Proceedings of the 8th International
Workshop on Discrete Event Systems
Ann Arbor, Michigan, USA, July 10-12, 2006

Adaptive stepsize selection for online Q-learning in a non-stationary environment

Kim Levy, Felisa J. Vázquez-Abad and Andre Costa

Abstract— **We consider the problem of real-time control of a discrete-time Markov decision process (MDP) in a non-stationary environment, which is characterized by large, sudden changes in the parameters of the MDP. We consider here an online version of the well-known Q-learning algorithm, which operates directly in its target environment. In order to track changes, the stepsizes (or learning rates) must be bounded away from zero. In this paper, we show how the theory of constant stepsize stochastic approximation algorithms can be used to motivate and develop an adaptive stepsize algorithm, that is appropriate for the online learning scenario described above. Our algorithm automatically achieves a desirable balance between accuracy and rate of reaction, and seeks to track the optimal policy with some pre-determined level of confidence.**

I. Introduction

The Q-learning algorithm is a popular and well-established tool for solving Markov decision problems (MDPs) [4] when the state transition probabilities and reward structure are not known explicitly. It estimates the average rewards associated with different actions at each state via simulation, or by performing direct measurements on the system to be controlled. An estimate of the optimal policy is then constructed using the maximum (estimated) average reward.

In this paper, we address theoretical and practical issues surrounding Q-learning in the case where

1) the algorithm is implemented *online*, and
2) the environment is *non-stationary*.

In particular, we consider environments where the MDP parameters are subject to sudden, large changes – a situation referred to as *regime switching* [9]. We assume that regime changes occur on a slow time scale compared with the state transitions of the MDP (we formalize this later in Assumption 1), however, we do not require that they be rare events.

Online Q-learning is consistent with the case of a single learning agent operating in its actual target environment. In this context, the learner does not have the benefit of a separate training phase, and every action that is taken by the learning agent incurs a real cost or reward. Furthermore, although the agent can partially influence its own trajectory by adopting different action selection strategies, its state trajectory is nonetheless governed by the MDP. This contrasts with more common implementations of Q-learning, where it is assumed that the algorithm designer can choose to evaluate several or all state-action pairs synchronously, or in some pre-specified order [1], [5].

We note that a great deal is known about the convergence properties of Q-learning in a stationary environment [1]; in particular, researchers have made extensive use of the theory of stochastic approximation algorithms in order to establish almost sure convergence results when the learning rates, or stepsize parameters, approach zero at an appropriate rate. On the other hand, it is also well-known that in order to track changes in a non-stationary environment, the stepsize parameters must be bounded away from zero, and therefore, the latter results do not hold [1]. In response to this situation, the primary aim of this paper is to address the questions of (a) which theoretical results from the field of stochastic approximation are applicable to Q-learning for the case of non-decreasing stepsizes, as is necessary for tracking, and (b) how this theory can be used to inform the design of adaptive stepsize selection schemes.

Our work differs from existing literature on adaptive stepsize schemes for tracking (see [2] and references therein), in that we consider a non-stationary environment comprising sudden and discontinuous changes to the system parameters, rather than slow time-variations. In this respect, the environment that we consider is closer in spirit to the regime switching scenario presented in [8], [9] (however, the latter employ a fixed rather than adaptive stepsize in their tracking algorithms). Furthermore, we do not assume the existence of a stationary process governing the parameter changes. For example, in Section 3.2 of [2], the system parameters are assumed to change gradually via frequent, small, independent and identically distributed normal increments, and in [8], [9], the regime changes are governed by a modulating hidden Markov model. In contrast, we do not make any assumptions regarding regime changes other than that regime changes occur on a slow time scale compared with the state transitions of the MDP (see Assumption 1). As such, our algorithm is not designed to attain stepsizes that are optimal, in the long-run average sense, with respect an underlying modulating process. Instead, our algorithm is designed to anticipate (arbitrary) regime changes by always selecting the largest possible stepsizes for the current regime, as this **maximises the rate of adaptation** when a regime change eventually does occur. However, this maximisation is **subject to constraints on the probability of incorrectly identifying the optimal actions**.

The paper is structured as follows. In Section II, we formulate the problem. In Section III we describe and

Kim Lévy and Felisa J. Vázquez-Abad are with the Department of Mathematics and Statistics, University of Melbourne, 3010, Australia {k.levy,vazquez}@ms.unimelb.edu.au

Andre Costa is with the ARC Centre of Excellence for Mathematics and Statistics of Complex Systems, University of Melbourne, 3010, Australia acosta@ms.unimelb.edu.au

1-4244-0053-8/06/$20.00 ©2006 IEEE

summarize the main known results for Q-learning in a stationary environment. In Section IV, we establish new convergence results for the case of a constant stepsize in a stationary environment. These results form the basis for an adaptive stepsize scheme for the non-stationary environment, which we present in Section V. Numerical experiments are presented in Section VI, and we conclude with a discussion in Section VII.

