LEARNING IN CHANGING ENVIRONMENTS

Reinforcement Learning in Environments with Asymptotically Bounded Variation

Balázs Csanád Csáji

Research Associate (a), Former Ph.D. Student (b)

(a) Computer and Automation Research Institute, Hungarian Academy of Sciences

(b) Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary

Gatsby Unit, University College London, 24 September, 2008

Preliminaries

- First, we investigate the effects of environmental changes on the value function. We show that the optimal value function Lipschitz continuously depends on the transition function and on the immediate cost function.
- Then, we analyze stochastic iterative algorithms with time-dependent update operators. A relaxed convergence theorem is presented and some numerical experiments are shown.
- Afterwards, we introduce (ε, δ)-MDPs, a class of non-stationary MDPs. In this model the transition and the cost functions may change over time, provided that the accumulated changes remain bounded in the limit.
- Finally, we consider reinforcement learning methods in (ε, δ)-MDPs. An approximate convergence theorem is deduced from the previous results.

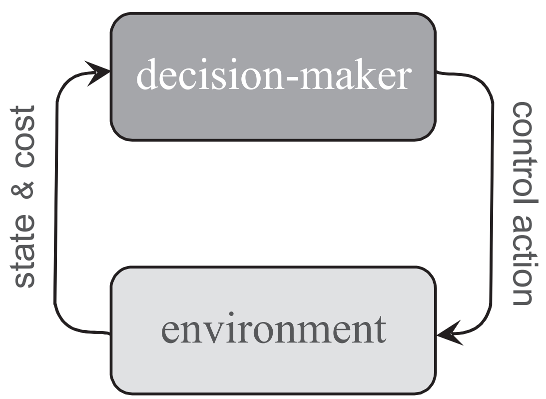
Contents

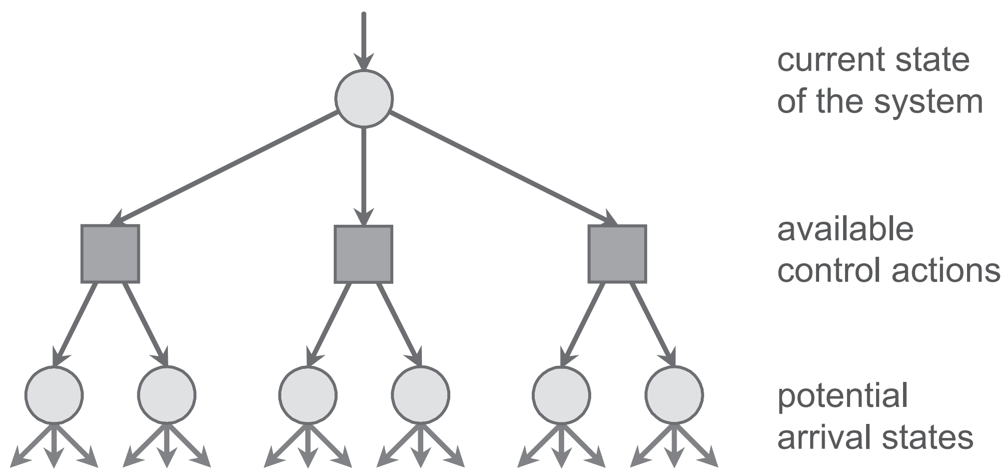
PART I.

Introduction

Reinforcement Learning

- Reinforcement learning (RL) is a computational approach to learn from the interaction with an environment based on feedbacks, e.g., rewards.
- An interpretation: consider an agent acting in an uncertain environment and receiving information on the actual states and immediate costs.
- The aim is to learn an efficient behavior (control policy), such that applying this strategy minimizes the expected costs in the long run.





Markov Decision Processes (MDPs)

By a (stationary, finite, discrete time, fully observable) Markov decision process (MDP) we mean a 6-tuple $\mathcal{M} = \langle \mathbb{X}, \mathbb{A}, \mathcal{A}, p, g, \alpha \rangle$, where:

- $\mathbb{X}$ is a finite set of discrete states
- $\mathbb{A}$ is a finite set of control actions
- $\mathcal{A} : \mathbb{X} \to \mathcal{P}(\mathbb{A})$ is an action constraint function
- $p : \mathbb{X} \times \mathbb{A} \to \Delta(\mathbb{X})$ is the transition function, $p(y \,|\, x, a)$ denotes the probability of arriving at state y after taking action $a \in \mathcal{A}(x)$ in a state x
- $g : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ is an immediate cost (or reward) function
- $\alpha \in [0, 1)$ is the discount rate or discount factor.

It is “Markov”, since p and g only depend on the current state and action.

