The adversarial stochastic shortest path problem with unknown transition probabilities

Gergely Neu
Budapest Univ. of Technology and Economics
and MTA SZTAKI, Budapest, Hungary

András György
University of Alberta
Edmonton, Canada

Csaba Szepesvári
University of Alberta
Edmonton, Canada

Abstract

We consider online learning in a special class of episodic Markovian decision processes, namely, loop-free stochastic shortest path problems. In this problem, an agent has to traverse through a finite directed acyclic graph with random transitions while maximizing the obtained rewards along the way. We assume that the reward function can change arbitrarily between consecutive episodes, and is entirely revealed to the agent at the end of each episode. Previous work was concerned with the case when the stochastic dynamics is known ahead of time, whereas the main novelty of this paper is that this assumption is lifted. We propose an algorithm called "follow the perturbed optimistic policy" that combines ideas from the "follow the perturbed leader" method for online learning of arbitrary sequences and "upper confidence reinforcement learning", an algorithm for regret minimization in Markovian decision processes (with a fixed reward function). We prove that the expected cumulative regret of our algorithm is of order $L|\mathcal{X}||\mathcal{A}|\sqrt{T}$ up to logarithmic factors, where L is the length of the longest path in the graph, $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space and T is the number of episodes. To our knowledge this is the first algorithm that learns and controls stochastic and adversarial components in an online fashion at the same time.

1 Introduction

In this paper we study reinforcement learning problems where the performance of learning algorithms is measured by the total reward they collect during learning. We are interested in algorithms with theoretically guaranteed bounds on their performance. Variants of this problem have received considerable attention during the last decade (see Brafman and Tennenholtz (2002); Kakade (2003); Bartlett and Tewari (2009a); Jaksch et al. (2010), or Section 4.2.4 of Szepesvári (2010) for a summary of available results and the references therein). Most works consider the case when the learner controls a finite Markovian Decision Process (MDP). While there exists a few works that extend the theoretical analysis beyond finite MDPs, these come at strong assumptions on the MDP (e.g., Kakade et al., 2003; Strehl and Littman, 2008; Abbasi-Yadkori and Szepesvári, 2011). Still, all these approaches assume that the state of the environment is completely Markovian and thus one can construct a predictive model of the environment. However, in practice, the environment might be very complex, for example, in an inventory management problem, the "world's economy" may influence the prices at which one can buy or sell items, thus modeling the environment as an MDP would mean that the learner must model the world's economy, based on her limited information.

This paper belongs to the important stream of works, initiated by Even-Dar et al. (2005) and Yu et al. (2009), where *the assumption that the state is completely Markovian is relaxed.* The main idea relies on the observation that in a number of practical problems, such as the above-mentioned inventory management problem, *the hard-to-model, complex part of the environment influences only the rewards that the learner receives.*

The interaction between the learner and the environment is shown in Figure 1. The environment is split into two parts: One part that has Markovian dynamics and another one with an unrestricted, autonomous dynamics. In each discrete time step, the agent receives the state of the Markovian environment and the previous state of the autonomous dynamics. The learner then makes a decision about the next action, which is sent to the environment. The environment

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.

then makes a transition: the next state of the Markovian environment depends stochastically on the current state and the chosen action, and the other part has an autonomous dynamic which is not influenced by the learner's actions or the state of the Markovian environment. After this transition, the agent receives a reward depending on the *complete* state of the environment and the chosen action, and then the process continues. The goal of the learner is to collect as much reward as possible. The modeling philosophy is that whatever information about the environment's dynamics can be modeled should be modeled in the Markovian part and the remaining "unmodeled dynamics" is what constitutes the autonomous part of the environment.

A large number of practical operations research and control problems have the above outlined structure. These problems include production and resource allocation problems, where the major source of difficulty is to model prices, various problems in computer science, such as the k-server problem, paging problems, or web-optimization problems, such as ad-allocation problems with delayed information (see, e.g., Even-Dar et al., 2009; Yu et al., 2009). The contextual bandit setting considered by Lazaric and Munos (2011) can also be regarded as a simplified version of this model, with the restriction that the states are generated in an i.i.d. fashion.

Using the notations of Figure 1, let x_t be the Markovian state, y_t the autonomous state and $r(x_t, a_t, y_t)$ be the reward given for selecting action a_t in state (x_t, y_t). In what follows, for simplicity, by slightly abusing terminology, we call the state x_t of the Markovian part "the state" and regard dependency on y_t as dependency on t by letting $r_t(\cdot,\cdot) = r(\cdot,\cdot,y_t)$. With these notations, the learner's goal is to perform nearly as well as the best state-feedback policy in hindsight in terms of the total reward collected ($\sum_{t=1}^{T} r_t(x_t, a_t)$), i.e., to compete with the best controller of the form $\pi : \mathcal{X} \to \mathcal{A}$, where $\mathcal{X}$ is the state-space of the Markovian part of the environment and $\mathcal{A}$ is the set of actions.

Naturally, no assumptions can be made about the autonomous part of the environment as it is assumed that modeling this part of the environment lies outside of the capabilities of the learner. This leads to a *robust control guarantee*: The guarantee on the performance must hold no matter how the autonomous state sequence (y_t), or equivalently, the reward sequence (r_t) is chosen. We think that this built-in robustness might be a much desired feature in many applications.[1]

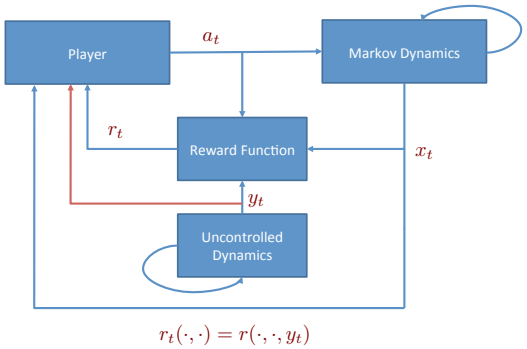

Figure 1: The interaction between the learner and the environment.