II. A Non-Stationary MDP Formulation

Consider a countable collection Ψ of maximum reward Markov decision problems, which share a common finite state and action set, and the same zero-reward absorbing state. The MDPs $\psi \in \Psi$ differ from one another only in the values of the state transition probabilities and rewards (see below). We associate each $\psi \in \Psi$ with a *regime*, and we introduce a general framework for non-stationarity in the form of regime switching, where changes occur on a time scale that is slow compared with the state transitions of the MDP (see Assumption 1 below). Furthermore, we assume that the system parameters associated with different regimes differ significantly from one another, as opposed to the *quasi-static* case [2], which is characterized by infrequent and small changes in the system parameters. Specifically, letting $k \in \mathbb{N}$ be a discrete-time index, we let $\psi_k \in \Psi$ denote the regime (MDP) which is active at time k. We note that unlike in [9], we do not assume that the "modulating state" ψ_k is a Markov process; we do not even require that the process which governs the sequence ψ_k be stationary.

Let the set of states common to all $\psi \in \Psi$ be denoted $\mathcal{S} = \{1, \ldots, S\}$, with S being the absorbing state, and let $\mathcal{A} = \{1, \ldots, A\}$, denote the set of all possible actions. For each non-absorbing state i, there is a subset of $\mathcal{A}$ containing the actions which are available at state i. These action sets are labeled $\mathcal{U}(i), i \in \mathcal{S} \setminus S$. As described above, these quantities are not time-dependent. On the other hand, the state transition probabilities and rewards *are* dependent on the current regime; let $p_{\psi_k}(i,u,j)$ be the probability of making a transition from state i to state $j \in \mathcal{S}$, when the action u is selected, under the current regime ψ_k. Similarly, let $r_{\psi_k}(i,u,j)$ be the single-stage reward incurred in a transition from state i to j when action u is selected under ψ_k.

For each regime $\psi \in \Psi$, and hence for each $\psi_k, k \in \mathbb{N}$, there is a corresponding optimal cost and optimal policy. A policy is a rule for selecting actions at each state [4], the most general case of which is a randomized policy, defined as follows: a randomized policy, Φ, is comprised of a set of probability distributions, such that each state $i \in \mathcal{S} \setminus S$, is associated with a probability distribution $\phi(i)$, which gives the probabilities $\phi(i,u)$ of selecting each action $u \in \mathcal{U}(i)$. Given a fixed policy Φ and a fixed regime $\psi \in \Psi$, the state evolution is a Markov process, X_k, with transition probabilities

$$P(X_{k+1} = j \mid X_k = i) = \sum_{u \in \mathcal{U}_i} \phi(i,u)\, p_\psi(i,u,j). \qquad (1)$$

Under the regime ψ, the cost of a policy Φ, starting from state i, is defined as

$$J^{\Phi,\psi}(i) = E\left[\sum_{k=0}^{N_{\Phi,\psi}-1} r_\psi(X_k, u_k, X_{k+1})\right], \qquad (2)$$

where $N_{\Phi,\psi}$ is the random time at which absorption at state S occurs, and u_k is the random sequence of actions taken at times k, under the policy Φ. Since state S is a zero-reward absorbing state, we let $J^{\Phi,\psi}(S) = 0$ for any policy and all $\psi \in \Psi$. An optimal policy under the regime ψ, is one which simultaneously maximizes (2) for every state $i \in \mathcal{S}$; the corresponding optimal cost at state i is given by $J^*_\psi(i) = \max_\Phi J^{\Phi,\psi}(i)$. For a given regime ψ, it is well-known that there always exists a deterministic policy that is optimal [4]. We label such an optimal policy μ^*_ψ, and note that its components are given by

$$\mu^*_\psi(i) = \arg\max_{u \in U(i)} \sum_{j \in S} p_\psi(i,u,j)\Big(r_\psi(i,u,j) + J^*_\psi(j)\Big), \qquad (3)$$

for $i \in \mathcal{S} \setminus S$ [4].

A critical feature of any policy is whether (or not) the absorbing state is reached with probability one from any state, under that particular policy. A policy for which this is true is called a *proper* policy [1], and a policy for which this is not true is *improper*. Here, we assume that for each $\psi \in \Psi$, the state transition probabilities of the MDP are such that every policy is proper.