The Bellman Equation

- A (stationary, Markovian, randomized) control policy $\pi : \mathbb{X} \to \Delta(\mathbb{A})$ is a function from states to probability distributions over actions.

- The value function of a policy $J^\pi : \mathbb{X} \to \mathbb{R}$ is defined as follows

$$J^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t g(X_t, A_t^\pi) \,\middle|\, X_0 = x\right],$$

  where $A_t^\pi \sim \pi(X_t)$ and $X_{t+1} \sim p(X_t, A_t)$ ("$\sim$" = "has distribution").

- The Bellman optimality equation is $TJ^* = J^*$ where

$$(TJ)(x) = \min_{a \in \mathcal{A}(x)} \left[g(x,a) + \alpha \sum_{y \in \mathbb{X}} p(y \,|\, x, a) J(y)\right]$$

- We aim at finding a policy that minimizes the expected costs.

Contractions and Value Iteration

- The action-value function of a policy $Q^{\pi} : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ is

$$Q^{\pi}(x,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t g(X_t, A_t^{\pi}) \,\middle|\, X_0 = x, A_0 = a\right]$$

- Function $f : \mathcal{X} \to \mathcal{Y}$ is Lipschitz continuous if there exists a $\beta \geq 0$: $\forall\, x_1, x_2 \in \mathcal{X} : \|f(x_1) - f(x_2)\|_{\mathcal{Y}} \leq \beta \,\|x_1 - x_2\|_{\mathcal{X}}$, where $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, respectively.
- The smallest such β is called the Lipschitz constant of f.
- If $\beta < 1$, then the function is called a contraction mapping.
- The Bellman operator is a contraction with Lipschitz constant α.
- Therefore, J^* is unique and it is the limit of the iteration $J_{t+1} = TJ_t$.

PART II.

Value Function Bounds for Environmental Changes

Value Function Bounds for Changes

Theorem 1. *Assume that two discounted MDPs differ only in their transition-probability functions, and let these two functions be denoted by* p_1 *and* p_2*. Let the corresponding optimal value functions be* J_1^* *and* J_2^**, then*

$$\| J_1^* - J_2^* \|_\infty \leq \frac{\alpha \, |\mathbb{X}| \, \|g\|_\infty}{(1-\alpha)^2} \, \| p_1 - p_2 \|_\infty$$

Theorem 2. *Assume that two discounted MDPs differ only in the immediate-cost functions, and let these two functions be denoted by* g_1 *and* g_2*. Let the corresponding optimal value functions be* J_1^* *and* J_2^**, then*

$$\| J_1^* - J_2^* \|_\infty \leq \frac{1}{1-\alpha} \, \| g_1 - g_2 \|_\infty$$

We applied the supremum norm: $\| f \|_\infty = \sup \{ | f(x) | : x \in dom(f) \}$.

Value Function Bounds for Changes

Theorem 3. *Assume that two discounted MDPs differ only in their transition-probability functions, and let these two functions be denoted by p_1 and p_2. Let the corresponding optimal value functions be J_1^* and J_2^*, then*

$$\|J_1^* - J_2^*\|_\infty \leq \frac{\alpha\,\|g\|_\infty}{(1-\alpha)^2}\,\|p_1 - p_2\|_1\,,$$

where $\|\cdot\|_1$ is a norm on $f : \mathbb{X} \times \mathbb{A} \times \mathbb{X} \to \mathbb{R}$ type functions:

$$\|f\|_1 = \max_{x,\,a} \sum_{y \in \mathbb{X}} |\, f(x,a,y)\,|$$

For example, $f(x,a,y) = p(y \,|\, x,a)$.

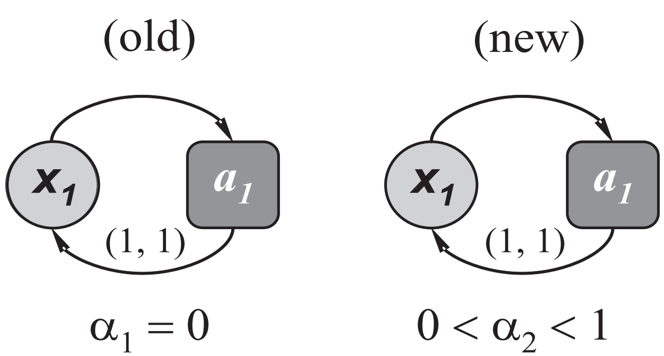
Since $\forall f : \|f\|_1 \leq n\,\|f\|_\infty$, where n is size of the state space, the estimation of Theorem 3 is at least as good as the estimation of Theorem 1.