The above described problem has been studied in a number of papers under varying assumptions. Existing results are summarized in Table 1. In the table and throughout the paper we use the notation $\tilde{\mathcal{O}}(t) = \mathcal{O}(t\,\mathrm{poly}(\log t))$. In all these works the state space $\mathcal{X}$ is assumed to be finite, as is the action space $\mathcal{A}$ and in fact it is assumed that tables proportional to the size of the state-action space $\mathcal{X}\times\mathcal{A}$ fit the memory. Although this might look restrictive, we must remind the reader that *(i)* studying the finite problem provides useful insight that can be exploited later and in fact, in many cases the solutions discovered by studying finite problems can be generalized; *(ii)* if an approach does not work for finite problems then it cannot be expected to work well on larger problems either.

1.1 Our contributions

Our main contribution is an algorithm with a regret bound (of optimal order in the number of episodes) against an arbitrary sequence of reward functions but for the case *when the Markovian dynamics is unknown*. As it turns out, this problem is considerably more challenging than the one, exclusively studied in the above cited previous works, when the Markovian dynamics is known *a priori*. That said, there exists an alternate thread of previous theoretical works that assume an unknown Markovian dynamics over finite state and action spaces (Jaksch et al., 2010; Bartlett and Tewari, 2009b). In these works, however, the reward function is fixed (known or unknown). That in the problem we study *both* the Markovian dynamics is unknown and the reward function is allowed to be an arbitrary sequence forces us to explore an algorithm design principle that remained relatively less-explored so far.[2]

[1] Sometimes, robustness is associated to conservative choices and thus poor "average" performance. Although we do not study this question here, we note in passing that the algorithms we build upon have "adaptive variants" that are known to adapt to the environment in the sense that their performance improves when the environment is "less adversarial".

[2] The reason why straightforward modifications and combinations of previous techniques do not work will be discussed later in this section.

paper	reward	feedback	loops	transition	regret bound
Even-Dar et al. (2005, 2009)	adversarial	r_t	MDP	fixed, known	$\tilde{\mathcal{O}}(\sqrt{T})$
Yu et al. (2009)	adversarial	r_t	MDP	fixed, known	$\tilde{\mathcal{O}}(T^{3/4+\epsilon}), \epsilon > 0$
Jaksch et al. (2010)	stochastic	$r_t(x_t, a_t)$	MDP	fixed, unknown	$\tilde{\mathcal{O}}(\sqrt{T})$
Neu et al. (2010a)	adversarial	$r_t(x_t, a_t)$	SSP	fixed, known	$\mathcal{O}(\sqrt{T})$
Neu et al. (2010b)	adversarial	$r_t(x_t, a_t)$	MDP	fixed, known	$\tilde{\mathcal{O}}(T^{2/3})$
this paper	adversarial	r_t	SSP	fixed, unknown	$\tilde{\mathcal{O}}(\sqrt{T})$

Table 1: Existing results related to our work. For each paper we describe the setup by specifying the type of the reward function and feedback, whether the results correspond to a general MDP with loops (we do not list other restrictions presented in the papers such as mixing) or the loop-free SSP, and the type of the transition function and if it is known. Finally, for each paper and setup we present the order of the obtained regret bound in terms of the time horizon T.

Our results are proven under the assumption that the learner repeatedly solves a loop-free stochastic shortest path problem when a new reward function is selected at the beginning of each episode. We have studied this problem very briefly in Section 4 of Neu et al. (2010a) under the assumption that before learning starts, the agent is given the transition probabilities underlying the controlled Markovian dynamics. We gave an algorithm and proved that its regret is $\mathcal{O}(L^2\sqrt{T \log |\mathcal{A}|})$, where L is the number of layers in the state space, T is the number of episodes and $\mathcal{A}$ denotes the (finite) set of actions. Here, the regret is the expected difference between the performance of the learning agent and the best policy in hindsight. In this paper we give a new algorithm for this problem when the learner does not initially know the dynamics and prove that the algorithm's regret is bounded by $\tilde{\mathcal{O}}(L|\mathcal{X}||\mathcal{A}|\sqrt{T})$. Note that using the techniques of Jaksch et al. (2010), one can show a regret bound of $\tilde{\mathcal{O}}(L|\mathcal{X}|\sqrt{T|\mathcal{A}|})$ for the easier problem when the reward function is fixed and known (this bound follows from a detailed look at our proof). Thus, the price we pay for playing against an arbitrary reward sequence is an $\mathcal{O}(\sqrt{|\mathcal{A}|})$ factor in the upper bound.

Our new algorithm combines previous work on online learning in Markovian decision processes and work on online prediction in adversarial settings in a novel way. In particular, we combine ideas from the UCRL-2 algorithm of Jaksch et al. (2010) developed for learning in finite MDPs, with the "follow the perturbed leader" (FPL) prediction method for arbitrary sequences (Hannan, 1957; Kalai and Vempala, 2005), while for the analysis we also borrow some ideas from our paper Neu et al. (2010a) (which in turn builds on the fundamental ideas of Even-Dar et al. 2005, 2009). However, in contrast to our previous work where we rely on using one *exponentially weighted average forecaster* (EWA, Cesa-Bianchi and Lugosi, 2006, Section 2.8) *locally* in each state, we use FPL to compute the policy to be followed *globally*. The main reason for this change is twofold: First, we need a Follow-the-Leader-Be-the-Leader-type inequality for the policies we use and we could not prove such an inequality when the policies are implicitly obtained by placing experts into the individual states. Second, we could not find a computationally efficient way of implementing EWA over the space of policies, hence we turned to the FPL approach, which gives rise to a computationally efficient implementation. The FPL technique has been explored by Yu et al. (2009) and Even-Dar et al. (2005, 2009), but, as it is shown in Table 1, they assumed that the environment is known. It is also interesting to note that even if the reward function is fixed, no previous work considered the analysis of the FPL technique for unknown MDP dynamics. Our analysis also sheds some light on the role of the confidence intervals and the principle of "optimism in face of uncertainty" in on-line learning.

