TWO TIME-SCALE FEASIBLE DIRECTION METHOD

Vladislav B. Tadić, [*] **Nenad Vlajković,** [**]
Efstratios C. Kyriakopoulos [*]

[*] *Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, United Kingdom*
[**] *Department of Electrical Engineering, University of Belgrade, 11000 Belgrade, Serbia and Montenegro*

Abstract: Stochastic constrained optimization problems with non-convex objective and convex feasible domain are considered for the case where the objective and constraint functions are available only throuth noisy observations. A general algorithm of the two time-scale stochastic approximation type is proposed for these problems. The proposed algorithm is applied to Markov decision problems with average cost, average constraints and parameterized randomized policy. The asymptotic behavior of the proposed algorithm is analyzed for the case where the algorithm step-sizes are constant and the noise in the observations of the objective and constraint functions depends on the algorithm iterates.

Keywords: Optimization, stochastic approximation, Markov decision problems, Monte Carlo simulation.

1. INTRODUCTION

Stochastic constrained optimization is typically solved by the following techniques: methods based on the Lagrange multipliers and duality theory, projection and feasible direction methods, and penalty function methods. The method based on the Lagrange multiplies and duality theory exhibit good convergence rates and can handle noise in the observations of the objective and constraint functions. However, these methods require the objective and constraint functions to be convex and admit the strong duality (for details, see (Bertsekas, 1999), (Kushner and Clark, 1978), (Pflug, 1996) and references cited therein). The projection and feasible direction methods also exhibit good convergence rate, but they do not require the objective function to be convex. However, these methods cannot handle noise in the observations of the constraint functions. Moreover, they require the feasible domain (i.e., the constraint functions) to have a simple structure (e.g., to be a ball or hyper-rectangle) in order to determine projections and feasible directions efficiently (for details, see (Bertsekas, 1999), (Kushner and Clark, 1978), (Pflug, 1996) and references cited therein). The penalty function methods can handle noise in the observations of the objective and constraint functions and require none of these functions to be convex. However, these methods usually exhibit poor convergence rates (particularly if the constraints are available only through their noisy observations) and suffer from ill-conditioning(for details, see (Bertsekas, 1999), (Kushner and Clark, 1978), (Pflug, 1996) and references cited therein).

In this paper, stochastic constrained optimization problems with non-convex objective and convex feasible domain are considered for the case where the objective and constraint functions are available only throuth noisy observations. A general algorithm of the two time-scale stochastic approximation type is proposed for these problems. The proposed algorithm is an extension of the existing feasible direction techniques and can be considered as a combination of the Frank-Wolf and gradient recursions (coupled through two time-scale implementation): the Frank-Wolf procedure looks for minima of the objective function, while the gradient recursion 'interprets' feasible search directions as solutions to a subsidiary convex optimization problem and uses Lagrange multipliers to solve it. Compared with the classical feasible direction methods, this algorithm exhibits similar convergence rates at a moderately increased computational complexity. It is successfully applied to Markov decision problems with average cost, average constraints and parameterized randomized policy (for more details on constrained Markov decision problems and their applications see e.g., (Altman, 1999) and references cited therein). The asymptotic behavior of the proposed algorithm is analyzed for the case where the algorithm stepsizes are constant and the noise in the observations of the objective and constraint functions depends on the algorithm iterates.

2. CONSTRAINED OPTIMIZATION PROBLEM

Let $f : R^p \to R$ and $g_1, \ldots, g_q : R^p \to R$ be differentiable functions. The optimization problem we are concerned with in this paper is defined as follows:

$$\begin{aligned} &\text{Minimize } f(x) \\ &\text{Subject to: } g_1(x) \le 0, \ldots, g_q(x) \le 0. \end{aligned} \quad (1)$$

Constrained optimization problem (1) is considered for the following case:

(i) $g_1(\cdot), \ldots, g_q(\cdot)$ are convex, but $f(\cdot)$ can be non-convex.

(ii) $f(\cdot)$, $g_1(\cdot), \ldots, g_q(\cdot)$ and their derivatives are available only through 'noisy' observations, i.e., for each $x \in R^p$, instead of exact values, asymptotically unbiased estimates of $f(x)$, $\nabla f(x)$, $g_1(x), \nabla g_1(x) \ldots, g_q(x), \nabla g_q(x)$ are available.

The following notation is used throughout the paper. $\|\cdot\|$ denotes the Euclidean vector norm. For $\rho \in [0, \infty)$, $k \ge 1$, $\mathcal{B}^k$ is the family of Borel sets from R^k and $B_\rho^k = \{x \in R^k : \|x\| \le \rho\}$. For $x = y = 0$, we use the convention that $x/y = 0$.