Discount Factor Changes

Theorem 4. *Assume that two MDPs differ only in the* *discount factors,* $\alpha_1, \alpha_2 \in [0, 1)$. *Let the optimal value functions be* J_1^* *and* J_2^**, then*

$$\|J_1^* - J_2^*\|_\infty \leq \frac{|\alpha_1 - \alpha_2|}{(1-\alpha_1)(1-\alpha_2)} \|g\|_\infty$$

However, as the following example shows, this dependence is non-Lipschitz.



$$J_1^*(x_1) = 1$$

$$|\alpha_1 - \alpha_2| < 1$$

$$J_2^*(x_1) \to \infty$$

$$as\ \alpha_2 \to 1$$

At the same time, if we fix an $\alpha_0 < 1$ and only allow discount factors from $[0, \alpha_0]$, then this dependence becomes Lipschitz continuous, as well.

Tracing Back to Cost Changes

Discount factor changes can be traced back to cost changes:

Lemma 5. *Assume that two discounted MDPs, $\mathcal{M}_1$ and $\mathcal{M}_2$, differ only in the discount factors, denoted by α_1 and α_2. Then, there exists an MDP, denoted by $\mathcal{M}_3$, such that it differs only in the immediate-cost function from $\mathcal{M}_1$, thus its discount factor is α_1, and it has the same optimal value function as $\mathcal{M}_2$. The immediate-cost function of $\mathcal{M}_3$ is*

$$\widehat{g}(x,a) = g(x,a) + (\alpha_2 - \alpha_1) \sum_{y \in \mathbb{X}} p(y \,|\, x, a) J_2^*(y),$$

where p is the transition function of all $\mathcal{M}_i$; g is the cost function of $\mathcal{M}_1$ and $\mathcal{M}_2$; and $J_2^*(y)$ denotes the optimal cost-to-go function of $\mathcal{M}_2$.

However, cost changes cannot be traced back to transition changes!

PART III.

Stochastic Iterative Algorithms with Time-Dependent Update

Stochastic Iterative Algorithms (SIAs)

- We denote the set of value functions by $\mathcal{V}$ which contains, in general, all bounded real-valued functions over an arbitrary set $\mathcal{X}$.
- Many learning and optimization methods can be written in a general form as a stochastic iterative algorithm (SIA). More precisely, for all $x \in \mathcal{X}$ as

$$\boxed{V_{t+1}(x) = (1 - \gamma_t(x))V_t(x) + \gamma_t(x)((K_t V_t)(x) + W_t(x)),}$$

where $V_t \in \mathcal{V}$, operator $K_t : \mathcal{V} \to \mathcal{V}$ acts on value functions, parameter γ_t denotes random stepsizes and W_t is a noise parameter.

- Note that the value function operator, K_t, is time-dependent.
- Q-learning, SARSA and TD-learning, e.g., can be formulated this way.

Two Classical Examples

- The Robbins-Monro stochastic approximation algorithm: let q_t be a sequence of independent identically-distributed (i.i.d.) random variables with unknown mean μ and finite variance. Let us define v_t by

$$v_{t+1} = (1 - \gamma_t)v_t + \gamma_t q_t$$

  Then, sequence v_t converges almost surely to μ if suitable assumptions on the stepsize parameters, γ_t, are made, e.g., $\gamma_t = 1/t$.

- Another example of a SIA is the stochastic gradient descent algorithm which aims at minimizing cost function f and is described by

$$v_{t+1} = (1 - \gamma_t)v_t + \gamma_t(v_t - \nabla f(v_t) + w_t),$$

  where w_t is a noise parameter and ∇f denotes the gradient of f.

Main Assumptions

(A1) There exits a constant $C > 0$ such that for all x and t, we have

$$\mathbb{E}\left[W_t(x) \,|\, \mathcal{F}_t\right] = 0 \qquad \text{and} \qquad \mathbb{E}\left[W_t^2(x) \,|\, \mathcal{F}_t\right] < C < \infty,$$

where $\mathcal{F}_t = \{V_0, \dots, V_t, W_0, \dots, W_{t-1}, \gamma_0, \dots, \gamma_t\}$ is the "history".

(A2) For all x and t: $\gamma_t(x) \in [0, 1]$ and we have with probability one

$$\sum_{t=0}^{\infty} \gamma_t(x) = \infty \qquad \text{and} \qquad \sum_{t=0}^{\infty} \gamma_t^2(x) < \infty$$

(A3) For all t, operator $K_t : \mathcal{V} \to \mathcal{V}$ is a supremum norm contraction mapping with Lipschitz constant $\beta_t < 1$ and with fixed point V_t^*:

$$\forall\, V_1, V_2 \in \mathcal{V} : \left\| K_t V_1 - K_t V_2 \right\|_\infty \leq \beta_t \left\| V_1 - V_2 \right\|_\infty$$

Let us introduce $\beta_0 = \limsup_{t \to \infty} \beta_t$, and we assume that $\beta_0 < 1$.