2 Problem formulation

Markovian decision processes (MDPs) are a widely studied framework for sequential decision making (Puterman, 1994). In an MDP a decision maker (or agent) observes its current state x, and based on its previous experiences, selects an action a. As a result, the decision maker receives some positive reward $r(x, a)$ and its state is changed randomly according to some probability distribution conditioned on x and a. The goal of the decision maker is to maximize its cumulative reward.

A special class of MDPs is the class of loop-free stochastic shortest path problems (loop-free SSP problems): Informally, the decision maker has to start from a fixed initial state x_0, and has to make its way to a terminal state x_L while maximizing the total rewards gathered during the passage. Once the terminal state is reached, the agent is placed back to the initial state x_0, and thus a new *episode* is started. We assume that during each episode, the agent can visit each state only once, or, in other words, the environment is loop-free.

In this paper we consider the online version of the loop-free SSP problem (as defined in Neu et al. 2010a), in which case the reward function is allowed to change between episodes, that is, instead of a single reward function r, we are given a sequence of reward functions (r_t) describing the rewards at episode t that is assumed to be an *individual sequence*, that is, no statistical assumption is made about the rewards.

More formally, the online loop-free SSP problem M is defined by a tuple $(\mathcal{X}, \mathcal{A}, P, (r_t))$, where

$\mathcal{X}$ is the state space, including two special states, the initial and the terminal states x_0 and x_L, respectively.

$\mathcal{A}$ is the action space, the set of available actions at each state.

$P : \mathcal{X} \times \mathcal{X} \times \mathcal{A} \to [0,1]$ is the transition function, where $P(x'|x,a)$ gives the probability of moving to state x' upon selecting action a in state x,

$r_t : \mathcal{X} \times \mathcal{A} \to [0,1]$ is the reward function in episode t, where $r_t(x,a)$ is the reward given for choosing action a in state x. No stochastic assumptions are made on the sequence (r_t).

Loop free condition: The state space is composed of *layers* $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_L$, where $\mathcal{X}_0 = \{x_0\}$, $\mathcal{X}_L = \{x_L\}$, $\mathcal{X}_l \cap \mathcal{X}_k = \emptyset$ if $l \neq k$, and transitions are only possible between consecutive layers, that is, if $P(x'|x,a) > 0$ then $x' \in \mathcal{X}_{l+1}$ and $x \in \mathcal{X}_l$ for some $0 \leq l \leq L-1$.

Note that our problem formulation implies that each path from the initial state to the terminal state consists of exactly L transitions. While this assumptions may seem restrictive, it is not: all loop-free state spaces (with varying path lengths) can easily be transformed to one that satisfies our assumption on uniform path length (a simple transformation algorithm is given in Appendix A of György et al., 2007).

Most papers on adversarial MDPs assume that the transition function is known (Even-Dar et al., 2009; Neu et al., 2010a,b) and the decision maker has to predict the rewards only. The only exception we know is the work of Yu and Mannor (2009a,b) who considered the problem of online learning in MDPs where the transition probabilities may also change arbitrarily after each transition, from a given set. This problem is significantly more difficult than the case when only the rewards have to be estimated; accordingly, the algorithms proposed in these papers fail to achieve consistency for this setting. We consider the simpler situation when the transitions in the MDP are governed by a single transition function which is *not known* before the beginning of the decision process, and has to be estimated from observed transitions in the MDP. Also, we assume that the reward function is fully observed after each finished episode for all state-action pairs. Thus, this is a significant extension of the full information problem investigated in Neu et al. (2010a) to the case of an unknown transition function.

3 Tools for SSPs

A stationary policy π (or, in short: a policy) is a mapping $\pi : \mathcal{X} \to \mathcal{A}$. We say that a policy π is followed in an SSP problem if the action in state $x \in \mathcal{X}$ is set to be $\pi(x)$, independently of previous states and actions. A random path

$$\mathbf{u} = (\mathbf{x}_0, \mathbf{a}_0, \ldots, \mathbf{x}_{L-1}, \mathbf{a}_{L-1}, \mathbf{x}_L)$$

is said to be generated by policy π under the transition model P if $\mathbf{x}_0 = x_0$ and $\mathbf{x}_{l+1} \in \mathcal{X}_{l+1}$ is drawn from $P(\cdot|\mathbf{x}_l, \pi(\mathbf{x}_l))$ for all $l = 0, 1, \ldots, L-1$. We denote this relation by $\mathbf{u} \sim (\pi, P)$. Using the above notation we can define the *value* of a policy π, given a fixed reward function r and a transition model P as

$$W(r, \pi, P) = \mathbb{E}\left[\sum_{l=0}^{L-1} r(\mathbf{x}_l, \pi(\mathbf{x}_l)) \middle| \mathbf{u} \sim (\pi, P) \right],$$

that is, the expected sum of rewards gained when following π in the MDP defined by r and P. In our problem, we are given a sequence of reward functions $(r_t)_{t=1}^T$. We will refer to the partial sum of these rewards as $R_t = \sum_{s=1}^t r_t$. Using the notations

$$v_t(\pi) = W(r_t, \pi, P) \text{ and } V_t(\pi) = W(R_t, \pi, P),$$

we can phrase our goal as coming up with an algorithm that generates a sequence of policies $(\boldsymbol{\pi}_t)_{t=1}^T$ that minimizes the *expected regret*

$$\hat{L}_T = \max_\pi V_T(\pi) - \mathbb{E}\left[\sum_{t=1}^T v_t(\boldsymbol{\pi}_t) \right].$$

At this point, we mention that regret minimization demands that we continuously incorporate each observation as it is made available during the learning process. Assume that we decide to use the following simple algorithm: run an arbitrary exploration policy π_{exp} for $0 < K < T$ episodes, estimate a transition function $\bar{\mathbf{P}}$, then, in the remaining episodes, run the algorithm described in Neu et al. (2010a) using $\bar{\mathbf{P}}$. It is easy to see that this method attains a regret of $\mathcal{O}(K + T/\sqrt{K})$, which becomes $\mathcal{O}(T^{2/3})$ when setting $K = \mathcal{O}(T^{2/3})$. Also, the regret of this algorithm would scale very poorly with the parameters of the SSP.

