JMLR: Workshop and Conference Proceedings 10 (2010) 1–2 Workshop on Active Learning and Experimental Design

Generating Informative Trajectories by using Bounds on the Return of Control Policies

Raphael Fonteneau RAPHAEL.FONTENEAU@ULG.AC.BE
University of Liège, Belgium

Susan A. Murphy SAMURPHY@UMICH.EDU
University of Michigan, USA

Louis Wehenkel L.WEHENKEL@ULG.AC.BE
University of Liège, Belgium

Damien Ernst DERNST@ULG.AC.BE
University of Liège, Belgium

Editor: I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov

Abstract

We propose new methods for guiding the generation of informative trajectories when solving discrete-time optimal control problems. These methods exploit recently published results that provide ways for computing bounds on the return of control policies from a set of trajectories.

Keywords: reinforcement learning, optimal control, sampling strategies

Introduction. Discrete-time optimal control problems arise in many fields such as finance, medicine, engineering as well as artificial intelligence. Whatever the techniques used for solving such problems, their performance is related to the amount of information available on the system dynamics and the reward function of the optimal control problem. In this paper, we consider settings in which information on the system dynamics must be inferred from trajectories and, furthermore, due to cost and time constraints, only a limited number of trajectories can be generated. We assume that a regularity structure - given in the form of Lipschitz continuity assumptions - exists on the system dynamics and the reward function. Under such assumptions, we exploit recently published methods for computing bounds on the return of control policies from a set of trajectories (Fonteneau et al. (2009, 2010)) in order to sample the state-action space so as to be able to discriminate between optimal and non-optimal policies.

Problem statement. We consider a discrete-time system whose dynamics over T stages is described by a time-invariant equation $x_{t+1} = f(x_t, u_t) \quad t = 0, \ldots, T-1$, where for all t, the state x_t is an element of the compact normed state space $\mathcal{X}$ and u_t is an element of the finite (discrete) action space $\mathcal{U}$. $T \in \mathbb{N}_0$ is referred to as the optimization horizon. An instantaneous reward $r_t = \rho(x_t, u_t) \in \mathbb{R}$ is associated with the action u_t taken while being in state x_t. The initial state of the system is fixed to $x_0 \in \mathcal{X}$. For every policy $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, the $T-$stage return of $(u_0, \ldots, u_{T-1})$ is defined as $J^{u_0,\ldots,u_{T-1}}(x_0) = \sum_{t=0}^{T-1} \rho(x_t, u_t)$, where $x_{t+1} = f(x_t, u_t)$,$\forall t \in \{0, \ldots, T-1\}$. An optimal policy is a policy that belongs to $\underset{(u_0,\ldots,u_{T-1}) \in \mathcal{U}^T}{\arg\max} \{J^{u_0,\ldots,u_{T-1}}(x_0)\}$. Here, f and ρ are Lipschitz continuous, and we have access to two constants $L_f, L_\rho > 0$ such that $\forall (x', x'') \in \mathcal{X}^2, \forall u \in \mathcal{U}, \|f(x',u) - f(x'',u)\|_{\mathcal{X}} \leq L_f \|x' - x''\|_{\mathcal{X}}$ and $|\rho(x',u) - \rho(x'',u)| \leq L_\rho \|x' - x''\|_{\mathcal{X}}$, where $\|.\|_{\mathcal{X}}$ denotes the chosen norm over the space $\mathcal{X}$. Initially, the values of f and ρ are only known for n state-action pairs. These values are given in a set of one-step transitions $\mathcal{F}_n = \{(x^l, u^l, r^l, y^l)\}_{l=1}^n$ where $y^l = f(x^l, u^l)$ and $r^l = \rho(x^l, u^l)$. We suppose that additional transitions can be sampled, and we detail hereafter a sampling strategy to select state-action pairs (x, u) for generating $f(x, u)$ and $\rho(x, u)$ so as to be able to discriminate rapidly $-$ as new one-step transitions are generated $-$ between optimal and non-optimal policies.