Approximate Convergence

Definition 6. *We say that a sequence of random variables X_t κ-approximates X with $\kappa > 0$ if for all $\varepsilon > 0$ there exits a t_0 such that*

$$\mathbb{P}\left(\sup_{t>t_0} (\|X_t - X\| \leq \kappa) \right) > 1 - \varepsilon$$

Theorem 7. *Suppose that Assumptions (A1), (A2) and (A3) hold and let V_t be the sequence generated by a SIA. Then, for any $V^*, V_0 \in \mathcal{V}$, the sequence V_t κ-approximates function V^* with*

$$\kappa = \frac{4\varrho}{1-\beta_0} \qquad \textit{where} \qquad \varrho = \limsup_{t\to\infty} \|V_t^* - V^*\|_\infty$$

Notice that V^* can be an *arbitrary* function!

A Simple Numerical Example

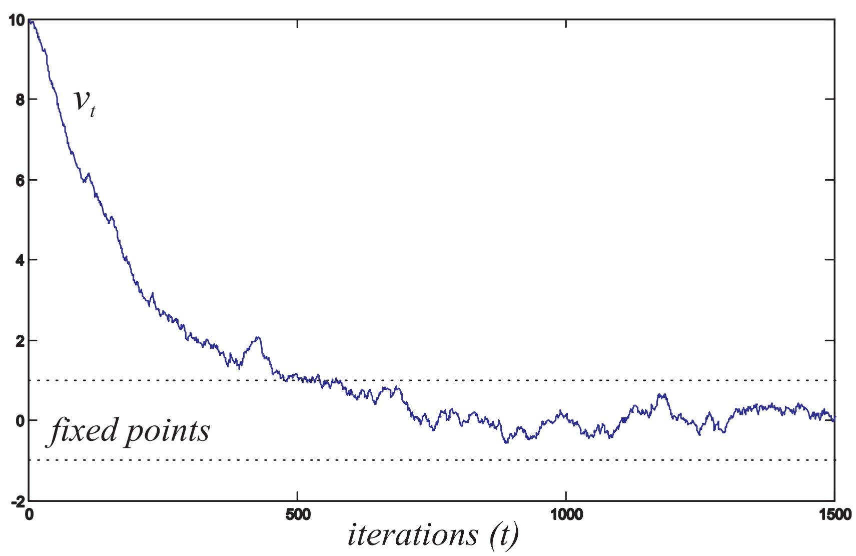
Consider a one dimensional stochastic process, v_t, characterized by

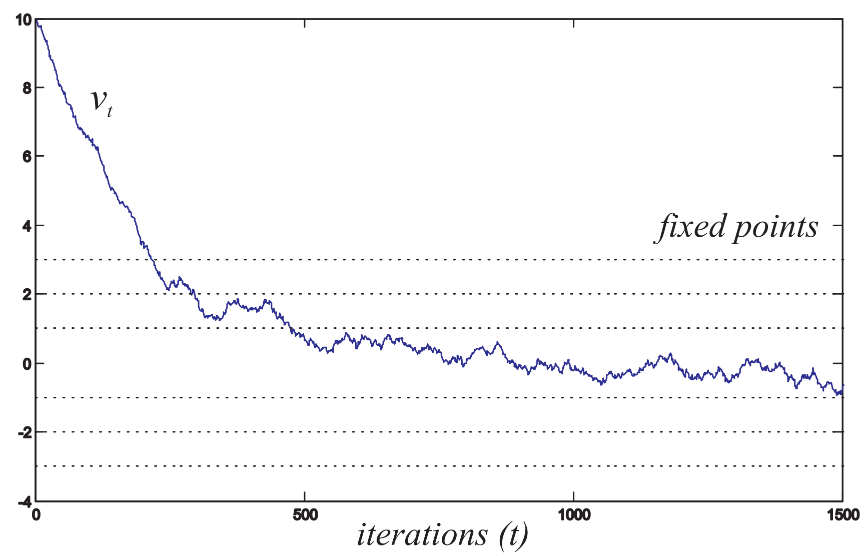
$$v_{t+1} = (1 - \gamma_t) v_t + \gamma_t (K_t(v_t) + w_t),$$

where γ_t is the learning rate and w_t is a noise term. Suppose we have n alternating operators k_i with Lipschitz constants $b_i < 1$ and fixed points v_i^*

$$k_i(v) = v + (1 - b_i)(v_i^* - v)$$

The current operator at time t is $K_t = k_i$ if $i \equiv t \pmod{n}$.