3. TWO TIME-SCALE FEASIBLE DIRECTION ALGORITHMS

In this section, we derive two-time scale feasible direction algorithms for the constrained optimization problem (1). First, a two time-scale algorithm is derived for the deterministic counterpart of (1). Then, using the obtained algorithm as a starting point, we derive a two time-scale algorithm for the stochastic version of (1).

3.1 Deterministic Constrained Optimization

Suppose that the functions $f(\cdot)$, $g_1(\cdot), \ldots, g_q(\cdot)$ and their gradients are known (i.e., the values $f(x)$, $\nabla f(x)$ and $g_1(x), \nabla g_1(x), \ldots, g_q(x), \nabla g_q(x)$ are available for each $x \in R^p$). Then, the optimization problem (1) can be solved efficiently by the feasible direction method (also known as the conditional gradient method; for details see (Bertsekas, 1999, Chapter 2) and references cited therein).

Let $\{\alpha_n\}_{n\ge 0}$ be a sequence from $(0, 1)$ (e.g., $\alpha_n = \alpha$ for each $n \ge 0$, or $\alpha_n = 1/(1+n)^a$ for $n \ge 0$, where α is a small positive constant and a is a constant from $(0, 1)$). Moreover, for $x \in R^p$, let

$$h(x) = \underset{g_i(y)\le 0, 1\le i\le q}{\arg\min} (\nabla f(x))^T (y - x). \quad (2)$$

Classical feasible direction algorithms of the Frank-Wolf type are defined by the following difference equations:

$$x_{n+1} = x_n + \alpha_n (y_n - x_n), \quad (3)$$
$$y_n = h(x_n), \quad n \ge 0. \quad (4)$$

Due to the fact that $g_1(\cdot), \ldots, g_q(\cdot)$ are convex, Lagrange multipliers and the duality theory can be used to solve the (subsidary) constrained optimization problem associated to (4) (see e.g., (Bertsekas, 1999, Chapter 5)).

Let $\{\beta_n\}_{n\ge 0}$ be a sequence of positive reals (e.g., $\beta_n = \beta$ for each $n \ge 0$, or $\beta_n = 1/(1+n)^b$ for $n \ge 0$, where β is a small positive constant and b is a constant from $(0, 1)$). Moreover, for $y \in R^p$, let

$$\begin{aligned} G(x) &= [g_1(y) \cdots g_q(y)]^T, \\ \nabla G(x) &= [\nabla g_1(y) \cdots \nabla g_q(y)], \\ \Pi_+(y) &= \underset{z\in[0,\infty)^q}{\arg\min} \|y - z\|, \end{aligned}$$

while

$$L_n(y, \lambda) = (\nabla f(x_n))^T (y - x_n) + G^T(y)\lambda$$

for $y \in R^p$, $\lambda \in R^q$, $n \ge 1$ ($L_n(\cdot, \cdot)$ is the Lagrangian corresponding to the subsidary optimization problem associated to (4)). Then, owing to the duality theory, the solution y_n of (2), (4)

can be estimated using a gradient search for a saddle point of $L_n(\cdot,\cdot)$:

$$\begin{aligned} y_{j+1}^n =& y_j^n - \beta_j \nabla_y L_n(y_j^n, \lambda_j^n) \\ =& y_j^n - \beta_j(\nabla f(x_n) + \nabla G(y_j^n)\lambda_j^n), \quad (5) \\ \lambda_{j+1}^n =& \Pi_+(\lambda_j^n + \beta_j \nabla_\lambda L_n(y_j^n, \lambda_j^n)) \\ =& \Pi_+(\lambda_j^n + \beta_j G(y_j^n)), \quad 1 \le j \le N, \quad (6) \end{aligned}$$

where N is the number of the executions of the 'inner' recursion (5), (6), while the initial values y_0^n, λ_0^n can be selected arbitrary from R^p, $[0,\infty)^q$ (natural choice would be $y_0^{n+1} = y_N^n$, $\lambda_0^{n+1} = \lambda_N^n$). Noticing that $x_n \approx x_{n+1}$ (due to $\alpha_n \approx 0$ for sufficiently large n), we can take $N = 1$. Then, we get the following recursive algorithm for the constrained optimization problem (1):

$$\begin{aligned} x_{n+1} =& x_n + \alpha_n(y_n - x_n), \quad (7) \\ y_{n+1} =& y_n - \beta_n(\nabla f(x_n) + \nabla G(y_n)\lambda_n), \quad (8) \\ \lambda_{n+1} =& \Pi_+(\lambda_n + \beta_n G(y_n)), \quad n \ge 0. \quad (9) \end{aligned}$$