4 The algorithm: follow the perturbed optimistic policy

In this section we present our FPL-based method that we call "follow the perturbed optimistic policy" (FPOP). The algorithm is shown as Algorithm 1.

One of the key ideas of our approach, borrowed from Jaksch et al. (2010), is to maintain a confidence set of transition functions that contains the true transition function with high probability. The confidence set is constructed in such a way that it remains constant over random time periods, called epochs, and the number of epochs ($\mathbf{K}_T$) is guaranteed to be small relative to the time horizon (details on how to select the confidence set are presented in Section 4.1). We denote the i-th epoch by E_i, while the corresponding confidence set is denoted by $\mathcal{P}_i$. Now consider the simpler problem where the reward function r is assumed to be constant and known throughout all episodes $t = 1, 2, \ldots, T$. One can directly apply UCRL-2 of Jaksch et al. (2010) and select policy and transition-function estimate as

$$\left(\boldsymbol{\pi}_i, \tilde{\mathbf{P}}_i\right) = \operatorname*{arg\,max}_{\pi \in \Pi, \bar{P} \in \mathcal{P}_i} \left\{W(r, \pi, \bar{P})\right\}$$

in epoch E_i, and follow $\boldsymbol{\pi}_i$ through that epoch. In other words, we *optimistically* select the model from the confidence set that maximizes the optimal value of the MDP (defined as the value of the optimal policy in the MDP) and follow the optimal policy for this model and the fixed reward function. However, it is well known that for the case of arbitrarily changing reward functions, optimistic "follow the leader"-style algorithms like the one described above are bound to fail even in the simple (stateless) expert setting. Thus, we need to adopt ideas from online learning of arbitrary sequences, a topic extensively covered by Cesa-Bianchi and Lugosi (2006).

In particular, our approach is to make this optimistic method more conservative by adding some carefully designed noise to the cumulative sum of reward functions that we have observed so far – much in the vein of FPL. To this end, introduce the perturbation function $\mathbf{Y}_i : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ with $\mathbf{Y}_i(x, a)$ being i.i.d. random variables for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ and all epochs $i = 1, 2, \ldots, \mathbf{K}_T$; in our algorithm $\mathbf{Y}_i$ will be selected according to $\mathrm{Exp}(\eta, |\mathcal{X}||\mathcal{A}|)$, the $|\mathcal{X}||\mathcal{A}|$-dimensional distribution whose components are independent having exponential distribution with expectation η. Using these perturbations, the key idea of our algorithm is to choose our policy and transition-function estimate as

$$\left(\boldsymbol{\pi}_t, \tilde{\mathbf{P}}_t\right) = \operatorname*{arg\,max}_{\pi \in \Pi, \bar{P} \in \mathcal{P}_{\mathbf{i}(t)}} \left\{W(R_{t-1} + \mathbf{Y}_{\mathbf{i}(t)}, \pi, \bar{P})\right\} \quad (1)$$

where $\mathbf{i}(t)$ is the epoch containing episode t. That is, after fixing a biased reward function based on our previous observations, we still act optimistically when taking stochastic uncertainty into consideration. We call this method "follow the perturbed optimistic policy", or, in short, FPOP.

Our main result concerning the performance of this algorithm is presented below. The performance bound is detailed in Theorem 2 in the appendix.

Theorem 1. *Assume that $T > (|\mathcal{X}||\mathcal{A}|)^2$. Then, for some appropriate setting of η and δ, the expected regret of FPOP can be bounded as*

$$V_T^* - \mathbb{E}\left[\sum_{t=1}^{T} v_t(\boldsymbol{\pi}_t)\right] = \tilde{\mathcal{O}}\left(L|\mathcal{X}||\mathcal{A}|\sqrt{T}\right).$$

4.1 The confidence set for the transition function

In this subsection, following Jaksch et al. (2010), we describe how the confidence set is maintained to ensure that it contains the real transition function with high probability yet does not change too often. To maintain simplicity, we will assume that the layer decomposition of the SSP is known in advance, however the algorithm can be easily extended to cope with unknown layer structure.

The algorithm proceeds in epochs of random length: the first epoch E_1 starts at episode $t = 1$, and each epoch E_i ends when *any* state-action pair (x, a) has been chosen at least as many times in the epoch as before the epoch (e.g., the first epoch E_1 consists of the single episode $t = 1$). Let $\mathbf{t}_i$ denote the index of the first episode in epoch E_i, $\mathbf{i}(t)$ be the index of the epoch that includes t, and let $\mathbf{N}_i(x, a)$ and $\mathbf{M}_i(x'|x, a)$ denote the number of times state-action pair (x, a) has been visited and the number of times this event has been followed by a transition to x' up to episode $\mathbf{t}_i$, respectively. That is

$$\begin{aligned} \mathbf{N}_i(x_l, a_l) &= \sum_{s=1}^{\mathbf{t}_i - 1} \mathbb{I}_{\left\{\mathbf{x}_l^{(s)} = x_l, \mathbf{a}_l^{(s)} = a_l\right\}} \\ \mathbf{M}_i(x_{l+1}|x_l, a_l) &= \sum_{s=1}^{\mathbf{t}_i - 1} \mathbb{I}_{\left\{\mathbf{x}_{l+1}^{(s)} = x_{l+1}, \mathbf{x}_l^{(s)} = x_l, \mathbf{a}_l^{(s)} = a_l\right\}}, \end{aligned} \quad (2)$$

where $l = 0, 1, \ldots, L-1, x_l \in \mathcal{X}_l, a_l \in \mathcal{A}$ and $x_{l+1} \in \mathcal{X}$ (clearly, $\mathbf{M}_i(x_{l+1}|x_l, a_l)$ can be non-zero only if $x_{l+1} \in \mathcal{X}_{l+1}$). Our estimate of the transition function in epoch E_i will be based on the relative frequencies

$$\bar{\mathbf{P}}_i(x_{l+1}|x_l, a_l) = \frac{\mathbf{M}_i(x_{l+1}|x_l, a_l)}{\max\left\{1, \mathbf{N}_i(x_l, a_l)\right\}}.$$

Note that $\bar{\mathbf{P}}_i(\cdot|x, a)$ belongs to the set of probability distributions $\Delta(\mathcal{X}_{l_x+1}, \mathcal{X})$ defined over $\mathcal{X}$ with support contained in $\mathcal{X}_{l_x+1}$. Define a confidence set of transition functions for epoch E_i as the following L_1-ball

Algorithm 1 The FPOP algorithm for the online loop-free SSP problem with unknown transition probabilities.