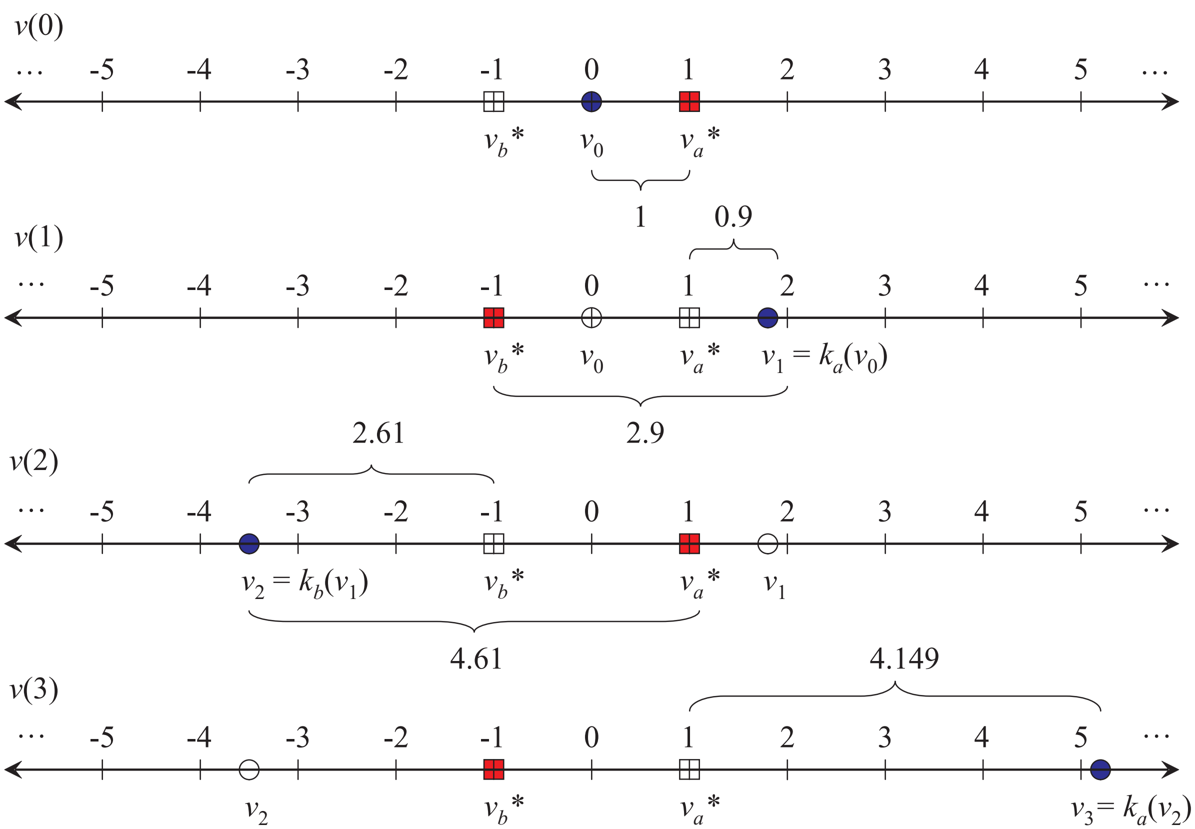
A Deterministic Pathological Example

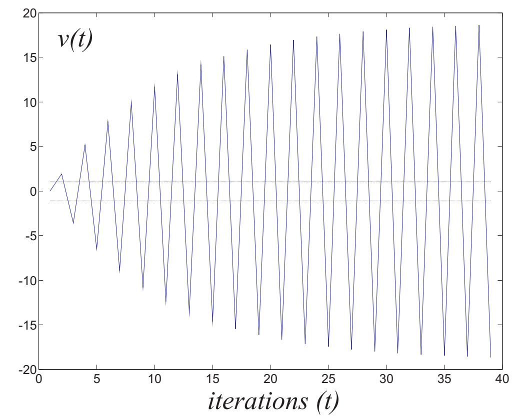
- According to the Banach fixed point theorem, if we have a contraction f over a complete metric space, $\mathcal{V}$, with fixed point $v^* = f(v^*)$ then for any initial v_0 the sequence $v_{t+1} = f(v_t)$ converges to v^*.
- Now, suppose we have n alternating contraction mappings k_i with Lipschitz constants $b_i < 1$ and fixed points v_i^*, respectively.
- Let $v_{t+1} = K_t(v_t)$ where $K_t = k_i$ if $i \equiv t \pmod{n}$, v_0 is arbitrary.
- In each iteration, K_t attracts the point towards its fixed point.
- Then, does v_t converge to the convex hull of the fixed points?
- No! Moreover, even if v_0 is in the middle of the convex hull v_t could starting moving farer and farer from the fixed points in each iteration.

