty \rho_s \le K\}$$

is positive. From projection properties and (18.15), we get the estimate

$$\|x^{s+1} - x^s\| \le \|\rho_s \xi^s\| \le \rho_s C_2 \text{ a.s.} \tag{18.18}$$

Stepsize rule (18.5) together with (18.18) yields:

$$\rho_{s+1} = \rho_s a^{\langle \xi^{s+1}, x^s - x^{s+1} \rangle - \delta \rho_s} \ge \rho_s a^{-\|\xi^{s+1}\| \|x^s - x^{s+1}\| - \delta \rho_s}$$

$$\ge \rho_s a^{-(C_2^2 + \delta)\rho_s} = \rho_s a^{-C_3 \rho_s}$$

where $C_3 = (C_2^2 + \delta)$. Therefore for $\omega \in A$ the following relation holds

$$\rho_{s+1} \ge \rho_s a^{-C_3 \rho_s} \ge \rho_0 a^{-C_3 \Sigma_0^s \rho_l} \ge \rho_0 a^{-C_3 K}$$

which implies $\sum_{s=0}^\infty \rho_s = \infty$ for $\omega \in A$ contradicting initial assumption. Therefore condition (18.13) is satisfied.

2. Consider now condition (18.12) and let us prove that $\rho_s \longrightarrow 0$ as $s \longrightarrow 0$ a.s. Denoting

$$C_s = \inf_{h \in \partial F(x^s)} \|\xi^s - h\|$$

we obtain the following estimate:

$$\begin{aligned}\langle \xi^{s+1}, x^s - x^{s+1}\rangle - \delta\rho_s &\le \langle \hat{F}_x(x^{s+1}), x^s - x^{s+1}\rangle \\ &\quad + \langle \xi^{s+1} - \hat{F}_x(x^{s+1}), x^s - x^{s+1}\rangle - \delta\rho_s \\ &\le F(x^s) - F(x^{s+1}) \\ &\quad + \|\xi^{s+1} - \hat{F}_x(x^{s+1})\|\|x^s - x^{s+1}\| - \delta\rho_s \\ &\le F(x^s) - F(x^{s+1}) \\ &\quad + C_2\rho_s\|\xi^{s+1} - \hat{F}_x(x^{s+1})\| - \delta\rho_s \text{ a.s.}\end{aligned}$$

Since $\hat{F}_x(x^{s+1})$ in the last relation is an arbitrary vector belonging to set $\partial F(x^{s+1})$, we obtain

$$\begin{aligned}\langle \xi^{s+1}, x^s - x^{s+1}\rangle - \delta\rho_s &\le F(x^s) - F(x^{s+1}) + C_1\rho_s \inf_{h \in \partial F(x^s)} \|\xi^s - h\| - \delta\rho_s \\ &= F(x^s) - F(x^{s+1}) + (C_2C_s - \delta)\rho_s \text{ a.s.}\end{aligned}$$

By substituting this estimate into (18.5), we obtain

$$\begin{aligned}\rho_{s+1} &\le \rho_s a^{F(x^s) - F(x^{s+1}) + (C_2C_s - \delta)\rho_s} \\ &\le \rho_0 a^{\sum_{\ell=0}^{s}(F(x^\ell) - F(x^{\ell+1})) + \sum_{\ell=0}^{s}(C_2C_s - \delta)\rho_\ell} \\ &= \rho_0 a^{F(x^0) - F(x^{s+1}) + \sum_{\ell=0}^{s}(C_2C_s - \delta)\rho_\ell}\end{aligned}$$

Taking into consideration $\sum_{s=0}^{\infty}\rho_s = \infty$ a.s. and relations (18.14), (18.17), we see that the expression in the exponent in the last relation tends to $-\infty$:

$$\lim_{s \to \infty}\left[f(x^0) - f(x^s)) + \sum_{\ell=0}^{s}(C_2C_s - \delta)\rho_\ell\right] \longrightarrow -\infty \text{ a.s.}$$

Since $a > 1$, this implies $\rho_s \to 0$ a.s.