In this paper, we consider the problem of tracking the optimal policy, $\mu^*_{\psi_k}$, abbreviated as μ^*_k. As noted earlier, we make no assumptions regarding the process ψ_k, other than Assumption 1, below. Let T_ψ denote the duration of regime ψ.

Assumption 1: For each regime $\psi \in \Psi$, $\mathbb{E}[T_\psi] \gg \sup_\Phi \mathbb{E}[N_{\Phi,\psi}]$.

In other words, on average, many trajectories to the absorbing state are completed within the duration of each regime.

Given a sequence $\psi_k, k = 1, 2, \ldots, K$, for some time horizon K, a natural performance measure for the online learning algorithm is the proportion of time steps during this interval for which the algorithm's *estimate* of the optimal policy is equal to the true optimal policy μ^*_k. We shall employ this measure in Section VI.

III. Review of Q-learning in a Stationary Environment

In this section, we summarise existing convergence results for Q-learning for a single regime $\psi \in \Psi$. We shall return to the regime switching scenario in Section V.

A. Background

The Q-learning method is a well-known reinforcement learning approach for solving MDPs when the expected state transition rewards $r(i,u,j)$ and state transition probabilities $p(i,u,j)$ are not known explicitly [1], [5]. This method involves the iterative "learning" of optimal Q-factors, $Q^*(i,u)$,

associated with state-action pairs $(i,u), i \in \mathcal{S}, u \in \mathcal{U}(i)$, via the iteration

$$Q_{k+1}(i,u) := Q_k(i,u) + \eta_k(i,u)\Big(r(i,u,\zeta_k) + \max_{v \in U(\zeta_k)} Q_{k+1}(\zeta_k, v) - Q_k(i,u)\Big), \quad (4)$$

where ζ_k is the random successor state of i at the k^{th} update, having probability mass function $p(i,u,\cdot)$, and $r(i,u,\zeta_k)$ is the corresponding single-stage reward. We note that (4) is a stochastic iterative method for solving the fixed point equation $HQ = Q$, where

$$(HQ)(i,u) = \sum_{j \in S} p(i,u,j)\Big(r(i,u,j) + \max_{v \in U(j)} Q(j,v)\Big). \quad (5)$$

In particular, the mapping H is a weighted maximum norm contraction provided that every policy is proper [1]. For each pair (i,u), define the sequence of random variables

$$Y_k(i,u) = r(i,u,\zeta_k) + \max_{v \in \mathcal{U}(\zeta_k)} Q_k(\zeta_k, v) - Q_k(i,u), \quad (6)$$

for $k = 0, 1, 2 \ldots$. Then (4) can be expressed as

$$Q_{k+1}(i,u) := Q_k(i,u) + \eta_k(i,u) Y_k(i,u). \quad (7)$$

Finally, the candidate policy $\tilde{\mu}_k$ with components

$$\tilde{\mu}_k(i) = \arg\max_{v \in \mathcal{U}(i)} Q_k(i,v), \quad i \in \mathcal{S} \setminus S, \quad (8)$$

is an *estimate* of the optimal policy.

B. Summary of known convergence results: decreasing stepsize parameters

The following is the very general result from [6], which covers a broad range of online and offline implementations.

Theorem 1: [6] If for every pair (i,u), the time interval separating consecutive updates of $Q_k(i,u)$ is finite with probability one, and the stepsize parameters satisfy $\sum_{i=1}^{\infty} \eta_k^2(i,u) < \infty$, $\sum_{i=1}^{\infty} \eta_k(i,u) = \infty$, then the $Q_k(i,u)$ converge to the optimal Q-factors $Q^*(i,u)$ with probability one, provided at least one of the following conditions is satisfied: (1) every policy is proper, or (2) the sequence $\{Q_k(i,u), k \in \mathbb{N}\}$ is bounded with probability one.

Thus, if the conditions of Theorem 1 are satisfied, it follows that the sequence of candidate policies $\tilde{\mu}_k, k = 1, 2, \ldots$, given in (8) converge with probability one to an optimal policy.

C. Online Q-learning

In this paper we consider an on-line implementation, where the Markov chain X_k represents the "real-time" state trajectory of the system to be controlled, and where u_k represents the action that is selected at stage k (discussed below). The state transition probabilities, $p(i,u,j)$, which drive the system are unknown, and the only information that is available to the learning agent is the state trajectory and the rewards that are generated at each stage. Specifically, once the action u_k is taken, the subsequent state X_{k+1} and reward $r(X_k, u_k, X_{k+1})$ are observed directly from the system.