3.2 Stochastic Constrained Optimization

Suppose that instead of exact values of $\nabla f(\cdot)$ and $g_1(\cdot), \nabla g_1(\cdot), \ldots, g_q(\cdot), \nabla g_q(\cdot)$, we have for any $x \in R^p$ their estimates $A(x)$, $B(x)$, $C(x)$, where $A : R^p \to R^p$, $B : R^p \to R^{p\times q}$, $C : R^p \to R^q$ are random functions. Then, provided that the estimates are unbiased (i.e., $E(A(x)) = \nabla f(x)$, $E(B(x)) = \nabla G(x)$, $E(C(x)) = G(x)$ for all $x \in R^p$), a reasonable way to get an algorithm for the constrained optimization problem (1) is to substitute $\nabla f(x_n)$, $\nabla G(y_n)$, $G(y_n)$ in (7) – (9) with their estimates $A(x_n)$, $B(y_n)$, $C(y_n)$. As a result, we get the following algorithm for the stochastic counterpart of (1):

$$\begin{aligned} X_{n+1} =& X_n + \alpha_n(Y_n - X_n), \quad (10) \\ Y_{n+1} =& Y_n - \beta_n(A(X_n) + B(Y_n)\lambda_n), \quad (11) \\ \lambda_{n+1} =& \Pi_+(\lambda_n + \beta_n C(Y_n))), \quad n \ge 0, \quad (12) \end{aligned}$$

Remark. In order to emphasise that the iterates (i.e., states) of the algorithm (10) – (12) are random variables, they are denoted with capital letters.

Setting

$$\begin{aligned} U_n =& A(X_n) - \nabla f(X_n), \\ V_n =& B(X_n) - \nabla G(X_n), \\ W_n =& C(X_n) - G(X_n) \end{aligned}$$

for $n \ge 0$, the algorithm (10) – (12) can be represented in the form of (two time-scale) stochastic approximation (for details see (Borkar, 1997), (Kushner and Yin, 1997) and references cited therein):

$$\begin{aligned} X_{n+1} =& X_n + \alpha_n(Y_n - X_n), \quad (13) \\ Y_{n+1} =& Y_n - \beta_n(\nabla f(X_n) + \nabla G(Y_n)\lambda_n \\ & + U_n + V_n\lambda_n), \quad (14) \\ \lambda_{n+1} =& \Pi_+(\lambda_n + \beta_n(G(Y_n) + W_n))), \quad n \ge 0. \\ & (15) \end{aligned}$$

Random variable U_n, V_n, W_n themselves can be interpreted as 'observation noise' in the estimates of $\nabla f(X_n)$, $\nabla G(Y_n)$, $G(Y_n)$, respectively.

If the sequences $\{\alpha_n\}_{n\ge0}$ and $\{\beta_n\}_{n\ge0}$ (the step-sizes of the algorithm (13) – (15)) are selected such that $\alpha_n = \alpha$, $\beta_n = \beta$ for all $n \ge 0$, and $\alpha \ll \beta$ (the case of constant step-sizes), or such that $\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \beta_n = 0$ and $\alpha_n = o(\beta_n)$ (the case of decreasing step-sizes), the algorithm (13) – (15) asymptotically behaves similarly as singularly perturbed ordinary differential equations (see e.g., (O'Malley, 1991) and references cited therein). More specifically, the algorithm (13) – (15) consists of two sub-recursions: slow (13) and fast one (14), (15). The iterates $\{X_n\}_{n\ge0}$ of the slow sub-recursion (13) are updated with smaller step-sizes ($\{\alpha_n\}_{n\ge0}$) and evolve on a slower time-scale. On the other hand, larger step-sizes ($\{\beta_n\}_{n\ge0}$) are used for updating the iterates $\{Y_n\}_{n\ge0}, \{\lambda_n\}_{n\ge0}$ of the fast sub-recursion (14), (15), while their evolution is characterized by a faster time-scale. Moreover, the fast sub-recursion (14), (15) sees the iterates of the slow one as 'static', while the slow sub-recursion (13) sees the iterates of the fast one as 'equilibrated'. Since $h(x)$ would be a globally stable equilibrium of (14) (under mild conditions; see A3, next section) if X_n were constant and equal to x, we can conclude that Y_n tracks $h(X_n)$ (i.e., $Y_n \approx h(X_n)$ asymptotically as $n \to \infty$).

4. ASYMPTOTIC RESULTS

In this section, the asymptotic analysis of the algorithm derived in the previous section is carried out for the case where the algorithm step-sizes are constant and the observation noise $\{U_n\}_{n\ge0}$, $\{V_n\}_{n\ge0}$, $\{W_n\}_{n\ge0}$ is state-dependent (i.e., for each $n \ge 0$, U_n, V_n, W_n are random functions of $X_0, Y_0, \lambda_0, \ldots, X_n, Y_n, \lambda_n$).