Input: State space $\mathcal{X}$, action space $\mathcal{A}$, perturbation parameter $\eta \geq 0$, confidence parameter $0 < \delta < 1$, time horizon $T > 0$.

Initialization: Let $R_0(x,a) = 0$, $\mathbf{n}_1(x,a) = 0$, $\mathbf{N}_1(x,a) = 0$, and $\mathbf{M}_1(x,a) = 0$ for all state-action pairs $(x,a) \in \mathcal{X} \times \mathcal{A}$, let $\mathcal{P}_1$ be the set of all transition functions, and let the episode index $\mathbf{i}(1) = 1$. Choose $\mathbf{Y}_1 \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|} \sim \mathrm{Exp}(\eta, |\mathcal{X}||\mathcal{A}|)$ randomly.

For $t = 1, 2, \ldots, T$

1. Compute $\boldsymbol{\pi}_t$ and $\tilde{\mathbf{P}}_t$ according to (1).
2. Traverse a path $\mathbf{u}_t = \left(\mathbf{x}_0^{(t)}, \mathbf{a}_0^{(t)}, \ldots, \mathbf{x}_{L-1}^{(t)}, \mathbf{a}_{L-1}^{(t)}, \mathbf{x}_L^{(t)}\right)$ following the policy $\boldsymbol{\pi}_t$, where $\mathbf{x}_l^{(t)} \in \mathcal{X}_l$ and $\mathbf{a}_l^{(t)} = \boldsymbol{\pi}_t\left(\mathbf{x}_l^{(t)}\right) \in \mathcal{A}$.
3. Observe reward function r_t and receive rewards $\sum_{l=0}^{L-1} r_t\left(\mathbf{x}_l^{(t)}, \mathbf{a}_l^{(t)}\right)$.
4. Set $\mathbf{n}_{\mathbf{i}(t)}\left(\mathbf{x}_l^{(t)}, \mathbf{a}_l^{(t)}\right) = \mathbf{n}_{\mathbf{i}(t)}\left(\mathbf{x}_l^{(t)}, \mathbf{a}_l^{(t)}\right) + 1$ for all $l = 0, \ldots, L-1$.
5. If $\mathbf{n}_{\mathbf{i}(t)}(x,a) \geq \mathbf{N}_{\mathbf{i}(t)}(x,a)$ for some $(x,a) \in \mathcal{X} \times \mathcal{A}$ then start a new epoch:
   (a) Set $\mathbf{i}(t+1) = \mathbf{i}(t) + 1$, $\mathbf{t}_{\mathbf{i}(t+1)} = t + 1$ and compute $\mathbf{N}_{\mathbf{i}(t+1)}(x,a)$ and $\mathbf{M}_{\mathbf{i}(t+1)}(x,a)$ for all (x,a) by (2).
   (b) Construct $\mathcal{P}_{\mathbf{i}(t+1)}$ according to (3) and (4).
   (c) Reset $\mathbf{n}_{\mathbf{i}(t+1)}(x,a) = 0$ for all (x,a).
   (d) Choose $\mathbf{Y}_{\mathbf{i}(t+1)} \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|} \sim \mathrm{Exp}(\eta, |\mathcal{X}||\mathcal{A}|)$ independently.

   Else set $\mathbf{i}(t+1) = \mathbf{i}(t)$.

around $\bar{\mathbf{P}}_i$:

$$\hat{\mathcal{P}}_i = \Bigg\{ \bar{P} : \left\| \bar{P}(\cdot|x,a) - \bar{\mathbf{P}}_i(\cdot|x,a) \right\|_1 \leq \sqrt{\frac{2|\mathcal{X}_{l_x+1}| \ln \frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}{\max\{1, \mathbf{N}_i(x,a)\}}} \\ \text{and } \bar{P}(\cdot|x,a) \in \Delta(\mathcal{X}_{l_x+1}, \mathcal{X}) \text{ for all } (x,a) \in \mathcal{X} \times \mathcal{A} \Bigg\}. \tag{3}$$

The following lemma ensures that the true transition function lies in these confidence sets with high probability.

Lemma 1 (Jaksch et al., 2010, Lemma 17)**.** *For any* $0 < \delta < 1$

$$\left\| \bar{\mathbf{P}}_i(\cdot|x,a) - P(\cdot|x,a) \right\|_1 \leq \sqrt{\frac{2|\mathcal{X}_{l_x+1}| \ln \frac{T|\mathcal{X}||\mathcal{A}|}{\delta}}{\max\{1, \mathbf{N}_i(x,a)\}}}$$

holds with probability at least $1 - \delta$ *simultaneously for all* $(x,a) \in \mathcal{X} \times \mathcal{A}$ *and all epochs.*

Since by the above result $P \in \hat{\mathcal{P}}_i$ holds with high probability for all epochs, defining our true confidence set $\mathcal{P}_i$ as

$$\mathcal{P}_i = \cap_{j=1}^{i} \hat{\mathcal{P}}_j, \tag{4}$$

we also have $P \in \mathcal{P}_i$ for all epochs with probability at least $1 - \delta$. This way we ensure that our confidence sets cannot increase between consecutive episodes with probability 1. Note that this is a delicate difference from the construction of Jaksch et al. (2010) that plays an important role in our proof.

