Anytime-Competitive Reinforcement Learning with Policy Prior

JIANYI YANG, University of California, Riverside, United States
PENGFEI LI, University of California, Riverside, United States
TONGXIN LI, The Chinese University of Hong Kong, Shenzhen, China
ADAM WIERMAN, California Institute of Technology, United States
SHAOLEI REN, University of California, Riverside, United States

ACM Reference Format:
Jianyi Yang, Pengfei Li, Tongxin Li, Adam Wierman, and Shaolei Ren. 2018. Anytime-Competitive Reinforcement Learning with Policy Prior. 1, 1 (May 2018), 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 PROBLEM FORMULATION

We study a novel MDP problem called Anytime-Competitive Markov Decision Process (A-CMDP). In A-CMDP, each episode has H rounds. The state at each round is denoted as $x_h \in \mathcal{X}, h \in [H]$. At each round of an episode, the agent selects an action a_h from an action set $\mathcal{A}$. The environment generates a reward $r_h(x_h, a_h)$ and a cost $c_h(x_h, a_h)$ with $r_h \in \mathcal{R}$ and $c_h \in \mathcal{C}$. We model the dynamics as $x_{h+1} = f_h(x_h, a_h)$ where $f_h \in \mathcal{F}$ is a random transition function drawn from an unknown distribution $g(f_h)$ with the density $g \in \mathcal{G}$. The agent has no access to the random function f_h but can observe the state x_h at each round h. Thus, the state transition model can be expressed as $\mathbb{P}(x_{h+1} \mid x_h, a_h) = \sum_{f_h} \mathbb{1}(f_h(x_h, a_h) = x_{h+1}) g(f_h)$. Let $V_h^\pi(x) = \mathbb{E}\left[\sum_{i=h}^H r_i(x_i, a_i)) \mid x_h = x\right]$ denote the expected value of the total reward from round h by a policy π. One objective of A-CMDP is to maximize the expected total reward starting from the first round which is denoted as $\mathbb{E}_{x_1}\left[V_1^\pi(x_1)\right] = \mathbb{E}\left[\sum_{h=1}^H r_h(x_h, a_h))\right]$. Besides A-CMDP guarantees the anytime competitive cost constraints compared with a policy prior $\pi^\dagger$. Denote $y_h = (f_h, c_h, r_h)$, and $y_{1:H} = \{y_h\}_{h=1}^H \in \mathcal{Y} = \mathcal{F} \times \mathcal{R} \times \mathcal{C}$ is a sampled sequence. For any round h in any model sequence $y_{1:H} \in \mathcal{Y}$, the anytime competitive constraints require that $J_h^\pi(y_{1:H}) \leq (1+\lambda) J_h^{\pi^\dagger}(y_{1:H}) + hb$. where $J_h^\pi(y_{1:H}) = \sum_{i=1}^h c_i(x_i, a_i)$ be the cost up to round $h \in [H]$ with states $x_i, i \in [h]$ and actions $a_i, i \in [h]$ of a policy π, and parameters are $\lambda \geq 0$ and $b \geq 0$. The objective of A-CMDP is

$$\max_{\pi \in \Pi} \mathbb{E}_{x_1}\left[V_1^\pi(x_1)\right], \quad s.t. \;\; J_h^\pi(y_{1:H}) \leq (1+\lambda) J_h^{\pi^\dagger}(y_{1:H}) + hb, \;\; \forall h \in [H], \forall y_{1:H} \in \mathcal{Y}. \tag{1}$$

Assumption 1.1. The cost functions have a minimum value $\epsilon \geq 0$, i.e. $\forall (x, a), \forall h \in [H], c_h(x, a) \geq \epsilon \geq 0$, and are L_c-Lipschitz continuous. The transition functions are L_f-Lipschitz continuous. The parameters ϵ, L_c and L_f are known to the agent.

Assumption 1.2. The prior policy $\pi^\dagger$ is Lipschitz continuous and satisfies the telescoping property, i.e. if $\pi^\dagger$ is applied from round h_1 to h_2 with initialized states x_{h_1} and x'_{h_1}, it holds at round h_2 that $\|x_{h_2} - x'_{h_2}\| \leq p(h_2 - h_1)\|x_{h_1} - x'_{h_1}\|$, where $p(h)$ is a perturbation function with $p(0) = 1$.

This extended abstract summarizes the paper [1].

Authors' Contact Information: Jianyi Yang, University of California, Riverside, United States; Pengfei Li, University of California, Riverside, United States; Tongxin Li, The Chinese University of Hong Kong, Shenzhen, China; Adam Wierman, California Institute of Technology, United States; Shaolei Ren, University of California, Riverside, United States.