Let $\alpha_n = \alpha$, $\beta_n = \beta$ for all $n \ge 0$, where α, β are small positive constants. In that case, the algorithm iterates (i.e., states) $\{X_n\}_{n\ge1}$, $\{Y_n\}_{n\ge1}$, $\{\lambda_n\}_{n\ge1}$, as well as the observation noise $\{U_n\}_{n\ge0}$, $\{V_n\}_{n\ge0}$, $\{W_n\}_{n\ge0}$ depend on the step-sizes α, β (notice that the initial values X_0, Y_0, λ_0 do not depend on α, β). In order to emphasize this fact, the following notation is used in this section: With the exception of the initial iterates, α and β appear in the superscripts of all variables of the algorithm (13) – (15), i.e., for $n \ge 1$, $X_n^{\alpha,\beta}$, $Y_n^{\alpha,\beta}$, $\lambda_n^{\alpha,\beta}$ denote X_n, Y_n, λ_n, and for $n \ge 0$, $U_n^{\alpha,\beta}$,

$V_n^{\alpha,\beta}$, $W_n^{\alpha,\beta}$ stand for U_n, V_n, W_n (notice that the notation for X_0, Y_0, λ_0 remains unchanged).

Let $\mathcal{F}_n^{\alpha,\beta} = \sigma\{U_i^{\alpha,\beta}, V_i^{\alpha,\beta}, W_i^{\alpha,\beta} : 0 \le i \le n\}$ for $n \ge 0$. The algorithm (13) – (15) is analyzed under the following assumptions:

A1. For all $\alpha, \beta \in (0,\infty)$, there exist R^p-valued stochastic processes $\{U_{1,n}^{\alpha,\beta}\}_{n\ge0}$, $\{U_{2,n}^{\alpha,\beta}\}_{n\ge0}$, $\{U_{3,n}^{\alpha,\beta}\}_{n\ge0}$, $R^{p\times q}$-valued stochastic processes $\{V_{1,n}^{\alpha,\beta}\}_{n\ge0}$, $\{V_{2,n}^{\alpha,\beta}\}_{n\ge0}$, $\{V_{3,n}^{\alpha,\beta}\}_{n\ge0}$, and R^q-valued stochastic processes $\{W_{1,n}^{\alpha,\beta}\}_{n\ge0}$, $\{W_{2,n}^{\alpha,\beta}\}_{n\ge0}$, $\{W_{3,n}^{\alpha,\beta}\}_{n\ge0}$, such that for $n \ge 0$,

$$\begin{aligned} U_{n+1}^{\alpha,\beta} =& U_{1,n+1}^{\alpha,\beta} + U_{2,n+1}^{\alpha,\beta} + U_{3,n+1}^{\alpha,\beta} - U_{3,n}^{\alpha,\beta}, \\ U_{n+1}^{\alpha,\beta} =& V_{1,n+1}^{\alpha,\beta} + V_{2,n+1}^{\alpha,\beta} + V_{3,n+1}^{\alpha,\beta} - U_{3,n}^{\alpha,\beta}, \\ W_{n+1}^{\alpha,\beta} =& W_{1,n+1}^{\alpha,\beta} + W_{2,n+1}^{\alpha,\beta} + W_{3,n+1}^{\alpha,\beta} - U_{3,n}^{\alpha,\beta}. \end{aligned}$$

A2. For all $\rho \in [1,\infty)$, there exist constants $\delta_\rho \in (0,1)$, $K_\rho \in [1,\infty)$ (not depending on α, β) such that for all α, β, $n \ge 0$, $1 \le i \le 3$,

$$\begin{aligned} &E(U_{1,n+1}^{\alpha,\beta} I_{\{\tau_\rho^{\alpha,\beta} > n\}} | \mathcal{F}_n^{\alpha,\beta}) = 0 \ w.p.1, \\ &E(V_{1,n+1}^{\alpha,\beta} I_{\{\tau_\rho^{\alpha,\beta} > n\}} | \mathcal{F}_n^{\alpha,\beta}) = 0 \ w.p.1, \\ &E(W_{1,n+1}^{\alpha,\beta} I_{\{\tau_\rho^{\alpha,\beta} > n\}} | \mathcal{F}_n^{\alpha,\beta}) = 0 \ w.p.1, \\ &E(\|U_{i,n}^{\alpha,\beta}\|^2 I_{\{\tau_\rho^{\alpha,\beta} \ge n\}}) \le K_\rho, \\ &E(\|V_{i,n}^{\alpha,\beta}\|^2 I_{\{\tau_\rho^{\alpha,\beta} \ge n\}}) \le K_\rho, \\ &E(\|W_{i,n}^{\alpha,\beta}\|^2 I_{\{\tau_\rho^{\alpha,\beta} \ge n\}}) \le K_\rho, \\ &E(\|U_{2,n}^{\alpha,\beta}\| I_{\{\tau_\rho^{\alpha,\beta} \ge n\}}) \le K_\rho(\alpha+\beta), \\ &E(\|V_{2,n}^{\alpha,\beta}\| I_{\{\tau_\rho^{\alpha,\beta} \ge n\}}) \le K_\rho(\alpha+\beta), \\ &E(\|W_{2,n}^{\alpha,\beta}\| I_{\{\tau_\rho^{\alpha,\beta} \ge n\}}) \le K_\rho(\alpha+\beta), \end{aligned}$$