An important consequence of the online model is that the sequence of successive indices of the Q-factors that are updated is a stochastic process. In particular, the update of a given Q-factor, $Q(i,u)$, is triggered at each time point k for which the event $\{X_k = i, u_k = u\}$ occurs. The Q-learning update rule for the online case is

$$Q_{k+1}(i,u) := Q_k(i,u) + \eta_k(i,u)\hat{Y}_k(i,u), \quad (9)$$

where

$$\hat{Y}_k(i,u) = \mathbf{1}_{\{X_k = i, u_k = u\}}\Big(r(X_k, u_k, X_{k+1}) + \max_{v \in \mathcal{U}(X_{k+1})} Q_k(X_{k+1}, v) - Q_k(X_k, u_k)\Big). \quad (10)$$

Because of the indicator function in (10), the update (9) reduces to $Q_{k+1}(i,u) := Q_k(i,u)$ for all state-action pairs $(i,u) \neq (X_k, u_k)$.

Naturally, the selection of actions u_k constitutes a key aspect of the algorithm design. In order to guarantee that all state-action pairs are "visited" by the algorithm, it is necessary to employ randomized (rather than deterministic) policies, despite the fact that randomized policies are typically suboptimal [4]. Here, we employ the well-known "ϵ-soft" method for action selection [5]: given the current state $X_k = i, i \in \mathcal{S} \setminus S$, the action u_k is selected according to

$$\begin{aligned} P(u_k = \tilde{\mu}_k(i)) &= 1 - \epsilon \\ P(u_k = u \in \mathcal{U}(i) \setminus \tilde{\mu}_k(i)) &= \frac{\epsilon}{|\mathcal{U}(i)| - 1}, \end{aligned} \quad (11)$$

where $\epsilon \in (0,1)$ parameterizes the exploration level, and $\tilde{\mu}_k(i)$ is the candidate action defined in (8). Each time the absorbing state is reached, we take the restart state to be state 1, and we also assume that there exists a path of positive probability from state 1 to every other state under at least one policy. Thus, the convergence result of Theorem 1 applies to the online implementation of the algorithm described here if the stepsize condition of the theorem is met. However, the stepsize condition of Theorem 1 implies that the stepsize parameters approach zero, which precludes the possibility of tracking the optimal policy in a non-stationary environment. In order to perform tracking, the stepsizes must be bounded away from zero – the remainder of this paper addresses this case.

IV. Convergence results for the constant stepsize case in a stationary environment

Consider the case where the Q-factors are constant. Given a state-action pair (X_k, u_k) at stage k, the state X_{k+1} follows the distribution $p(X_k, u_k, \cdot)$. Since Q is constant, then $\tilde{\mu}_k$ is constant, and u_{k+1} is determined from (11). It follows that $(X_k, u_k)_Q$ is a homogeneous Markov chain. That it is ergodic follows from the assumptions in Section III-C, provided that the exploration parameter ϵ is strictly positive. Therefore, $(X_k, u_k)_Q$ possesses a unique stationary

distribution, which we denote using π_Q. To keep the notation simple for the theoretical analysis in this section, we assume that the constant stepsize parameters all have the same value, that is, $\eta(i,u) = \eta$ for all pairs (i,u). This is not a strong assumption and can be easily generalized. Define the continuous-time interpolation $Q^\eta(i,u;t) = Q_k(i,u)$ for $t \in [\eta k, \eta(k+1))$.

Theorem 2: The continuous-time interpolation $Q^\eta(i,u;t)$ converges in distribution, as $\eta \to 0$, to the unique solution of the ODE

$$\frac{dQ(i,u;t)}{dt} = \pi_{Q(t)}(i,u)\left[(HQ)(i,u;t) - Q(i,u;t)\right]. \quad (12)$$

Proof: The result follows from a straightforward application of the method in Chapter 8 of [2] as well as [7]. The details are given in [3]. ∎

We note that since H is a contraction mapping [1] which is bounded and continuous [3], the system (12) has as its unique stable point the set of optimal Q-factors, $Q^*(i,u)$, provided that ϵ is strictly positive (see (11)).

The following result follows from Theorem 10.1.1 of [2].

Theorem 3: Under some regularity assumptions [2], as $\eta \to 0$, there exists a time $\tau < \infty$ beyond which the fluctuations around the limit ODE (12) are approximately normally distributed, that is, $Q^\eta(t) - Q^* \overset{\mathrm{d}}{\approx} \mathcal{N}(0, \eta\,\Lambda)$, for $t \geq \tau$, where Λ is a stationary variance-covariance matrix that depends on the MDP parameters and is independent of η.