Algorithm. Fonteneau et al. (2010) proposes a method for computing from any set of transitions $\mathcal{F}$ such that each action $u \in \mathcal{U}$ appears at least once in $\mathcal{F}$ and for any policy $(u_0, \ldots, u_{T-1})$ a lower bound $L_{\mathcal{F}}^{u_0,\ldots,u_{T-1}}(x_0)$ and an upper bound $U_{\mathcal{F}}^{u_0,\ldots,u_{T-1}}(x_0)$ on $J^{u_0,\ldots,u_{T-1}}(x_0)$. Furthermore, these bounds converge towards $J^{u_0,\ldots,u_{T-1}}(x_0)$ when the sparsity of $\mathcal{F}$ decreases towards zero. Before describing our proposed sampling strategy, note that a policy can only be optimal given a set of one-step transitions $\mathcal{F}$ if its upper bound is not lower than the lower bound of any element of $\mathcal{U}^T$. We denote by $\Pi(\mathcal{F})$ the set of policies which, given $\mathcal{F}$, satisfy this property. Our sampling strategy generates new one-step transitions iteratively. Given an existing set $\mathcal{F}_m$ of m one-step transitions, which is made of the elements of the initial set $\mathcal{F}_n$ and the m-n one-step transitions generated during the first m-n iterations of this algorithm, it selects as next sampling point $(x^{m+1}, u^{m+1}) \in \mathcal{X} \times \mathcal{U}$, the point that minimizes in the worst conditions the largest bound width among the candidate optimal policies at the next iteration:

$$(x^{m+1}, u^{m+1}) \in \underset{(x,u)\in X\times U}{\arg\min} \left\{ \max_{\substack{(r,y)\in\mathbb{R}\times\mathcal{X}\ s.t.\\ (x,u,r,y)\in\mathcal{C}(\mathcal{F}_m)}} \left\{ \max_{\substack{(u_0,\ldots,u_{T-1})\in\\ \Pi(\mathcal{F}_m\cup\{(x,u,r,y)\})}} \Delta^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_m\cup\{(x,u,r,y)\}}(x_0) \right\} \right\}$$

where $\Delta_{\mathcal{F}}^{u_0,\ldots,u_{T-1}}(x_0) = U_{\mathcal{F}}^{u_0,\ldots,u_{T-1}}(x_0) - L_{\mathcal{F}}^{u_0,\ldots,u_{T-1}}(x_0)$ and $\mathcal{C}(\mathcal{F}) \subset \mathcal{X} \times \mathcal{U} \times \mathbb{R} \times \mathcal{U}$ gathers all transitions that are compatible with the set of transitions $\mathcal{F}$. A transition (x, u, r, y) is said compatible with the set of transitions $\mathcal{F}$ if $\forall (x^l, u^l, r^l, y^l) \in \mathcal{F}$ satisfying $u^l = u$, $\left|r - r^l\right| \leq L_\rho \|x - x^l\|_{\mathcal{X}}$ and $\left\|y - y^l\right\|_{\mathcal{X}} \leq L_f \|x - x^l\|_{\mathcal{X}}$. Based on the convergence properties of the bounds, we conjecture that the sequence $(\Pi(\mathcal{F}_m))_{m\in\mathbb{N}}$ converges towards the set of all optimal policies in a finite number of iterations. The analysis of the theoretical properties of the sampling strategy and its empirical validation are left for future work.

Acknowledgements. Raphael Fonteneau acknowledges the financial support of the FRIA (Fund for Research in Industry and Agriculture). Damien Ernst is a research associate of the FRS-FNRS. This paper presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. We also acknowledge financial support from NIH grants P50 DA10075 and R01 MH080015. The scientific responsibility rests with its authors.

References

R. Fonteneau, S. Murphy, L. Wehenkel, and D. Ernst. Inferring bounds on the performance of a control policy from a sample of trajectories. In *Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (IEEE ADPRL 09)*, Nashville, TN, USA, 2009.

R. Fonteneau, S.A. Murphy, L. Wehenkel, and D. Ernst. A cautious approach to generalization in reinforcement learning. In *Proceedings of the Second International Conference on Agents and Artificial Intelligence (ICAART 2010)*, Valencia, Spain, 2010.