where

$$\begin{aligned} \tau_\rho^{\alpha,\beta} =& \tau_{1,\rho}^{\alpha,\beta} \wedge \tau_{2,\rho}^{\alpha,\beta} \\ \tau_{1,\rho}^{\alpha,\beta} =& \inf(\{n \ge 0 : \|X_n^{\alpha,\beta}\| \vee \|Y_n^{\alpha,\beta}\| \\ & \vee \|\lambda_n^{\alpha,\beta}\| > \rho\} \cup \{\infty\}), \\ \tau_{2,\rho}^{\alpha,\beta} =& \inf(\{n \ge 1 : \|X_n^{\alpha,\beta} - X_{n-1}^{\alpha,\beta}\| \\ & \vee \|Y_n^{\alpha,\beta} - Y_{n-1}^{\alpha,\beta}\| \\ & \vee \|\lambda_n^{\alpha,\beta} - \lambda_{n-1}^{\alpha,\beta}\| > \delta_\rho\} \cup \{\infty\}). \end{aligned}$$

A3. $g_1(\cdot), \ldots, g_q(\cdot)$ are convex, and there exists $x \in R^p$ such that $g_i(x) < 0$ for each $1 \le i \le q$.

Remark. A1 and A2 are standard noise conditions for the asymptotic analysis of stochastic approximation algorithms (see e.g., (Benveniste *et al.*, 1990), (Kushner and Yin, 1997) and references cited therein). A3 is basically the Slater constraint qualification condition (see e.g., (Bertsekas, 1999, Section 5.3)). It ensures that there is no duality gap in the subsidiary optimization problems (2).

Let $\bar{X}_0^\alpha = X_0$ and

$$\bar{X}_{n+1}^\alpha = \bar{X}_n^\alpha + \alpha(h(\bar{X}_n^\alpha) - \bar{X}_n^\alpha) \tag{16}$$

for $n \ge 0$.

As a main result on the asymptotic behavior of the algorithm (13) – (15), the following theorem is obtained:

Theorem 1. Let A1 – A3 hold. Suppose that $\nabla f(\cdot)$ and $\nabla g_1(\cdot), \ldots, g_q(\cdot)$ are locally Lipschitz continuous. Then,

$$\lim_{\substack{\alpha,\beta\to0 \\ \alpha/\beta,\beta^2/\alpha\to0}} \mathcal{P}\left(\sup_{0\le n\le t/\alpha} \|X_n^{\alpha,\beta} - \bar{X}_n^\alpha\| \ge \delta\right) = 0$$

for all $\delta, t \in (0,\infty)$.

For the proof, see (Tadić *et al.*, 2004).

Remark. Theorem 1 basically claims that the iterates $\{X_n^{\alpha,\beta}\}_{n\ge0}$ of the algorithm (13) – (15) asymptotically behave as the classical feasible direction algorithm (3), (4).

5. MARKOV DECISION PROBLEMS WITH AVERAGE COST AND AVERAGE CONSTRAINTS

In this section, Markov decision problems with average cost, average constraints and parameterized randomized policy is considered. Using the general results obtained in Sections 3, 4, we develop and analyze simulation based (i.e., Monte-Carlo) algorithms for this class of Markov decision problems.