Although the Q-factors converge weakly to the ODE, and hence eventually converge weakly to the optimal Q-values, we can nonetheless obtain the following strong convergence result regarding the candidate policies, given in Theorem 4.

Theorem 4: For the online Q-learning algorithm (9) - (11) with constant step size η and constant exploration rate ϵ, the sequence of candidate policies $\tilde{\mu}_k$ converges almost surely to the optimal policy μ^* as $\eta \to 0$.

Proof: Define for each $i \in \mathcal{S} \setminus S$, $\tilde{\mu}^\eta(i;t) = \arg\max_{v \in \mathcal{U}_i} Q^\eta(i,v;t)$. Assume that for each i, all the optimal Q-values are unique, that is, there is a unique $u^*(i) = \arg\max_v Q^*(i,v)$. Were this not the case, two actions $u_1^*(i)$ and $u_2^*(i)$ would be equivalent and the problem can be restated without loss of generality. From Theorem 3, the process $Q^\eta(\cdot)$ converges in distribution to Q^*, and thus for each i, there exists a $T < \infty$ such that $\mu(i;t) = \arg\max_v(Q(i,v;t)) = u^*(i)$, for all $t \geq T$. From the weak convergence of Q^η, it follows that for any time t, the sequence $\tilde{\mu}^\eta(i;t)$ also converges in distribution to $\mu(i;t)$. In particular, for $t > T$, this limit is a single point mass. Any integer-valued random variable that converges in distribution to a constant also converges with probability one, which proves the claim. ∎

V. Online Q-learning in a non-stationary environment

In this section, we present an adaptive stepsize scheme that is appropriate for online Q-learning in a non-stationary environment which is characterized by the regime switching behaviour described in Section II. Importantly, no *apriori* knowledge of the regime switching times or of the regime parameter values is assumed; the only assumptions about the regime switching that we make are those stated in Assumption 1.

Recall from Section II that for each time step k, μ_k^* denotes the optimal policy corresponding to the current regime ψ_k. We extend the notation of Sections III and IV to reflect this time-dependence; let $Q_k^*(i,u)$, for all pairs (i,u), denote the optimal Q-values under the regime ψ_k.

Our algorithm uses a sequence of distinct stepsize parameters $\eta_k(i), k = 1, 2, \ldots,$ for each state i, where $\eta_k(i)$ is used for updating each of $Q_k(i,u), u \in \mathcal{U}_i$, at iteration k. In order to track the optimal policy, the sequences of stepsizes $\eta_k(i)$ must be bounded away from zero. Indeed, the simplest and most common approach for tracking would be to employ a constant stepsize [2], [7], which then naturally raises the question of which fixed value should be selected. There exists a tradeoff between the following objectives: (1) minimising the magnitude of the fluctuations of the Q-factors *within a given regime* (achieved using small stepsizes), and (2) maximising the rate at which the Q-factors are able to *adapt in response to a regime change* (achieved using large stepsizes). The reason for the first objective is that larger fluctuations in the Q-factors increase the likelihood that they are not in the correct order with respect to the optimal Q-factors, which in turn can lead to the event $E_{i,k} = \{\arg\max_{u \in \mathcal{U}_i} Q_k(i,u) \neq \mu_k^*(i)\}$ for some state i and iteration k, that is, an error in the identification of an optimal policy. Indeed, the result of Theorem 3 shows that within a given regime, as $\eta \to 0$, the fluctuations of the Q-factors about their optimal values approach a normal distribution having variance proportional to η, and Theorem 4 shows that the candidate policy $\tilde{\mu}_k$ is equal to the optimal policy μ_k^* with probability 1 in this limit.

However, immediately following a regime change, the rate of convergence of the Q-factors to the new optimal values is proportional to η, and thus larger values of η should be favored to speed up the rate of convergence in response to regime changes. The key to our approach for balancing these objectives is as follows: we anticipate (arbitrary) regime changes by always selecting the largest possible stepsizes for the current regime, as this maximises the rate of adaptation when a regime change eventually does occur. However, this maximisation is subject to constraints on the probability of incorrectly identifying the optimal actions. Specifically, for each state i, and at each iteration k, we update $\eta_k(i)$ using the solution to the optimisation problem

$$\max \eta(i) \quad \text{subject to} \quad \mathbb{P}(E_{i,k}) \leq \gamma \qquad (13)$$

where $\gamma \in (0,1)$ is a pre-determined "error tolerance". We impose the additional constraint $0 < \eta(i) \leq \overline{\eta}$, where $\overline{\eta} \leq 1$.