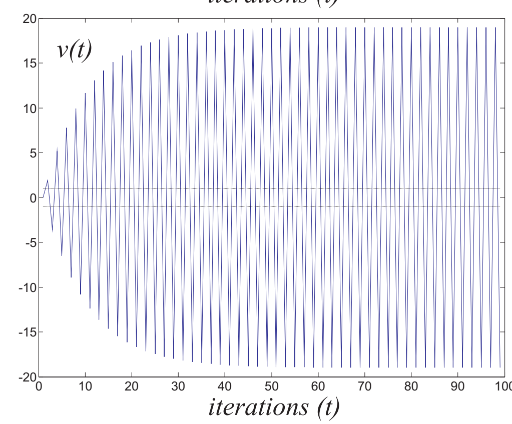
A Deterministic Pathological Example

$$k_i(v) = \begin{cases} v + (1-b_i)(v_i^* - v) & \text{if } sgn(v_i^*) = sgn(v), \\ v_i^* + (v_i^* - v) + (1-b_i)(v - v_i^*) & \text{otherwise,} \end{cases}$$

where $sgn(\cdot)$ denotes the signum function and $i \in \{a, b\}$.







PART IV.

Reinforcement Learning in Non-Stationary MDPs

Varying Environments: (ε, δ)-MDPs

- Now, a class of non-stationary MDPs is defined. In this model the transition-probabilities and the immediate-costs may change over time, as long as the accumulated changes remain asymptotically bounded.

Definition 8. *A tuple* $\langle \mathbb{X}, \mathbb{A}, \mathcal{A}, \{p_t\}_{t=1}^{\infty}, \{g_t\}_{t=1}^{\infty}, \alpha \rangle$*, which represents a sequence of MDPs, is called an* (ε, δ)*-MDP where* $\varepsilon, \delta > 0$*, if there exists an MDP,* $\mathcal{M} = \langle \mathbb{X}, \mathbb{A}, \mathcal{A}, p, g, \alpha \rangle$*, called the base MDP, such that*

- $\limsup_{t \to \infty} \|p - p_t\| \leq \varepsilon$
- $\limsup_{t \to \infty} \|g - g_t\| \leq \delta$

- The optimal cost-to-go function of the base MDP, $\mathcal{M}$, and of the current MDP at time t, $\mathcal{M}_t$, are denoted by J^* and J_t^*, respectively.

Relaxed Convergence in (ε, δ)-MDPs

Assume we have an (ε, δ)-MDP, then (from Theorems 2 and 3)

$$\limsup_{t \to \infty} \| J^* - J_t^* \|_\infty \leq d(\varepsilon, \delta)$$

$$d(\varepsilon, \delta) = \frac{\alpha\, \varepsilon\, (\|g\|_\infty + \delta)}{(1-\alpha)^2} + \frac{\delta}{1-\alpha}$$

where J_t^* and J^* are the optimal value functions of $\mathcal{M}_t$ and $\mathcal{M}$.

Corollary 9. *Consider an (ε, δ)-MDP and assume that (A1), (A2) and (A3) hold. Let V_t be the sequence generated by a SIA. Assume that the fixed point of each K_t is J_t^*. Then, V_t κ-approximates J^* with*

$$\kappa = \frac{4\, d(\varepsilon, \delta)}{1 - \beta_0}$$

Async Value Iteration in (ε, δ)-MDPs

- "Classical" value iteration $\forall J_0 : J_{t+1} = T J_t$ converges to J^*.
- A small stepsize variant of asynchronous value iteration in (ε, δ)-MDPs:

$$J_{t+1}(x) = (1 - \gamma_t(x))J_t(x) + \gamma_t(x)(T_t J_t)(x),$$

  where T_t is the Bellman operator of the current MDP at time t.

- Corollary 9 can be applied to prove convergence in (ε, δ)-MDPs:
  - There is no noise term $\Rightarrow$ (A1) is trivially satisfied.
  - Each operator T_t is an α contraction $\Rightarrow$ (A3) holds.
  - Thus, (A2) $\Rightarrow$ J_t κ-approximates J^* with $\kappa = 4\, d(\varepsilon, \delta)/(1 - \alpha)$.