Let $\mu(\cdot)$, $\nu(\cdot)$ be a non-negative measure on $(R^r, \mathcal{B}^r)$, $(R^s, \mathcal{B}^s)$, respectively. Moreover, let $p : R^r \times R^s \times R^r \to [0,\infty)$, $q : R^p \times R^r \times R^s \to [0,\infty)$ be Borel-measurable functions such that $\int p(\xi,\zeta,\xi')\mu(d\xi') = \int q(x,\xi,\zeta')\nu(d\zeta') = 1$ for all $x \in R^p$, $\xi \in R^r$, $\zeta \in R^s$. A Markov controlled processes with a parameterized randomized stationary policy can be defined as a parameterized $R^r \times R^s$-valued Markov chain $\{(\xi_n^x, \zeta_n^x)\}_{n\ge0}$ ($x \in R^p$ is the parameter) satisfying

$$\begin{aligned} &\mathcal{P}(\xi_{n+1}^x \in B_\xi | \xi_0^x, \ldots, \xi_n^x, \zeta_0^x, \ldots, \zeta_n^x) \\ &= \int_{B_\xi} p(\xi_n^x, \zeta_n^x, \xi)\mu(d\xi) \quad w.p.1, \\ &\mathcal{P}(\zeta_{n+1}^x \in B_\zeta | \xi_0^x, \ldots, \xi_n^x, \xi_{n+1}^x, \zeta_0^x, \ldots, \zeta_n^x) \\ &= \int_{B_\zeta} q(x, \xi_{n+1}^x, \zeta)\nu(d\zeta) \quad w.p.1 \end{aligned}$$

for all $x \in R^p$, $B_\xi \in \mathcal{B}^r$, $B_\zeta \in \mathcal{B}^s$, $n \ge 0$.

Let $\phi : R^r \times R^s \to R$ and $\psi_1, \ldots, \psi_q : R^r \times R^s \to R$ be Borel-measurable functions, while

$$\begin{aligned} f^n(x) =& E(\phi(\xi_n^x, \zeta_n^x)), \\ g_i^n(x) =& E(\psi_i(\xi_n^x, \zeta_n^x)) \end{aligned} \tag{17}$$

for $x \in R^p$, $1 \le i \le q$, $n \ge 0$. Suppose that

$$f(x) = \lim_{n\to\infty} f^n(x), \tag{18}$$

$$g_i(x) = \lim_{n\to\infty} g_i^n(x) \tag{19}$$

are well-defined and finite for all $x \in R^p$, $1 \le i \le q$ (which holds if $\{(\xi_n^x, \zeta_n^x)\}_{n\ge 0}$ is geometrically ergodic for all $x \in R^p$). Markov decision problems with average cost, average constraints and parameterized randomized policy can be defined as the optimization problem (1) with $f(\cdot)$ and $g_1(\cdot), \ldots, g_q(\cdot)$ defined in (18), (19).

In order to apply the algorithm (13) – (15) to the previously described Markov decision problem, we need to find estimates for $\nabla f(\cdot)$ and $\nabla g_1(\cdot), \ldots, \nabla g_q(\cdot)$. It is straightforward to verify that

$$\nabla f^n(x) = E\left(\phi(\xi_n^x, \zeta_n^x)\left(\sum_{i=1}^{n} \frac{\nabla_x q(x, \xi_i^x, \zeta_i^x)}{q(x, \xi_i^x, \zeta_i^x)}\right)\right), \tag{20}$$

$$\nabla g_i^n(x) = E\left(\psi_i(\xi_n^x, \zeta_n^x)\left(\sum_{i=1}^{n} \frac{\nabla_x q(x, \xi_i^x, \zeta_i^x)}{q(x, \xi_i^x, \zeta_i^x)}\right)\right) \tag{21}$$

for all $\theta \in R^p$, $1 \le i \le q$, $n \ge 0$ (under some regularity conditions; see B1 – B3 below).

For $\xi \in R^r$, $\zeta \in R^s$, let

$$\Psi(\xi, \zeta) = [\psi_1(\xi, \zeta) \ldots \psi_q(\xi, \zeta)]^T.$$

Moreover, for $x \in R^p$, $\gamma \in (0, 1)$, let $S_0^{x,\gamma} = 0$, while

$$S_{n+1}^{x,\gamma} = (1-\gamma) S_n^{x,\gamma} + \frac{\nabla_x q(x, \xi_{n+1}^x, \zeta_{n+1}^x)}{q(x, \xi_{n+1}^x, \zeta_{n+1}^x)}$$

for $n \ge 0$. Owing to (17), (19), $\psi_i(\xi_n^x, \zeta_n^x)$ is an asymptotically unbiased estimate of $g_i(x)$ for all $x \in R^p$, $1 \le i \le q$, while (20), (21) imply that for all $x \in R^p$, $1 \le i \le q$, $\phi(\xi_n^x, \zeta_n^x) S_n^{x,\gamma}$, $\psi_i(\xi_n^x, \zeta_n^x) S_n^{x,\gamma}$ are asymptotically unbiased estimates of $\nabla f(x)$, $\nabla g_i(x)$ as $n \to \infty$, $\gamma \to 0$. This (together with the results of Section 3) suggests the following simulation based (i.e., Monte-Carlo) algorithm for the Markov decision problem described above:

$$X_{n+1} = X_n + \alpha_n (Y_n - X_n), \tag{22}$$

$$Y_{n+1} = Y_n - \beta_n (S_n \phi(\xi_n, \zeta_n) + \tilde{S}_n \Psi^T(\tilde{\xi}_n, \tilde{\zeta}_n) \lambda_n), \tag{23}$$

$$\lambda_{n+1} = \Pi_+(\lambda_n + \beta_n \Psi(\tilde{\xi}_n, \tilde{\zeta}_n)), \tag{24}$$

$$S_{n+1} = (1-\gamma) S_n + \frac{\nabla_x q(X_n, \xi_{n+1}, \zeta_{n+1})}{q(X_n, \xi_{n+1}, \zeta_{n+1})}, \tag{25}$$

$$\tilde{S}_{n+1} = (1-\gamma) \tilde{S}_n + \frac{\nabla_x q(Y_n, \tilde{\xi}_{n+1}, \tilde{\zeta}_{n+1})}{q(Y_n, \tilde{\xi}_{n+1}, \tilde{\zeta}_{n+1})}. \tag{26}$$

In the difference equations (22) – (26), $\{\alpha_n\}_{n\ge 0}$, $\{\beta_n\}_{n\ge 0}$ have the same meaning as in the algorithm (13) – (14), while $\gamma \in (0, 1)$ is a constant. Moreover, for each $n \ge$, ζ_n, $\tilde{\zeta}_n$ are samples from $q(X_n, \xi_n, \cdot)$, $q(Y_n, \tilde{\xi}_n, \cdot)$ (respectively) drawn independently from $\xi_0, \tilde{\xi}_0, \ldots, \xi_{n-1}, \tilde{\xi}_{n-1}$, $\zeta_0, \tilde{\zeta}_0, \ldots, \zeta_{n-1}, \tilde{\zeta}_{n-1}$, while ξ_{n+1}, $\tilde{\xi}_{n+1}$ are samples from $p(\xi_n, \zeta_n, \cdot)$, $p(\tilde{\xi}_n, \tilde{\zeta}_n, \cdot)$ (respectively) drawn independently from $\xi_0, \tilde{\xi}_0, \ldots, \xi_{n-1}, \tilde{\xi}_{n-1}$, $\zeta_0, \tilde{\zeta}_0, \ldots, \zeta_{n-1}, \tilde{\zeta}_{n-1}$.

$\{\alpha_n\}_{n\ge 0}$, $\{\beta_n\}_{n\ge 0}$ are the step-sizes in the algorithm (22) – (26). γ can be considered as a forgetting factor in (25), (25) ensuring the stability of $\{S_n\}_{n\ge 0}$, $\{\tilde{S}_n\}_{n\ge 0}$, while for $n \ge 0$, $S_n \phi(\xi_n, \zeta_n)$, $\Psi(\tilde{\xi}_n, \tilde{\zeta}_n)$, $\tilde{S}_n \Psi^T(\tilde{\xi}_n, \tilde{\zeta}_n)$ are estimators of $\nabla f(X_n)$, $G(Y_n)$, $\nabla G(Y_n)$ (respectively).

The asymptotic behavior of the algorithm (22) – (26) is analyzed for the case when $\alpha_n = \alpha$, $\beta_n = \beta$ for each $n \ge 0$, where α, β are small positive constants. In that case, the algorithm iterates $\{X_n\}_{n\ge 1}$, $\{Y_n\}_{n\ge 1}$, $\{\lambda_n\}_{n\ge 1}$ depend on the step-sizes α, β and the forgetting factor γ (notice that the initial values X_0, Y_0, λ_0 do not depend on α, β, γ). In order to emphasize this fact, the following notation is used in the rest of the section: With the exception of the initial ones, α, β, γ appear in the superscripts of all algorithm iterates, i.e., for $n \ge 1$, $X_n^{\alpha,\beta,\gamma}$, $Y_n^{\alpha,\beta,\gamma}$, $\lambda_n^{\alpha,\beta,\gamma}$ denote X_n, Y_n, λ_n (notice that the notation for the initial iterates X_0, Y_0, λ_0 remains unchanged).

The algorithm (22) – (26) is analyzed under the following assumptions:

B1. For all $x \in R^p$, $\{(\xi_n^x, \zeta_n^x)\}_{n\ge 0}$ has a unique invariant probability measure $\tilde{\pi}(x, \cdot)$. There exists a Borel-measurable function $\tilde{p} : R^p \times R^r \times R^s \to [0, \infty)$ such that $\tilde{p}(\cdot, \xi, \zeta)$ is differentiable for all $\xi \in R^r$, $\zeta \in R^s$, and

$$\tilde{\pi}(x, B) = \int I_B(\xi, \zeta) \tilde{p}(x, \xi, \zeta) \mu(d\xi) \nu(d\zeta)$$

for all $x \in R^p$, $B \in \mathcal{B}^{r+s}$.