Now, we have to show that condition (18.2) of Theorem 1 is satisfied. The following relation is satisfied:

$$\frac{\rho_{s+1}}{\rho_s} = a^{\langle \xi^{s+1}, x^s - x^{s+1}\rangle - \delta\rho_s}.$$

Since $\rho_s \to 0$ a.s., then

$$\langle \xi^{s+1}, x^s - x^{s+1}\rangle - \delta\rho_s \longrightarrow 0 \text{ a.s. and}$$

$$\frac{\rho_{s+1}}{\rho_s} \longrightarrow 0 \text{ a.s.}$$

after all conditions of Theorem 1 are tested, the statement of this theorem follows from it.

18.4 Implementation Strategies

In this paragraph problems which arise during the implementation of stochastic quasigradient algorithm (18.2), (18.5) are discussed. Its implementation includes some heuristical elements. The implemented method can be used for the fast finding of good initial approximation in the vicinity of solution. The implemented algorithm described below performed essentially better than the method with programmed rule for step size selection. First, we shall present the algorithm and then discuss some of its features.

Algorithm. Set $s = 0$ at the beginning of the computation.

Step 1. Computation of stochastic quasigradient ξ^s.

Step 2. Averaging of the stochastic quasigradient norm $\|\xi^s\|$

$$G_s = G_{s-1} + (\|\xi^s\| - G_{s-1}) \cdot D.$$

At the beginning of the computation $G_1 = 0$.

Step 3. The computation of the average current point drift

$$Q_s = G_s\rho_s$$

Step 4. Check the stopping criterion: if $Q_s < Q_*$ or $s > s_*$, finish the computation, otherwise go to the next step.

Step 5. The computation of scalar production T_s:

$$T_s = (\xi^s, x^{s-1} - x^s).$$

Step 6. Averaging of the T_s absolute value:

$$Z_s = Z_{s-1} + (|T_s| - Z_{s-1}) \cdot D.$$

At the beginning of the computation $Z_{-1} = 0$.

Step 7. Rule for the step size ρ_s selection:

$$\rho_s = \rho_{s-1}R^{\frac{T_s}{Z_s}} \times \begin{cases} 1 & \text{if } t_s > 0 \\ U & \text{if } T_s \le 0. \end{cases}$$

Step 8. Reducing the step size change.

$$\rho_s = \begin{cases} 3\rho_{s-1} & \text{if } \rho_s\rho_{s-1}^{-1} > 3, \\ \frac{\rho_{s-1}}{4} & \text{if } \rho_s\rho_{s-1}^{-1} < 4^{-1}, \\ \rho_s & \text{otherwise.} \end{cases}$$

Step 9. Finding the next approximation

$$x^{s+1} = x^s - \rho_s\xi^s.$$

Step 10. Projection on the feasible region X

$$x^{s+1} = \pi_X(x^{s+1})$$

Step 11. Take $s = s + 1$ and go to Step 1.

Two stopping criteria are implemented in the method. The first one is by the number of iterations. The second stopping criterion is by the value of the mean point trend which is equal to the product of the mean norm of the quasigradient ξ^s by the step size ρ_s. When the value of the shift becomes less than the threshold value Q_*, the method stops (steps 3,4). The step size control differs from theoretical one (18.5) in several aspects (step 7). For one thing, value T_s is divided by the averaged absolute value of T_s. Additional reduction of the step size by means of factor U, $0 < U \leq 1$ is introduced. The additional reduction takes place only if

$$T_s = \langle \xi^s, x^{s-1} - x^s \rangle \leq 0.$$

Since T_s/Z_s is some random value, step size ρ_s can increase or decrease, sometimes by too large a factor (step 7). In order that the next step does not differ too strongly from the preceding one ρ_{s-1}, some bounding coefficients are provided for increase or decrease of the step size (step 8).

Recommendations on the choice of the algorithm parameters. The following recommendations are obtained as a result of numerical experiments.

