Online Learning in Markov Decision Processes with Arbitrarily Changing Rewards and Transitions*

Jia Yuan Yu and Shie Mannor
McGill University
jia.yu@mcgill.ca; shie.mannor@mcgill.ca

We consider problems where an agent controls a system subject to uncertainty. The system under control is modelled as a Markov decision process (MDP). Markov decision processes (MDPs) model a number of practical problems [2]. We consider a generalized version of MDPs that captures an additional feature of real-life problems: the rewards and transition probabilities may change over time in an arbitrary and non-stationary manner (*e.g.*, under the influence of an adversary or Nature). This generalized version is reminiscent of stochastic games [3]. However, in stochastic games, one usually assumes that there is an adversary whose utility is well-defined. In our setup, as in [4], [5], and more generally in the online learning setting [6], we do not assume that such an adversary exists, but rather that the changes in the environment occur arbitrarily.

Uncertainty can be intrinsic and unavoidable to the system (probabilistic transitions, uncertainty principles), or may arise from measurements. In many decision-making problems, uncertainty exists in the rewards and transition probabilities (cf. [7] and references therein). When this uncertainty follows a stochastic model [8, 9], sampling can give estimates on the parameters of the transition model, but some residual uncertainty always remains as a result of limited samples. Under these circumstances, if it is imperative to take precautions against the worst possible occurrence, then a standard approach is to handle this uncertainty through *robust optimization*. This approach typically assumes that the uncertainty has a *fixed*—albeit unknown—realization and constructs a policy that performs well in the worst case. Since the uncertainty evolves in a non-stationary fashion in our setting, the standard robust approach does not offer a satisfying solution. In particular, it only guarantees optimal performance against the *worst* realization of the *environment*. It does not promise anything about the relative performance compared to the *best* alternative *policy* in *hindsight*. The goal of this paper is to address this *relative* performance.

In this paper, we distinguish between two notions of uncertainty: arbitrary variations in the *reward function*, and arbitrary, but bounded, variations in the *transition probabilities*. These notions have been studied separately. Online learning in MDPs makes the solution robust against arbitrary variation in the reward functions when the transition probabilities are fixed [4, 10]. Robust dynamic programming has been used to control MDPs where the transition probabilities may vary arbitrarily, but where the reward functions may not [7]. This is the first work to address both uncertainties simultaneously. To this end, we propose a new approach that uses online learning to adapt time-varying reward functions and uses robust dynamic programming to deal with the range of possible transition models.

*A version of this work will be presented at GameNets 2009 [1].

To model uncertainty, the reward and transition functions are defined by arbitrary individual sequences. First, we show through an example how arbitrary change in transition probabilities, no matter how small, can cause large performance loss for online learning algorithms that compute policies using a nominal transition function. To protect against such breakdowns, we present algorithms that combine online learning and robust control, and then establish guarantees on their performance evaluated in retrospect against alternative policies—*i.e.*, their regret. These guarantees depend critically on the range of uncertainty in the transition probabilities, but otherwise hold universally for all sequences of reward and transition functions. We present a version of the main algorithm in the setting where the decision-maker's observations are limited to its state-action trajectory, and another version that allows a trade-off between performance and computational complexity.

It is important to note that, in contrast to most online learning settings, the average regret does *not* vanish with time when the transition probabilities can change arbitrarily over time [4]. Therefore, the setting of interest is where the uncertainty in the transition probabilities is limited, whereas the uncertainty in the rewards is not. Our results will quantify the extent to which the regret may grow as the transition probabilities are given more leeway to change.

References

[1] J. Y. Yu and S. Mannor, "Online learning in Markov decision processes with arbitrarily changing rewards and transitions," in *GameNets*, 2009, http://www.cim.mcgill.ca/~jiayuan/olsg.pdf.

[2] M. L. Puterman, *Markov Decision Processes.* Wiley, 1994.

[3] L. Shapley, "Stochastic games," *PNAS*, vol. 39, no. 10, pp. 1095–1100, 1953.

[4] E. Even-Dar, S. Kakade, and Y. Mansour, "Experts in a Markov decision process," in *NIPS*, 2004, pp. 401–408.

[5] S. Mannor and N. Shimkin, "The empirical Bayes envelope and regret minimization in competitive Markov decision processes," *Mathematics of Operations Research*, vol. 28, no. 2, pp. 327–345, 2003.

[6] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games.* Cambridge University Press, 2006.

[7] A. Nilim and L. E. Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[8] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[9] R. I. Brafman and M. Tennenholtz, "R-max—a general polynomial time algorithm for near-optimal reinforcement learning," *Journal of Machine Learning Research*, vol. 3, pp. 213–231, 2003.

[10] J. Y. Yu, S. Mannor, and N. Shimkin, "Markov decision processes with arbitrary reward processes," in *Lecture Notes in Computer Science*, vol. 5323, 2009.