B2. For all $\rho \in [1, \infty)$, there exists a Borel-measurable function $\varphi_\rho : R^{r+s} \to [1, \infty)$ such that

$$\begin{aligned}
&\max\{|\phi(\xi, \zeta)|, |\psi_i(\xi, \zeta)\|\} \le \varphi_\rho^{1/4}(\xi, \zeta),\\
&\max\left\{\left\|\frac{\nabla_x \tilde{p}(x, \xi, \zeta)}{\tilde{p}(x, \xi, \zeta)}\right\|, \left\|\frac{\nabla_x q(x, \xi, \zeta)}{q(x, \xi, \zeta)}\right\|\right\}\\
&\le \varphi_\rho^{1/4}(\xi, \zeta),\\
&\left\|\frac{\nabla_x \tilde{p}(x', \xi, \zeta)}{\tilde{p}(x', \xi, \zeta)} - \frac{\nabla_x \tilde{p}(x'', \xi, \zeta)}{\tilde{p}(x'', \xi, \zeta)}\right\|\\
&\le \varphi_\rho^{1/4}(x)\|x' - x''\|,\\
&\left\|\frac{\nabla_x q(x', \xi, \zeta)}{q(x', \xi, \zeta)} - \frac{\nabla_x q(x'', \xi, \zeta)}{q(x'', \xi, \zeta)}\right\|\\
&\le \varphi_\rho^{1/4}(x)\|x' - x''\|
\end{aligned}$$

for all $x, x', x'' \in R^p$, $\xi \in R^r$, $\zeta \in R^s$, $1 \le i \le q$.

B3. For all $\rho \in [1, \infty)$, there exist constants $r_\rho \in (0, 1)$, $K_\rho \in [1, \infty)$ such that

$$\Big| E(\varphi(\xi_n^x, \zeta_n^x)|\xi_0^x = \xi, \zeta_0^x = \zeta)$$
$$- \int \varphi(\xi', \zeta')\tilde{p}(x, \xi', \zeta')\mu(d\xi')\nu(d\zeta') \Big|$$
$$\leq K_\rho r_\rho^n \varphi_\rho(\xi, \zeta)$$

for all $x \in B^p$, $\xi \in R^r$, $\zeta \in R^s$, $n \geq 0$, and any Borel-measurable function $\varphi : R^r \times R^s \to R$ satisfying $0 \leq \varphi(\xi, \zeta) \leq \varphi_\rho(\xi, \zeta)$ for all $\xi \in R^r$, $\zeta \in R^s$.

Remark. B1 – B3 imply that $f(\cdot)$, $g_1(\cdot), \ldots, g_q(\cdot)$ are well-defined, finite and differentiable.

Let $\{\bar{X}_n^\alpha\}_{n\geq 0}$ has the same meaning as in Section 4. As a main result on the asymptotic behavior of the algorithm (22) – (26), the following theorem is obtained:

Theorem 2. Let B1 – B3 hold. Suppose that A3 is satisfied with $g_1(\cdot), \ldots, g_q(\cdot)$ defined in (19). Then,

$$\lim_{\substack{\alpha,\beta,\gamma\to 0 \\ \alpha/\beta,\beta^2/\alpha\to 0}} \mathcal{P}\left(\sup_{0\leq n\leq t/\alpha} \|X_n^{\alpha,\beta,\gamma} - \bar{X}_n^\alpha\| \geq \delta \right) = 0$$

for all $\delta, t \in (0, \infty)$.

For the proof, see (Tadić *et al.*, 2004).

REFERENCES

Altman, E. (1999). *Constrained Markov Decision Processes*. Chapman & Hall.

Benveniste, A., M. Metivier and P. Priouret (1990). *Adaptive Algorithms and Stochastic Approximation*. Springer Verlag.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific.

Borkar, V. K. (1997). Stochastic approximation with two time scales. *Systems and Control Letters* **29**, 291–294.

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

Kushner, H. J. and G. G. Yin (1997). *Stochastic Approximation Algorithms and Applications*. Springer.

O'Malley, R. E. (1991). *Singular Perturbations Methods for Ordinary Differential Equations*. Springer.

Pflug, G. C. (1996). *Optimization of Stochastic Models: The Interface between Simulation and Optimization*. Kluwer.

Tadić, V. B., N. Vlajković and E. C. Kyriakopoulos (2004). Two time-scale feasible direction method. *to be submitted*.