- The value of the mean change of step size $R(1 < R < 3)$ is usually set to $R = 2$.
- The value of the initial step size has no essential effect on the method convergence rate. However, if additional information is available, the initial value of the step size factor ρ_0 can be set approximately

$$\|x^0 - x^*\|(E(\|\xi^0\|))^{-1},$$

  where x^0- initial approximation, x^*- estimated location of extremum point;
- Parameter k defines averaging factor $D = \frac{1}{k}$ in the averaging formulas (Steps 2 and 6). Usually k is selected within the range $4 \leq k \leq 6$
- Parameter U (additional coefficient of step size reduction) is selected within the range $0.8 \leq U \leq 1$. With $k > 1$ coefficient U can be equal to 1 since step size decreases fast without additional decrease.
- The value of mean shift Q_* in stopping criterion is to be set approximately to the required solution accuracy for components of x.

18.5 Results of Numerical Experiments

Let us note firstly that it is advisable to average the values of variables and of the objective function during fixed number of the last iterations and take these quantities as the final approximation to the solution. The averaged value of coordinates x^s will now be designated as x and the averaged value $f(x^s, \omega^s)$ as $f(x)$.

Problem 1. The following problem is an example of multi-commodity facility location problem [7]. It is necessary to minimize

$$F(x) = E\sum_{i=1}^{5} \max\{a_i(x_i - \theta_i); b_i(\theta_i - x_i)\},$$

under constraints

$$\begin{array}{ccccccccccc} x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & + & x_5 & = & 200 \\ x_1 & & & & & & & & & \leq & 50 \\ & & x_2 & & & & & & & \leq & 7 \\ & & & & x_3 & & & & & \leq & 7 \\ & & & & & & x_4 & & & \leq & 80 \\ & & & & & & & & x_5 & \leq & 25 \end{array}$$

$$x_i \geq 0, i = \overline{1,5}.$$

Here θ_i are random values uniformly distributed on intervals $[A_i, B_i]$, $i = 1,\ldots,5$. Vectors $a = (a_1,\ldots,a_5)$, $b = (b_1,\ldots,b_5)$, $A = (A_1,\ldots,A_5)$, $B = (B_1,\ldots,B_5)$ are defined as follows:

$$\begin{array}{rclrcl} A & = & (0,0,0,0,0); & B & = & (60,15,17,90,40); \\ a & = & (1,0,3,1,2); & b & = & (3,4,1,2,3). \end{array}$$

This problem allows analytical solution, which makes it possible to compare solution obtained by algorithm with exact one. The analytical form of the objective function is the following:

$$f(x) = \tfrac{1}{3}x_1^2 + \tfrac{2}{15}x_2^2 + \tfrac{2}{17}x_3^2 + \tfrac{1}{60}x_4^2 + \tfrac{1}{16}x_5^2 - 3x_1 - 4x_2 - x_3 - 2x_4 - 3x_5 + 278.5.$$

Stochastic quasigradient is computed by formula

$$\xi^s = (\xi_1^s,\ldots,\xi_5^s),$$
$$\xi_i^s = \begin{cases} a_1, & \text{if } x_i^s \geq \theta_i^s, \\ -b_1, & \text{if } x_i^s < \theta_i^s, \end{cases} \quad i = 1,\ldots,5$$

The following exact solution was obtained using quadratic programming methods:

$$x^* = (41.88057; 7.00000; 2.48092; 41.27456; 22.33456),$$
$$f(x^*) = 98.100089.$$

Algorithm parameters are

$$R = 1.5; k = 4; U = 0.9; \rho_0 = 1.0.$$

Initial point is

$$x^0 = (0,0,0,0,0); f(x^0) = 278.5.$$

Step size on the 91st iteration $\rho_{91} = 0.1532$.