4.2 Extended dynamic programming

FPOP needs to compute an optimistic transition function and an optimistic policy in each episode with respect to some reward function r and some confidence set $\mathcal{P}$ of transition functions. That is, we need to solve the problem

$$(\pi^*, P^*) = \arg\max_{\pi, \bar{P} \in \mathcal{P}} W(r, \pi, \bar{P}). \tag{5}$$

We will use an approach called *extended dynamic programming* to solve this problem, a simple adaptation of the extended value iteration method proposed by Jaksch et al. (2010). The method is presented as Algorithm 2 in the appendix. Computing in a backward manner on the states (that is, going from layer $\mathcal{X}_l$ to $\mathcal{X}_0$), the algorithm maximizes the transition probabilities to the direction of the largest reward-to-go. This is possible since the L_1-balls allow to select the optimistic transition functions independently for all state-action pairs $(x,a) \in \mathcal{X} \times \mathcal{A}$. Following the proof of Theorem 7 of Jaksch et al. (2010), Lemma 6 in the appendix shows that Algorithm 2 indeed solves the required minimization problem, and can be implemented with $\mathcal{O}(|\mathcal{A}||\mathcal{X}|^2)$ time and space complexity.

5 Analysis

The proof of Theorem 2 (and thus that of our main result, Theorem 1) mainly follows the regret analysis

of FPL combined with ideas from the regret analysis of UCRL-2. First, let us consider the policy-transition model pair

$$\left(\hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t\right) = \underset{\pi\in\Pi, \bar{P}\in\mathcal{P}_{\mathbf{i}(t)}}{\arg\max} \left\{W(R_t + \mathbf{Y}_{\mathbf{i}(t)}, \pi, \bar{P})\right\}.$$

In other words, $\hat{\boldsymbol{\pi}}_t$ is an optimistic policy that knows the reward function before the episode would begin. Define

$$\tilde{\mathbf{v}}_t = W(r_t, \boldsymbol{\pi}_t, \tilde{\mathbf{P}}_t) \quad \text{and} \quad \hat{\mathbf{v}}_t = W(r_t, \hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t).$$

Furthermore, let

$$\boldsymbol{\pi}_t^* = \underset{\pi\in\Pi}{\arg\max} \left\{W(R_t + \mathbf{Y}_{\mathbf{i}(t)}, \pi, P)\right\},$$

the optimal policy with respect to the perturbed rewards and the true transition function. The perturbation added to the value of policy π in episode t will be denoted by $\mathbf{Z}_t(\pi) = W(\mathbf{Y}_{\mathbf{i}(t)}, \pi, P)$. The true optimal value up to episode t and the optimal policy attaining this value will be denoted by

$$V_t^* = \max_{\pi\in\Pi} V_t(\pi) \text{ and } \sigma_t^* = \underset{\pi\in\Pi}{\arg\max}\, V_t(\pi).$$

We proceed by a series of lemmas to prove our main result. The first one shows that our optimistic choice of the estimates of the transition model enables us to upper bound the optimal value V_t^* with a reasonable quantity.

Lemma 2.

$$V_T^* \le \sum_{t=1}^{T} \mathbb{E}\left[\hat{\mathbf{v}}_t\right] + \delta\, T\, L + |\mathcal{X}||\mathcal{A}| \log_2\left(\frac{8T}{|\mathcal{X}||\mathcal{A}|}\right) \frac{\sum_{l=0}^{L-1} \ln\left(|\mathcal{X}_l||\mathcal{A}|\right) + L}{\eta}.$$

Note that while the proof of this result might seem a simple reproduction of some arguments from the standard FPL analysis, it contains subtle details about the role of our optimistic estimates that are of crucial importance for the analysis.

Proof. Assume that $P \in \mathcal{P}_{\mathbf{i}(T)}$, which holds with probability at least $1-\delta$, by Lemma 1. First, we have

$$\begin{aligned} V_T^* + \mathbf{Z}_T(\sigma_T^*) &\le V_T(\boldsymbol{\pi}_T^*) + \mathbf{Z}_T(\boldsymbol{\pi}_T^*) \\ &= W(R_T + \mathbf{Y}_{\mathbf{i}(T)}, \boldsymbol{\pi}_T^*, P) \\ &\le W(R_T + \mathbf{Y}_{\mathbf{i}(T)}, \hat{\boldsymbol{\pi}}_T, \hat{\mathbf{P}}_T) \end{aligned} \tag{6}$$

where the first inequality follows from the definition of $\boldsymbol{\pi}_T^*$ and the second from the optimistic choice of $\hat{\boldsymbol{\pi}}_T$ and $\hat{\mathbf{P}}_T$. Let $\mathbf{dY}_{\mathbf{i}(s)} = \mathbf{Y}_{\mathbf{i}(s)} - \mathbf{Y}_{\mathbf{i}(s-1)}$ for $s = 1, \dots, t$. Next we show that, given $P \in \mathcal{P}_{\mathbf{i}(T)}$,

$$W(R_t + \mathbf{Y}_{\mathbf{i}(t)}, \hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t) \le \sum_{s=1}^{t} W(r_s + \mathbf{dY}_{\mathbf{i}(s)}, \hat{\boldsymbol{\pi}}_s, \hat{\mathbf{P}}_s) \tag{7}$$

where we define $\mathbf{Y}_0 = 0$. The proof is done by induction on t. Equation (7) holds trivially for $t = 1$. For $t > 1$, assuming $P \in \mathcal{P}_{\mathbf{i}(T)}$ and (7) holds for $t-1$, we have

