Online Markov Decision Processes under Bandit Feedback

Gergely Neu*,†
gergely.neu@gmail.com
*Department of Computer Science and Information Theory
Budapest University of Technology and Economics, Hungary

András György
gya@szit.bme.hu
†Machine Learning Research Group
MTA SZTAKI Institute for Computer Science and Control, Hungary

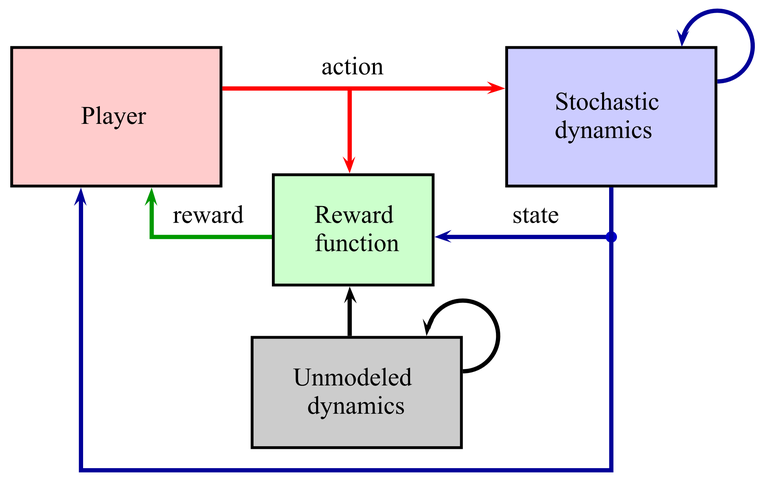
Csaba Szepesvári
szepesva@ualberta.ca
Department of Computing Science
University of Alberta, Canada

András Antos
antos@szit.bme.hu
Machine Learning Research Group
MTA SZTAKI Institute for Computer Science and Control, Hungary

Abstract

We consider online learning in finite, stochastic Markovian environments where in each time step a new reward function is chosen by an oblivious adversary. The goal of the learning agent is to compete with the best stationary policy in terms of the total reward received. In each time step the agent observes the current state and the reward associated with the last transition, however, the agent does not observe the rewards associated with other state-action pairs. The agent is assumed to know the transition probabilities. The state of the art result for this setting is a no-regret algorithm. In this paper we propose a new learning algorithm and, assuming that stationary policies mix uniformly fast, we show that the expected regret of the new algorithm in T time steps is $\mathcal{O}\left(T^{2/3}(\ln T)^{1/3}\right)$, giving the first rigorously proved regret bound for the problem.

Online Markov Decision Processes



- Reward sequence: an arbitrary sequence $r_1, r_2, \ldots, r_T$ fixed in advance, where $r_t : \mathcal{X} \times \mathcal{A} \to [0,1]$.
- History up to time t:
  $\mathbf{u}_t = (\mathbf{x}_1, \mathbf{a}_1, r_1(\mathbf{x}_1, \mathbf{a}_1), \mathbf{x}_2, \mathbf{a}_2, r_2(\mathbf{x}_2, \mathbf{a}_2), \ldots, \mathbf{x}_t, \mathbf{a}_t, r_t(\mathbf{x}_t, \mathbf{a}_t))$
- At time step t:
  - agent selects action $\mathbf{a}_t$ based on $\mathbf{u}_{t-1}$ and $\mathbf{x}_t$;
  - observes reward $r_t(\mathbf{x}_t, \mathbf{a}_t)$;
  - observes new state $\mathbf{x}_{t+1} \sim P(\cdot|\mathbf{x}_t, \mathbf{a}_t)$.

The learning problem

- Policy: $\boldsymbol{\pi}_t(a|x) = \mathbb{P}\left[\mathbf{a}_t = a | \mathbf{x}_t = x, \mathbf{u}_{t-1}\right]$
- Expected total reward of the player:

$$\hat{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{x}_t, \mathbf{a}_t)\right],$$

  where $\mathbf{a}_t \sim \boldsymbol{\pi}_t(\cdot|\mathbf{x}_t)$ and $\mathbf{x}_{t+1} \sim P(\cdot|\mathbf{x}_t, \mathbf{a}_t)$
- Expected total reward of a fixed policy π:

$$R_T^\pi = \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{x}'_t, \mathbf{a}'_t)\right],$$

  where $\mathbf{a}'_t \sim \pi(\cdot|\mathbf{x}'_t)$ and $\mathbf{x}'_{t+1} \sim P(\cdot|\mathbf{x}'_t, \mathbf{a}'_t)$
- Goal: minimize regret

$$\hat{L}_T = \sup_\pi R_T^\pi - \hat{R}_T.$$

Previous work

- Full information: $\hat{L}_T = O(\sqrt{T})$: Even-Dar et al. [2009]
- Bandit information: $\hat{L}_T = o(T)$: Yu et al. [2009]
- Bandit information for episodic loop-free MDPs: $\hat{L}_T = O(\sqrt{T})$ Neu et al. [2010]