The results for averaged values of the coordinates and of the function for 91st to 100th iteration are as follows

$$\overline{x}_i = \frac{1}{10}\sum_{s=91}^{100} x_i^s, i = \overline{1,5}; \overline{x}_1 = 40.5485; \overline{x}_2 = 6.9981;$$

$$\overline{x}_j = 2.4381; \overline{x}_4 = 42.2561; \overline{x}_5 = 20.3561;$$

$$\overline{f}(x) = \frac{1}{10}\sum_{s=91}^{100} f(x^s, \theta^s) = 97.4185.$$

For comparison, below are given results of the solution of the same problem using the method with programmed control of step size. Initial approximation was the same. In asymptotically optimal [11] off-line step size rule $\rho_s = 1/\ell(s+a)$, parameter ℓ must be equal to the least eigenvalue of the objective function Hessian, i.e., $\ell = 1/30$. In this case we selected $a = 10$ and got approximately the same performance. However, our choice was based on exact information on objective function. If such information is not available, the off-line decision rule works in a much worse way.

Problem 2. A random locational equilibrium problem (Weber problem) [12]. The classical statement of Weber problem is as follows: given n points $\omega_i, i = \overline{1,n}$ in two-dimensional Euclidean space R^2, find a point $x \in R^2$ which minimizes the sum of distances $\|\omega_i - x\|$. In generalized statement of the problem [12] each point $\omega_i, i = \overline{1,n}$ is considered to be a random variable represented by some probability measure $\theta_i(\omega)$ over R^2. The problem now is to find the location of a point $x \in R^2$ which minimizes the weighted sum of expectations of distances between point x and points $\omega_i, i = \overline{1,n}$, i.e.

$$F(x) = \sum_{i=1}^{n} \beta_i \int\int_{R^2} \|x - \omega\| \theta_i(d\omega) \longrightarrow \min_{x \in R^2},$$

where $\beta_i > 0, i = \overline{1,n}$. The stochastic quasigradient at point x^s can be chosen as follows:

$$\xi^s = \sum_{i=1}^{n} \beta_i \gamma_i$$

where

$$\gamma_i = \begin{cases} \frac{x - \omega_i^s}{\|x - \omega_i^s\|} \\ 0 \end{cases} \quad \text{otherwise}$$

and ω_i^s is distributed according to $\theta_i(\omega)$.

In this particular example, the number of destination points was chosen as $n = 30$, and $\theta_i, i = \overline{1,n}$ were taken as bivariate normal density functions whose means and standard deviations were generated randomly in the range 0–20. The weights β_j were also generated randomly in the range 0–10.

Exact value of the extremum is $x^* = (8,36; 9.36)$ initial approximation $x^0 = (41,87)$. The results for averaged values of the variables x^s for 50-th to 60-th iteration are as follows

$$\overline{x} = \frac{1}{10}\sum_{s=51}^{60} x^s = (9.1, 10.2),$$

and for 190-th to 200-th are

$$\overline{x} = \frac{1}{10}\sum_{s=191}^{200} x^s = (8.9, 9.0).$$

If the initial approximation is $x^0 = (54,30)$. The results for averaged values of the variables x^s for 20-th to 30-th iteration are as follows

$$\overline{x} = \frac{1}{10}\sum_{s=20}^{30} x^s = (8.0, 10.1),$$

and for 190-th to 200-th are

$$\overline{x} = \frac{1}{10}\sum_{s=190}^{200} x^s = (7.9, 9.7)$$

The following table contains detailed description of the problem.

x_1	3.02	6.07	9.77	16.26	6.12	14.80	7.24	7.52	15.91	13.57
means	2.08	12.70	0.16	15.78	3.95	11.89	4.68	6.11	9.19	11.56
	12.43	19.98	15.33	18.20	7.84	1.16	4.54	17.48	10.78	1.45
x_2	7.63	6.62	15.40	10.83	4.85	17.14	2.20	9.30	17.30	14.60
means	5.68	4.77	19.10	17.17	0.80	10.82	11.48	18.99	0.36	2.52
	10.00	1.93	11.39	16.41	16.21	2.09	16.69	8.70	12.04	2.93
x_1	18.65	18.95	0.45	13.50	17.55	1.12	18.42	1.59	15.65	9.49
devs.	19.13	18.19	19.56	19.14	11.93	7.26	1.72	11.37	7.09	16.05
	15.62	4.31	15.44	1.40	5.82	8.56	16.72	5.29	10.36	12.49
x_2	3.77	15.79	8.68	6.29	7.97	9.23	5.81	3.17	17.91	7.02
devs.	16.27	15.08	5.12	6.11	1.55	19.25	8.24	17.78	13.48	9.80
	5.49	15.13	7.07	16.83	15.86	9.90	19.44	16.35	0.37	15.31
weights	8.50	9.48	6.03	8.16	9.05	1.80	8.17	7.57	3.43	9.62
	2.87	3.77	4.34	4.88	0.11	2.13	7.75	1.64	5.75	6.12
	4.57	4.45	2.95	0.17	7.53	9.39	7.38	1.15	2.09	7.20