The result of Theorem 3 suggests using a normal distribution to approximate $\mathbb{P}(E_{i,k})$. To illustrate our approach, consider the particular case of a state i which has only two available actions, so that $\mathcal{U}_i = \{1,2\}$. The associated Q-factors are the random variables $Q_k(i,1)$ and $Q_k(i,2)$.

Within a given regime, these Q-factors are subject to fluctuations about their optimal values $Q_k^*(i,1)$ and $Q_k^*(i,2)$, respectively. Suppose that the stepsize parameter for these Q-factors is set to $\eta(i)$. Following Theorem 3, we assume that the Q-factors are distributed according to the joint distribution $\mathcal{N}(U_k, V_k)$, where $U_k = (Q_k^*(i,1), Q_k^*(i,2))$, and

$$V_k = \begin{pmatrix} \eta(i)\lambda_k^2(1,1) & \eta(i)\lambda_k^2(1,2) \\ \eta(i)\lambda_k^2(1,2) & \eta(i)\lambda_k^2(2,2) \end{pmatrix}$$

is the corresponding variance-covariance matrix, for some $\lambda_k^2(1,1), \lambda_k^2(2,2)$ and $\lambda_k^2(1,2)$. Now define the sequence of random variables $D_k(i) = Q_k(i,1) - Q_k(i,2)$, and observe that within a given regime, the expectation of $D_k(i)$ is given by $d_k(i) = Q_k^*(i,1) - Q_k^*(i,2)$, and that its variance is given by $\eta(i)\lambda_k^2(i)$, where $\lambda_k^2(i) = \lambda_k^2(1,1) + \lambda_k^2(2,2) - 2\lambda_k^2(1,2)$. It follows that the constraint in (13) is equivalent to

$$\frac{|d_k(i)|}{\sqrt{\eta(i)\lambda_k^2(i)}} \geq z_{1-\gamma}, \tag{14}$$

where $z_{1-\gamma}$ is the one-sided $100(1-\gamma)$-percentile for the standard normal distribution.

Let $\eta^{target}(i)$ denote the largest value of $\eta(i)$ on the interval $(0, \overline{\eta}]$ such that (14) is satisfied. Algorithm 1 below estimates $\eta^{target}(i)$ for the current regime, by first estimating the quantities $d_k^2(i)$ and $\lambda_k^2(i)$ via a "moving window" approach, using the M most-recent iterates. This information is then used to update the stepsize $\eta_k(i)$.

Algorithm 1: Initialise $M \in \mathbb{Z}^+$, $\overline{\eta} \in (0,1]$, $\alpha \in (0,1], \gamma \in (0,1)$. For each state i, and for each iteration $k \geq M$:

1) Calculate the sample mean $\widehat{d_k^2}(i)$ and sample covariance matrix $\widehat{V}_k$, using the M most-recent iterates of $Q_k(i,1)$ and $Q_k(i,2)$.
2) Calculate $\widehat{\lambda_k^2}(i) = \frac{1}{\eta_k(i)}(\widehat{V}_k(1,1) + \widehat{V}_k(2,2) - 2\widehat{V}_k(1,2))$ and $\hat{\eta}^{target}(i) = \frac{1}{\widehat{\lambda_k^2}(i)} \frac{\widehat{d_k^2}(i)}{z_{1-\gamma}^2}$.
3) Update $\eta_{k+1}(i) = \min\{\overline{\eta}, (1-\alpha)\eta_k(i) + \alpha\hat{\eta}^{target}(i)\}$.

The smoothing in Step 3 of the algorithm improves the stability of the iterates $\eta_k(i)$ (for our numerical experiments, we set $\alpha = 0.1$). We also note that since $\hat{\eta}^{target}(i) \propto \frac{\widehat{d_k^2}(i)}{\widehat{\lambda_k^2}(i)}$, and in general, $\mathbb{E}[\frac{X}{Y}] \neq \frac{\mathbb{E}[X]}{\mathbb{E}[Y]}$ for any random variables X and Y, then $\hat{\eta}^{target}(i)$ is a biased estimator of $\eta^{target}(i)$. We are currently working on improving the statistical properties of this estimator via appropriate correction terms.