ACM XXXX-XXXX/2018/5-ART
https://doi.org/XXXXXXX.XXXXXXX

2 MAIN RESULTS

First, we propose an Anytime-Competitive Decision-making (ACD) algorithm to provably guarantee the anytime competitive constraints for each episode. The key idea to satisfy the anytime competitive constraints is a projection to a safe action set $\mathcal{A}_h(D_h)$ in each round. The design of the safe action set has the following two challenges. First, since in MDPs, the agent can only observe the *real* states $\{x_h\}_{h=1}^H$ corresponding to the truly-selected actions $\{a_h\}_{h=1}^H$, the agent cannot evaluate the prior cost $J_h^{\pi^\dagger}$ in the anytime competitive constraint at each round h. Second, The impacts of the actions on future costs are based on random transition models f_i. Thus, besides satisfying the constraints in the current round, we need to have a good planning for the future rounds to avoid any possible constraint violations without the exact knowledge of transition and/or cost models. To address the challenges, we derive a sufficient condition for the satisfaction of the anytime competitive constraints. We prove that $(\lambda, b)-$anytime competitive constraints are satisfied if it holds at each round h that

$$\Gamma_{h,h}\|a_h - \pi^\dagger(x_h)\| \leq D_h, \tag{2}$$

where $D_1 = \lambda\epsilon+b$, $D_h = \max\left\{D_{h-1} + \lambda\epsilon + b - \Gamma_{h-1,h-1}d_{h-1},\ R_{h-1} + \lambda\epsilon + b\right\}$ is the allowed deviation at round h, and the parameters are calculated as $\Gamma_{j,n} = \sum_{i=n}^H q_{j,i}, (j \in [H], \forall n \geq j)$, with $q_{j,i} = L_c\mathbb{1}(j=i)+L_c(1+L_{\pi^\dagger})L_f p(i-1-j)\mathbb{1}(j<i), (\forall j \in [H], i \geq j), R_{h-1} = \sum_{i=1}^{h-1}\left((1+\lambda)\hat{c}_i^\dagger - c_i - \Gamma_{i,h}d_i\right)$, $\hat{c}_i^\dagger = \max\left\{\epsilon, c_i - \sum_{j=1}^i q_{j,i}d_j\right\}, (\forall i \in [H])$.

Then, we develop a new RL algorithm (ACRL) to optimize the average reward while satisfying the anytime competitive constraints. The anytime constrained inference ACD actually define a new MDP where it is the projected action instead of the ML output that directly interact with the environment. Thus, we design a model-based RL framework which learns the dynamic distribution g by interacting with the new environment defined by ACD. At each inference, the policy based on the learned dynamic distribution is projected to meet the constraints in (2) which further leads to the satisfaction of the anytime constraints.

We rigorously prove that the anytime competitive constraints are satisfied, and analyze the reward regret of ACRL compared with the optimal-unconstrained policy.

Theorem 2.1. *The $(\lambda, b)-$anytime competitive constraints are satisfied if (2) holds for each round h.*

Theorem 2.2. *Assume that the optimal-unconstrained policy π^* has a value function $Q_h^{\pi^*}(x, a)$ which is $L_{Q,h}$-Lipschitz continuous with respect to the action a for all x. The regret between the optimal ACD policy and the optimal-unconstrained policy π^* is bounded as*

$$\mathbb{E}_{x_1}\left[V_1^{\pi^*}(x_1) - V_1^{\pi^\circ}(x_1)\right] \leq \mathbb{E}_{y_{1:H}}\left\{\sum_{h=1}^H L_{Q,h}\left[\eta - \frac{1}{\Gamma_{h,h}}(\lambda\epsilon + b + \Delta G_h)\right]^+\right\}, \tag{3}$$

where $\eta = \sup_{x\in\mathcal{X}}\|\pi^*(x) - \pi^\dagger(x))\|$ is the maximum action discrepancy between the policy prior $\pi^\dagger$ and optimal-unconstrained policy π^*, and $\Delta G_h = [R_{h-1}]^+$.

The analysis shows that there exists a fundamental trade-off between the optimization of the average reward and the satisfaction of the anytime competitive constraints. We further bound the regret of ACRL comparing against the optimal policy, which shows the ACD policy performs as asymptotically well as the optimal ACD policy as episode number $K \to \infty$.

REFERENCES

[1] Jianyi Yang, Pengfei Li, Tongxin Li, Adam Wierman, and Shaolei Ren. 2024. Anytime-Competitive Reinforcement Learning with Policy Prior. *Advances in Neural Information Processing Systems* 36 (2024).