Assumptions

- **Assumption A1** *Every policy π has a well-defined unique stationary distribution μ^π.*
- **Assumption A2** *The stationary distributions are uniformly bounded away from zero:* $\inf_{\pi,x} \mu^\pi(x) \geq \beta > 0$.
- **Assumption A3** *There exists some fixed positive mixing time τ such that for any two arbitrary μ and μ' over $\mathcal{X}$,*

$$\sup_\pi \|(\mu - \mu')P^\pi\|_1 \leq e^{-1/\tau}\|\mu - \mu'\|_1.$$

Definitions

- Value functions and average rewards:

$$\rho_t^\pi = \lim_{S\to\infty} \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}\left[r_t(\mathbf{x}'_s, \mathbf{a}'_s)\right],$$

$$q_t^\pi(x,a) = \mathbb{E}\left[\sum_{s=1}^{\infty}\left(r_t(\mathbf{x}'_s, \mathbf{a}'_s) - \rho_t^\pi\right) \middle| \mathbf{x}'_1 = x, \mathbf{a}'_1 = a\right],$$

$$v_t^\pi(x,a) = \mathbb{E}\left[\sum_{s=1}^{\infty}\left(r_t(\mathbf{x}'_s, \mathbf{a}'_s) - \rho_t^\pi\right) \middle| \mathbf{x}'_1 = x\right].$$

- At time t, use only experience gathered up to time step $t - N$ and define

$$\boldsymbol{\mu}_{t,x,a}^N(x') = \mathbb{P}\left[\mathbf{x}_t = x' \,|\, \mathbf{x}_{t-N} = x, \mathbf{a}_{t-N} = a, \boldsymbol{\pi}_{t-N+1}, \ldots, \boldsymbol{\pi}_{t-1}\right],$$

  so that $\boldsymbol{\mu}_{t,x,a}^N$ is positive.
- Estimate reward as

$$\hat{\mathbf{r}}_t(x,a) = \begin{cases} \frac{r_t(x,a)}{\boldsymbol{\pi}_t(a|x)\boldsymbol{\mu}_t^N(x|\mathbf{x}_{t-N},\mathbf{a}_{t-N})}, & \text{if } (x,a) = (\mathbf{x}_t, \mathbf{a}_t); \\ 0, & \text{otherwise.} \end{cases}$$

  Mixing ensures that the probability of visiting state x at time t is positive for all x and t, that is,

$$\boldsymbol{\pi}_t(a|x)\boldsymbol{\mu}_t^N(x|\mathbf{x}_{t-N}, \mathbf{a}_{t-N}) > 0.$$

- Let $\hat{\boldsymbol{\rho}}_t = \sum_{x,a} \mu^{\boldsymbol{\pi}_t}(x)\boldsymbol{\pi}_t(a|x)\hat{\mathbf{r}}_t(x,a)$ and solve, for all x, a, the Bellman equations

$$\hat{\mathbf{q}}_t(x,a) = \hat{\mathbf{r}}_t(x,a) - \hat{\boldsymbol{\rho}}_t + \sum_{x',a'} P(x'|x,a)\boldsymbol{\pi}_t(a'|x')\hat{\mathbf{q}}_t(x,a).$$

Algorithm

Set $N \geq 1$, $\mathbf{w}_1(x,a) = \mathbf{w}_2(x,a) = \cdots = \mathbf{w}_{2N}(x,a) = 1$, $\gamma \in (0,1)$, $\eta \in (0,\gamma]$.
For $t = 1, 2, \ldots, T$, repeat

1. Set
$$\boldsymbol{\pi}_t(a|x) = (1-\gamma)\frac{\mathbf{w}_t(x,a)}{\sum_b \mathbf{w}_t(x,b)} + \frac{\gamma}{|\mathcal{A}|}$$
for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.
2. Draw an action $\mathbf{a}_t$ randomly, according to the policy $\boldsymbol{\pi}_t(\cdot|\mathbf{x}_t)$.
3. Receive reward $r_t(\mathbf{x}_t, \mathbf{a}_t)$ and observe $\mathbf{x}_{t+1}$.
4. If $t \geq N+1$
   (a) Compute $\boldsymbol{\mu}_t^N(x|\mathbf{x}_{t-N}, \mathbf{a}_{t-N})$ for all $x \in \mathcal{X}$.
   (b) Construct estimates $\hat{\mathbf{r}}_t$ and compute $\hat{\mathbf{q}}_t$ using the Bellman equations for $\boldsymbol{\pi}_t$.
   (c) Set $\mathbf{w}_{t+N}(x,a) = \mathbf{w}_{t+N-1}(x,a)e^{\eta\hat{\mathbf{q}}_t(x,a)}$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.