VI. Numerical experiments

In this section, we present numerical experiments using Algorithm 1 applied in a non-stationary environment, and we evaluate its performance. We take a collection of regimes (MDPs) $\Psi = \{1,2,3,4,5\}$, having different state transition probabilities and single-stage rewards. All MDPs have $S = 4$ states, with state 4 being the absorbing state, and two available actions at each state (that is, $\mathcal{U}_i = \{1,2\}$, for $i = 1,2,3$). We take the following (arbitrary) deterministic sequence of regimes $\mathcal{R} = \{1,2,3,4,5,1,4\}$, over a total of $K = 1500$ time steps, with the regime changes occurring at the times $\mathcal{T} = \{300, 450, 600, 800, 1000, 1250\}$, respectively. Note that for this sequence, the ratio $\frac{T_\psi}{\mathbb{E}[N_\psi]}$ lies in the range $[40, 60]$ for $\psi \in \Psi$ (c.f. Assumption 1). Algorithm 1 is used to to adapt the stepsizes $\eta_k(i), i = 1,2,3$, in real-time. We use the algorithm parameter values $M = 50, \alpha = 0.1$, $\overline{\eta} = 0.8$, and for the constraint (14), we set $\gamma = 0.05$. For the exploration level in the online Q-learning algorithm, we fix $\epsilon = 0.1$ in (11).

To better understand the results of the experiment, Figure 1 shows a single sample path for the difference of Q-factors at state 1, that is, the sequence of differences $Q_k(1,1) - Q_k(1,2)$ (solid lines). The dotted line shows the corresponding sequence of differences of the optimal Q-factors, $Q_k^*(1,1) - Q_k^*(1,2)$, which changes accordingly at times $\mathcal{T}$. Thus, at all time points k for which these sequences have the same sign, the Q-factors are in the "correct order", and the candidate action $\tilde{\mu}_k(1)$ is equal to the optimal action $\mu_k^*(1)$. In particular, the sequence of optimal actions at state 1 which correspond to the sequence $\mathcal{R}$ is $\{2,2,1,2,1,2,2\}$.

As a benchmark against which to compare the performance of Algorithm 1, we take the fixed stepsize version of the on-line Q-learning algorithm, that is, (9) – (11), with $\eta_k(i) = \eta$ for all k and states $i = 1,2,3$. As a measure of performance, we take for each state i the random variable $F_i = \frac{1}{K}\sum_{k=1}^K \mathbf{1}_{\{\tilde{\mu}_k(i) = \mu_k^*(i)\}}$, that is, the proportion of time-steps for which the candidate action at state i is equal to the true optimal action at i. We use F_i to compare the performance of Algorithm 1 with that of the fixed stepsize algorithm.

For the fixed stepsize algorithm, we consider a range of values of η, from 0.1 to 0.8. For each fixed value of η, we perform 100 replications of the experiment (using the same $\mathcal{R}$ and $\mathcal{T}$), and we construct 95% confidence intervals for F_1, F_2 and F_3, which are shown on the left-hand portion of Figures 2 – 4. The right hand portion of these figures shows the corresponding 95% confidence intervals obtained using Algorithm 1.

We see that for each state, Algorithm 1 compares favourably with the best performance that could be obtained using a fixed stepsize. In particular, the fixed values of η which yield the best performance for states 1, 2 and 3, are approximately 0.3, 0.7, and 0.5, respectively. We note that these values cannot be determined in advance, without *apriori* knowledge of $\mathcal{R}$ and $\mathcal{T}$ (here, we have determined them with hindsight, using many replications of the experiment). Furthermore, the above values differ significantly from one another, highlighting the advantage of using Algorithm 1 over using any fixed stepsize, when no prior information about regime changes is available; for example, if we were to select $\eta = 0.3$, which happens to be the best-performing value for state 1, then a significant drop in performance would be incurred at states 2 and 3, and vice versa if we selected $\eta = 0.5$ or $\eta = 0.7$. On the other hand, the adaptive stepsize algorithm is able to achieve a performance that is close to or better than what is possible using a fixed stepsize, and more importantly, is it able to do so *simultaneously* over

all states, without prior knowledge of the regime switching process.

We note that while the use of Algorithm 1 (as opposed to using a fixed stepsize) entails the introduction of a new set of "tunable" parameters M, α and $\overline{\eta}$, we have found its good performance to be robust to a broad range of values of these parameters. In contrast, the fixed stepsize algorithm is not as robust to changes in the fixed value η. Our experience suggests that the values $M = 50, \alpha = 0.1$ and $\overline{\eta} = 0.8$ produce consistently good results for experiments of the kind presented here.