Q-learning in (ε, δ)-MDPs

- The one-step version of Watkins' Q-learning rule in (ε, δ)-MDPs is

$$Q_{t+1}(x,a) = (1-\gamma_t(x,a))Q_t(x,a) + \gamma_t(x,a)(\widetilde{T}_t Q_t)(x,a),$$

$$(\widetilde{T}_t Q_t)(x,a) = g_t(x,a) + \alpha \min_{B \in \mathcal{A}(Y)} Q_t(Y,B),$$

  where Y is a random variable generated from (x,a) by simulation.

- The $\widetilde{T}_t$ operator can be rewritten in a form as follows

$$(\widetilde{T}_t Q)(x,a) = (\widetilde{K}_t Q)(x,a) + \widetilde{W}_t(x,a),$$

  where $\widetilde{W}_t(x,a)$ is a noise with zero mean and finite variance, and

$$(\widetilde{K}_t Q)(x,a) = g_t(x,a) + \alpha \sum_{y \in \mathbb{X}} p_t(y \mid x,a) \min_{b \in \mathcal{A}(y)} Q(y,b).$$

Q-learning in (ε, δ)-MDPs

- W_t has zero mean and finite variance $\Rightarrow$ (A1) is satisfied.
- Each operator $\widetilde{K}_t$ is an α contraction $\Rightarrow$ (A3) holds.
- Thus, (A2) $\Rightarrow$ Q_t κ-approximates Q^* with $\kappa = 4\,d(\varepsilon, \delta)/(1-\alpha)$.
- Similarly, the approximate convergence of TD(λ) (temporal difference learning) policy evaluation in (ε, δ)-MDPs can be proven.

Lemma 10. *Assume we have two discounted MDPs which differ only in the transition-probability functions or only in the immediate-cost functions or only in the discount factors. Let the corresponding optimal action-value functions be Q_1^* and Q_2^*, respectively. Then the bounds for $\|J_1^* - J_2^*\|_\infty$ of Theorems 3, 2 and 4 are also bounds for $\|Q_1^* - Q_2^*\|_\infty$.*