References

[1] Yu.M. Ermoliev, "Methods of stochastic Programming", Nauka, 1976, p.150 (in Russian).

[2] H. Robbins and S. Monro, "A stochastic approximation method", *Ann. Math. Statist.* **22**(1951), 400–407.

[3] H. Kesten, "Accelerated stochastic approximation", *Ann. Math. Statist.* **29**(1958), 41–59.

[4] G. Pflug, "On the determination of the step size in stochastic quasigradient methods". Collaborative Paper, CP-83-25, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1983.

[5] S.P. Uriasiev, "Step regulation for direct methods of stochastic programming", *Kibernetika* **6**(1980), 85–87 (in Russian).

[6] S.P. Uriasiev, "Adaptive stepsize rules for stochastic optimization methods", Ph.D. thesis. Institute of Cybernetics, Kiev, 1983.

[7] F. Mirzoakhmedov and Urjasév, "Adaptive Step Size Control for Stochastic Optimization Algorithm", *Ih urn. vych.mat.i mat. fisiki*, **6**(1983), 1314–1325 (in Russian).

[8] R.E. Bruck, "On weak convergence of an Ergodic iteration the solution of variational inequalities for Monotone Operators in Hilbert Space".

CHAPTER 19

A NOTE ABOUT PROJECTIONS IN THE IMPLEMENTATION OF STOCHASTIC QUASIGRADIENT METHODS

*R.T. Rockafellar and R.J-B Wets**

Given a stochastic optimization problem find $x \in X \subset R^n$ that minimizes $F(x) = E\{f(x,\xi)\}$ where $f : R^n x \Xi \to R$ is a real-valued function, the quasi-gradient algorithm generates a sequence $\{x^1, x^2, \ldots\}$ of points of X (converging to the optimal solution with probability 1) through the recursion:

$$x^{\nu+1} := \underset{X}{\text{prj}}(x^\nu - \rho_\nu z^\nu)$$

where prj_X denotes the projection on X, $\{\rho_\nu, \nu = 1, \ldots\}$ is a sequence of positive scalars that tend to 0, and z^ν is a stochastic quasi-gradient of F at x^ν; see Chapter 5.

Unless X is a simple convex set, e.g. a rectangle or a ball, the projection operation may be too onerous to allow for a straightforward implementation of the iterative step; one would have to find at each step

$$x^{\nu+1} = \text{argmin}[\text{dist}^2(x^\nu - \rho_\nu z^n, x) | x \in X],$$

which means solving a mathematical program with quadratic objective function. Therefore the implementations of the stochastic quasi-gradient method rely usually on various schemes to bypass this projection operation, through penalization or primal-dual methods, for example. There are however a few cases when it is possible to design a very effective subroutine to perform the projection operation.

We describe a simple method for projecting a point $\hat{y} \in R^n_+$ on a convex set X, assumed to be nonempty, that is the intersection of a rectangle $C \subset R^n$ and a set determined by a single linear or more generally by a separable nonlinear constraint of the type:

$$\sum_{j=1}^{n} a_j(x_j) \leq b, \tag{19.1}$$

where the a_j are convex differentiable functions such that for every $j = 1, \ldots, n$, the derivative a'_j of a $a_j(\cdot)$ is positive and bounded away from zero on C where

$$C := \{x \in R^n | \ell_j \leq x_j \leq u_j, \quad j = 1, \ldots, n\} \tag{19.2}$$

* *Supported in part by grants of the National Science Foundation.*