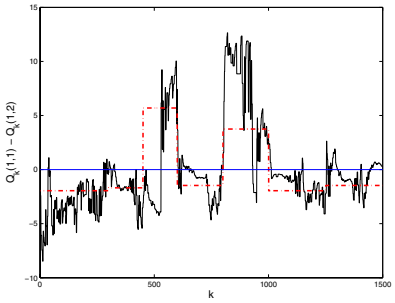

Fig. 1. Sample path for difference of Q-factors at state 1 (solid line), together with the difference of the corresponding optimal Q-factors (dashed line). The sequence of optimal actions is $\{2, 2, 1, 2, 1, 2, 2\}$.

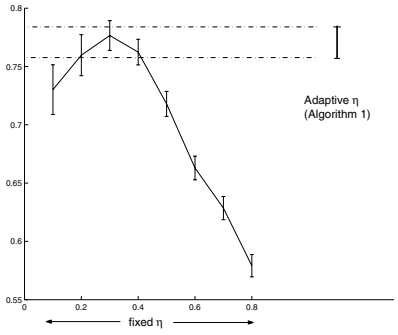

Fig. 2. Fixed stepsize versus adaptive stepsize: 95% confidence intervals for F_1.

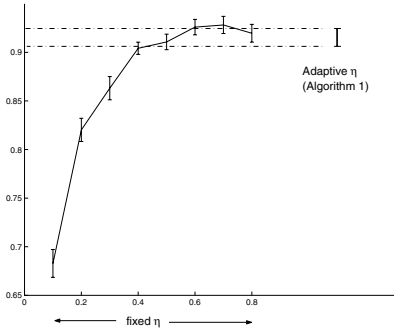

Fig. 3. Fixed stepsize versus adaptive stepsize: 95% confidence intervals for F_2.

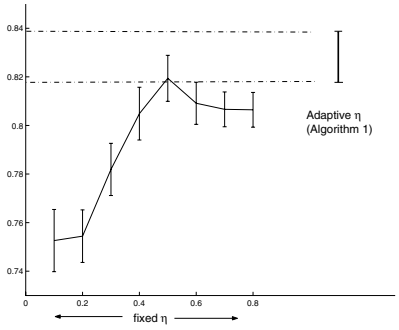

Fig. 4. Fixed stepsize versus adaptive stepsize: 95% confidence intervals for F_3.

VII. Discussion

We have presented a new adaptive stepsize algorithm that is appropriate for online learning in a non-stationary environment that is characterized by regime switching. The work presented in this paper has focused on the well-known Q-learning algorithm. However, our approach, which is essentially based on the mathematical program (13) and the weak convergence results for constant stepsize stochastic approximation algorithms, can be used for any iterative learning algorithm. We are currently implementing Algorithm 1 for the case of more than two actions per state, and we are refining the estimator $\hat{\eta}^{target}(i)$ (see Step 2 of Algorithm 1).

Finally, we note that an important aspect of online learning is the tradeoff between exploration of alternative actions and the exploitation of the current best estimates. In this paper, we implemented ϵ-soft policies (see (11)) with a fixed ϵ. For learning in a stationary environment, it is common practice to gradually let $\epsilon \to 0$ [1], [5]. However, this is not possible in a non-stationary environment, as it would eventually prevent adaptation (just as occurs if the stepsizes sequences $\eta_k(i)$ approach zero). Thus, it is important that ϵ be bounded away from zero. The investigation of adaptive exploration schemes and their integration with the present material constitutes a useful topic for further research.

References

[1] D.P. Bertsekas and J.N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

[2] H.J. Kushner and G.Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, 2003.

[3] K. Levy, *Apprentissage par simulation stochastique: étude de convergence et application à un probleme de tarification aérienne modélisé par un processus de décision markovien*, M Sc Thesis, Department of Computer Science and OR, University of Montreal, 2005.

[4] M. Puterman, *Markov Decision Processes*, John Wiley, 1994.

[5] R.S. Sutton and A.G. Barto, *Reinforcement Learning*, MIT Press, 1998.

[6] J.N. Tsitsiklis, *Asynchronous Stochastic Approximation and Q-Learning*, Machine Learning, Vol. 16, 185–202, 1994.

[7] F.J. Vázquez-Abad (1999) "Strong Points of Weak Convergence: A Study Using RPA Gradient Estimation for Automatic Learning", *Automatica*, Vol. 35, No. 7: 1255–1274.

[8] G.Yin and Q. Zhang, *Discrete-Time Markov Chains: Two-TIme-Scale Methods and Applications*, Springer, 2005.

[9] G.Yin, V. Krishnamurthy and C. Ion *Regime Switching Stochastic Approximation Algorithms with Application to Adaptive Discrete Stochastic Optimization*, SIAM Journal on Optimization, Vol.14, No.4: 1187–1215, 2004.