$$\begin{aligned} &W(R_t + \mathbf{Y}_{\mathbf{i}(t)}, \hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t) \\ &= W(R_{t-1} + \mathbf{Y}_{\mathbf{i}(t-1)}, \hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t) + W(r_t + \mathbf{dY}_{\mathbf{i}(t)}, \hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t) \\ &\le W(R_{t-1} + \mathbf{Y}_{\mathbf{i}(t-1)}, \hat{\boldsymbol{\pi}}_{t-1}, \hat{\mathbf{P}}_{t-1}) + W(r_t + \mathbf{dY}_{\mathbf{i}(t)}, \hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t) \\ &\le \sum_{s=1}^{t} W(r_s + \mathbf{dY}_{\mathbf{i}(s)}, \hat{\boldsymbol{\pi}}_s, \hat{\mathbf{P}}_s), \end{aligned}$$

where the first inequality follows from the fact that $\left(\hat{\boldsymbol{\pi}}_{t-1}, \hat{\mathbf{P}}_{t-1}\right)$ is selected from a wider class[3] than $\left(\hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t\right)$ and is optimistic with respect to rewards $R_{t-1} + \mathbf{Y}_{\mathbf{i}(t-1)}$, while the second inequality holds by the induction hypothesis for $t-1$. This proves (7).

Now the non-negativity of $\mathbf{Z}_T(\sigma_T^*)$, (6) and (7) imply that, given $P \in \mathcal{P}_{\mathbf{i}(T)}$,

$$\begin{aligned} V_T^* &\le \sum_{t=1}^{T} W(r_t + \mathbf{dY}_{\mathbf{i}(t)}, \hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t) \\ &= \sum_{t=1}^{T} \hat{\mathbf{v}}_t + \sum_{t=1}^{T} W(\mathbf{dY}_{\mathbf{i}(t)}, \hat{\boldsymbol{\pi}}_t, \hat{\mathbf{P}}_t). \end{aligned}$$

Since $P \in \mathcal{P}_{\mathbf{i}(T)}$ holds with probability at least $1-\delta$, $V_T^* \le TL$ trivially and the right hand side of the above inequality is non-negative, we have

$$V_T^* \le \sum_{t=1}^{T} \mathbb{E}\left[\hat{\mathbf{v}}_t\right] + \mathbb{E}\left[\sum_{t=1}^{T} W(\mathbf{dY}_{\mathbf{i}(t)}, \hat{\boldsymbol{\pi}}_{\mathbf{t}_j}, \hat{\mathbf{P}}_{\mathbf{t}_j})\right] + \delta TL \tag{8}$$

The elements in the second sum above may be non-zero only if $\mathbf{i}(t) \neq \mathbf{i}(t-1)$. Furthermore, by Proposition 18 of Jaksch et al. (2010), the number of epochs $\mathbf{K}_T$ up to episode T is bounded from above as

$$\mathbf{K}_T \overset{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}_{\{\mathbf{i}(t)\neq\mathbf{i}(t-1)\}} \le |\mathcal{X}||\mathcal{A}| \log_2\left(\frac{8T}{|\mathcal{X}||\mathcal{A}|}\right).$$

Therefore,

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} W(\mathbf{dY}_{\mathbf{i}(t)}, \hat{\boldsymbol{\pi}}_{\mathbf{t}_j}, \hat{\mathbf{P}}_{\mathbf{t}_j})\right] \le \mathbb{E}\left[\sum_{j=1}^{m} W(\mathbf{Y}_j, \hat{\boldsymbol{\pi}}_{\mathbf{t}_j}, \hat{\mathbf{P}}_{\mathbf{t}_j})\right] \\ &\le |\mathcal{X}||\mathcal{A}| \log_2\left(\frac{8T}{|\mathcal{X}||\mathcal{A}|}\right) \sum_{l=0}^{L-1} \mathbb{E}\left[\max_{(x,a)\in\mathcal{X}_l\times\mathcal{A}} \mathbf{Y}_1(x,a)\right]. \end{aligned}$$

[3] This follows from the definition of the confidence sets in Equation (4).

Using the upper bound on the expectation of the maximum of a number of exponentially distributed variables (see, e.g., the proof of Corollary 4.5 in Cesa-Bianchi and Lugosi 2006), a combination of the above inequality with (8) gives the desired result. □

Next, we show that peeking one episode into the future does not change the performance too much. The following lemma is a standard result used for the analysis of FPL and we include the proof in the appendix only for completeness.

Lemma 3. *Assume that* $\eta \le (|\mathcal{X}||\mathcal{A}|)^{-1}$. *Then,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \hat{\mathbf{v}}_t\right] \le \mathbb{E}\left[\sum_{t=1}^{T} \tilde{\mathbf{v}}_t\right] + \eta T\, L\,(e-1)|\mathcal{X}||\mathcal{A}|.$$

Now consider $\boldsymbol{\mu}_t(x) = \mathbb{P}\left[\mathbf{x}_{l_x} = x \middle| \mathbf{u} \sim (\boldsymbol{\pi}_t, P)\right]$, that is, the probability that a trajectory generated by $\boldsymbol{\pi}_t$ and P includes x. Note that given a layer $\mathcal{X}_l$, the restriction $\boldsymbol{\mu}_{t,l} : \mathcal{X}_l \to [0,1]$ is a distribution. Define an estimate of $\boldsymbol{\mu}_t$ as $\boldsymbol{\mu}_t(x) = \mathbb{P}\left[\mathbf{x}_l = x \middle| \mathbf{u} \sim (\boldsymbol{\pi}_t, \tilde{\mathbf{P}}_t)\right]$. Note that this estimate can be efficiently computed using the recursion

$$\tilde{\boldsymbol{\mu}}_t(x_{l+1}) = \sum_{x_l, a_l} \tilde{\mathbf{P}}_t(x_{l+1}|x_l, a_l)\boldsymbol{\pi}_t(a_l|x_l)\tilde{\boldsymbol{\mu}}_t(x_l),$$

for $l = 0, 1, 2, \ldots, L-1$, with $\tilde{\boldsymbol{\mu}}_t(x_0) = 1$. The following result will ensure that if our estimate of the transition function is close enough to the true transition function in the L_1 sense, than these estimates of the visitation probabilities are also close to the true values that they estimate. The proof follows from elementary algebraic manipulations and is included in the appendix for completeness.