Changes During Scheduling

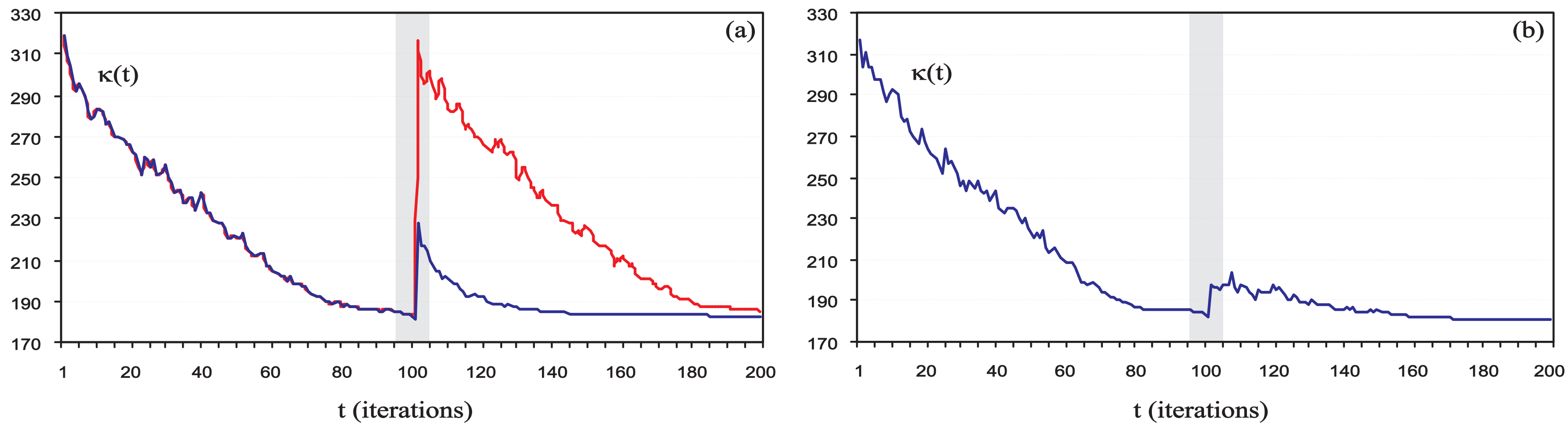


Figure 1: disturbance: (a) machine breakdown, (b) new machine availability

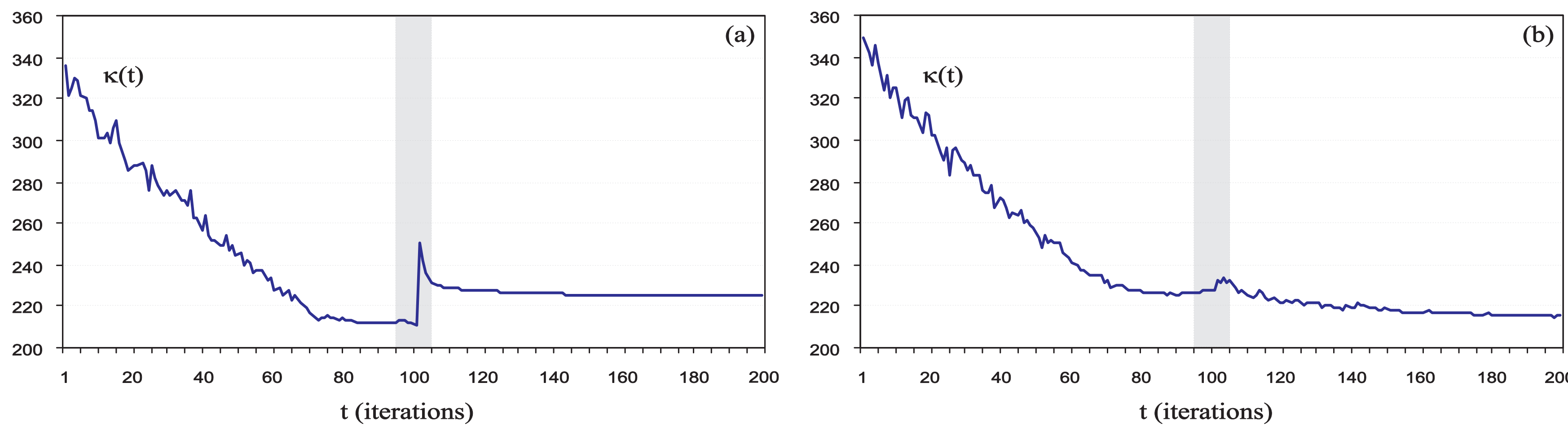


Figure 2: disturbance: (a) new job arrival, (b) job cancellation

Approximate Dynamic Programming

- Approximate dynamic programming (ADP) methods often take the form

$$\Phi(r_{t+1}) = \Pi\big((1-\gamma_t)\,\Phi(r_t) + \gamma_t(K_t(\Phi(r_t)) + W_t)\big),$$

  where $r_t \in \Theta$ is a parameter, Θ is the parameter space, e.g., $\Theta \subseteq \mathbb{R}^p$, $\Phi : \Theta \to \mathcal{F}$ is an approximation function where $\mathcal{F} \subseteq \mathcal{V}$ is a Hilbert space. Function $\Pi : \mathcal{V} \to \mathcal{F}$ is a projection mapping and K_t, W_t, γ_t are the same as the previously (cf. stochastic iterative algorithms).

- In order to apply the previous results, Π should be
  - Additive: $\forall\, V_1, V_2 : \Pi(V_1 + V_2) = \Pi(V_1) + \Pi(V_2)$
  - Homogeneous: $\forall\, V : \forall\, \alpha : \Pi(\alpha\, V) = \alpha\, \Pi(V)$
  - Nonexpansive: $\forall\, V_1, V_2 : \|\Pi(V_1) - \Pi(V_2)\| \leq \|V_1 - V_2\|$

- Then, Theorem 7 provides convergence results for ADPs.

PART V.

Conclusion

Conclusion

1. The value functions of discounted MDPs Lipschitz continuously depend on the transition-probability and the immediate-cost functions.

2. In (ε, δ)-MDPs these function may vary over time, provided that the accumulated changes remain asymptotically bounded.

3. A convergence theorem for stochastic iterative algorithms with time-dependent update was given. Under suitable assumptions, this theorem guarantees convergence to an environment of a target function.

4. These results were combined to deduce a convergence theorem for reinforcement learning algorithms working in changing environments.

5. Some numerical experiments were also presented to demonstrate working in changing environments.

Related Literature

- Csáji, B. Cs.; Monostori, L.: Value Function Based Reinforcement Learning in Changing Markovian Environments, Journal of Machine Learning Research (JMLR), 2008, Vol. 9, 2008, 1679–1709
- Csáji, B. Cs.; Monostori, L.: Adaptive Stochastic Resource Control: A Machine Learning Approach, Journal of Artificial Intelligence Research (JAIR), AAAI Press, Vol. 32, 2008, 453–486
- Csáji, B. Cs.: Adaptive Resource Control; Machine Learning Approaches to Resource Allocation in Uncertain and Changing Environments, Ph.D. Thesis, Supervisor: Monostori, L., Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary, 2008

Thank you for your attention!

If you have further questions, you can contact me at:

csaji@sztaki.hu, http://www.sztaki.hu/∼csaji