Main result

Theorem 1. *Let* $N = \lceil \tau \ln T \rceil$,

$$C = (2\tau+4)\tau|\mathcal{A}|\ln T + (3\tau+1)^2,$$

$$\eta = T^{-2/3} \cdot (\ln|\mathcal{A}|)^{2/3} \cdot \left(\frac{4C(\tau+2)}{\beta}\right)^{-1/3},$$

$$\gamma = T^{-1/3} \cdot (2\tau+4)^{-2/3} \cdot \left(\frac{2C\ln|\mathcal{A}|}{\beta}\right)^{1/3}.$$

Then

$$\hat{L}_T \leq 3\,T^{2/3} \cdot \left(\frac{(4\tau+8)\ln|\mathcal{A}|}{\beta}C\right)^{1/3} + \mathcal{O}\left(T^{1/3}\right).$$

Proof

The proof is based on ideas from Even-Dar et al. [2009]. The complication is of course that rewards are estimated. The regret can be decomposed as

$$R_T^\pi - \hat{R}_T = \underbrace{\left(R_T^\pi - \sum_{t=1}^{T}\rho_t^\pi\right)}_{(i)} + \underbrace{\left(\sum_{t=1}^{T}\rho_t^\pi - \sum_{t=1}^{T}\rho_t^{\boldsymbol{\pi}_t}\right)}_{(ii)} + \underbrace{\left(\sum_{t=1}^{T}\rho_t^{\boldsymbol{\pi}_t} - \hat{R}_T\right)}_{(iii)}.$$

Bounding (i)

After Even-Dar et al. [2009]:

$$R_T^\pi - \sum_{t=1}^{T}\rho_t^\pi \leq 2\tau + 2.$$

The policies $\boldsymbol{\pi}_t$ change slowly

Lemma 1. *Let* $c = \frac{2\eta}{\beta}\left(\frac{1}{\gamma} + 4\tau + 6\right)$. *Assume that* $c(3\tau+1)^2 < \beta/2$ *and* $N \geq \left\lceil \tau \ln\left(\frac{4}{\beta - 2c(3\tau+1)^2}\right)\right\rceil$. *Then, for all* $N < t \leq T$, $x, x' \in \mathcal{X}$ *and* $a \in \mathcal{A}$, *we have*

$$\boldsymbol{\mu}_{t,x,a}^N(x') \geq \beta/2$$

and

$$\max_{x'}\sum_{a'}|\boldsymbol{\pi}_{t+1}(a'|x') - \boldsymbol{\pi}_t(a'|x')| \leq c.$$

Bounding (ii)

After Even-Dar et al. [2009]:

$$\rho_t^\pi - \rho_t^{\boldsymbol{\pi}_t} = \sum_{x,a}\mu^\pi(x)\pi(a|x)\left[q_t^{\boldsymbol{\pi}_t}(x,a) - v_t^{\boldsymbol{\pi}_t}(x)\right].$$

A simple modification of the proof of the regret bound of **Exp3** yields the following:

Proposition 1. *Let* $N \geq \lceil \tau \ln T \rceil$. *For any policy* π *and for all* T *large enough, we have*

$$\sum_{t=1}^{T}\mathbb{E}\left[\rho_t^\pi - \boldsymbol{\rho}_t\right] \leq (4\tau+10)N + \frac{\ln|\mathcal{A}|}{\eta} + (2\tau+4)\,T\left(\gamma + \frac{2\eta}{\beta}|\mathcal{A}|\,(Nc + (e-2)(2\tau+4))\right).$$

Bounding (iii)

Proposition 2. *Let* $N \geq \lceil \tau \ln T \rceil$. *For any* T *large enough,*

$$\sum_{t=1}^{T}\mathbb{E}\left[\boldsymbol{\rho}_t\right] - \hat{R}_T \leq T\,c(3\tau+1)^2 + 2Te^{-N/\tau} + 2N.$$

Follows from the slow change of policies $\boldsymbol{\pi}_t$.

References

E. Even-Dar, S.. M. Kakade, and Y. Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

G. Neu, A. György, and Cs. Szepesvári. The online loop-free stochastic shortest-path problem. In *COLT-10*, pages 231–243, 2010.

J. Y. Yu, S. Mannor, and N. Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.