Lemma 4. *Assume that there exists some function* $\mathbf{a}_t : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ *such that*

$$\left\|\tilde{\mathbf{P}}_t(\cdot|x,a) - P(\cdot|x,a)\right\|_1 \le \mathbf{a}_t(x,a)$$

holds for all $(x,a) \in \mathcal{X} \times \mathcal{A}$. *Then*

$$\sum_{x_l \in \mathcal{X}_l} |\tilde{\boldsymbol{\mu}}_t(x_l) - \boldsymbol{\mu}_t(x_l)| \le \sum_{k=0}^{l-1} \sum_{x_k \in \mathcal{X}_k} \boldsymbol{\mu}_t(x_k)\, \mathbf{a}_t\left(x_k, \boldsymbol{\pi}_t(a_k)\right)$$

for all $l = 1, 2, \ldots, L-1$.

Finally, we use the above result to relate the estimated policy values $\tilde{\mathbf{v}}_t$ to their true values $v_t(\boldsymbol{\pi}_t)$. The following lemma, largely based on Lemma 19 of Jaksch et al. (2010), is proved in the appendix.

Lemma 5. *Assume* $T \ge |\mathcal{X}||\mathcal{A}|$. *Then*

$$\mathbb{E}\left[\sum_{t=1}^{T} \tilde{\mathbf{v}}_t\right] \le \mathbb{E}\left[\sum_{t=1}^{T} v_t(\boldsymbol{\pi}_t)\right] + L|\mathcal{X}|\sqrt{2\,T\,\ln\frac{L}{\delta}} + 2\delta\, T\, L + \left(\sqrt{2}+1\right) L|\mathcal{X}|\sqrt{T\,|\mathcal{A}|\ln\frac{T|\mathcal{X}||\mathcal{A}|}{\delta L}}.$$

6 Conclusions and future work

We have considered the problem of learning in adversarial stochastic shortest path problems when the transition probabilities are unknown. We proposed an algorithm achieving optimal regret rates in terms of the time horizon, that is (to our knowledge) the first one that learns stochastic and adversarial components in an online fashion at the same time. The algorithm is a novel combination of common techniques used for online learning in stochastic and adversarial environments – namely, a combination of the "optimism in the face of uncertainty" principle and the "follow the perturbed leader" technique.

In the process of proving our results, much structure (of both the standard MDP learning problem and the online learning problem) was uncovered that was not transparent in previous works. In particular, it has become apparent that the *size* of confidence intervals for the unknown model parameters only influences the quality of the estimate of the *achieved* performance (cf. Lemma 5), while selecting our models optimistically helps only in estimating the *best possible* performance (cf. Lemma 2).

An interesting open question is whether the local-forecasting based approach of Neu et al. (2010a), pioneered by Even-Dar et al. (2005, 2009), can be extended to our problem. Another interesting question is whether the FPL approach of Yu et al. (2009) can also be extended (and their regret bounds strengthened). Finally, in this paper we considered the problem under the assumption that the regret function becomes fully known at the end of each episode. However, in some applications the agent only learns the rewards along the trajectory it traversed, a.k.a the "bandit setting". It remains an interesting and important open problem to extend the results obtained in this paper to this setting. Another direction for future work is to consider the non-episodic variants of the problem, studied in many of the previously mentioned papers and more recently, in the bandit setting, by Neu et al. (2010b).

Acknowledgements

This work was supported in part by the Hungarian Scientific Research Fund and the Hungarian National Office for Research and Technology (OTKA-NKTH CNK 77782), the PASCAL2 Network of Excellence under EC grant no. 216886, NSERC, AITF, the Alberta Ingenuity Centre for Machine Learning, the DARPA GALE project (HR0011-08-C-0110) and iCore.

References

Abbasi-Yadkori, Y. and Szepesvári, Cs. (2011). Regret bounds for the adaptive control of linear quadratic systems. In *COLT-11*.

Bartlett, P. L. and Tewari, A. (2009a). REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI 2009*.

Bartlett, P. L. and Tewari, A. (2009b). REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*.

Brafman, R. I. and Tennenholtz, M. (2002). R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA.

Even-Dar, E., Kakade, S. M., and Mansour, Y. (2005). Experts in a Markov decision process. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 401–408.

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

György, A., Linder, T., Lugosi, G., and Ottucsák, Gy. (2007). The on-line shortest path problem under partial monitoring. *J. Mach. Learn. Res.*, 8:2369–2403.

Hannan, J. (1957). Approximation to Bayes risk in repeated play. *Contributions to the theory of games*, 3:97–139.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 99:1563–1600.

Kakade, S. (2003). *On the sample complexity of reinforcement learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

Kakade, S., Kearns, M. J., and Langford, J. (2003). Exploration in metric state spaces. In Fawcett, T. and Mishra, N., editors, *ICML 2003*, pages 306–312. AAAI Press.

Kalai, A. and Vempala, S. (2005). Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.*, 71:291–307.

Lazaric, A. and Munos, R. (2011). Learning with stochastic inputs and adversarial outputs. *To appear in Journal of Computer and System Sciences*.

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

Neu, G., György, A., Szepesvári, Cs., and Antos, A. (2010b). Online Markov decision processes under bandit feedback. In *NIPS-10*.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience.

Strehl, A. L. and Littman, M. L. (2008). Online linear regression and its application to model-based reinforcement learning. In *NIPS-20*, pages 1417–1424. MIT Press.

Szepesvári, Cs. (2010). *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the l1 deviation of the empirical distribution.

Yu, J. Y. and Mannor, S. (2009a). Arbitrarily modulated Markov decision processes. In *Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference*, pages 2946–2953. IEEE Press.

Yu, J. Y. and Mannor, S. (2009b). Online learning in Markov decision processes with arbitrarily changing rewards and transitions. In *GameNets'09: Proceedings of the First ICST international conference on Game Theory for Networks*, pages 314–322, Piscataway, NJ, USA. IEEE Press.

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