Identification of Blackwell Policies for Deterministic MDPs

Victor Boone[1], Bruno Gaujal[2]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
victor.boone@univ-grenoble-alpes.fr

[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
bruno.gaujal@univ-grenoble-alpes.fr

Keywords : *Reinforcement Learning, Markov Decision Processes, Blackwell Optimality*

We consider the problem of the identification of Blackwell optimal policies for deterministic finite Markov Decision Processes (d-MDPs). Specifically, we are interested in algorithms that learn reward distributions by querying samples over time, that stop almost surely and return a Blackwell optimal policy with high probability. We provide a characterization of the class of MDPs over which such algorithms exist together with an algorithm identifying Blackwell optimal policies with arbitrarily high probability.

1 Blackwell Optimality & Identification Algorithms

Blackwell optimality. A *deterministic Markov Decision Process* (d-MDP) M is given by a state space $\mathcal{S}$, action space $\mathcal{A}$ with reward distributions $q(x,a) \in \mathcal{P}([0,1])$ and degenerate transition distributions, that is, $\forall (x,a,y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}, P(y|s,a) \in \{0,1\}$. In general, given $x, y \in \mathcal{S}$, there may be distincts actions a, b such that $P(y|x,a) = P(y|x,b) = 1$ but up to a state-wise elimination of actions, transitions may be seen as edges of a graph. One can think of $P(y|x,a) = 1$ as an edge (x,y) and write $q(x,y)$ for $q(x,a)$. The set of edges $\mathcal{E} := \{(x,y) \mid \exists a, P(y|x,a) = 1\}$ will be called *edge space*.

Upon choosing an edge (x,y) from a state x, the system changes state to y and produces a reward $r \sim q(x,y)$. A policy is an application $\pi : \mathcal{S} \to \mathcal{E}$ that, from each state $x \in \mathcal{S}$, selects an outgoing edge $\pi(x) \in \{(x,y) \in \mathcal{E} \mid y \in \mathcal{S}\}$. Iterating a policy over time gives a sequence of state-reward pairs $(x_t, r_t), t \geq 0$. Policies are usually discriminated with respect to the reward they score either at discounted infinite horizon $\nu_\beta^\pi(x_0) := \mathbb{E}_{x_0}^{M,\pi}[\sum_{t=0}^\infty \beta^t r_t]$ or at undiscounted infinite horizon $g_\pi(x_0) := \mathbb{E}_{x_0}^{M,\pi}[\lim \frac{1}{T}\sum_{t=0}^{T-1} r_t]$. The discounted and undiscounted infinite settings are linked by the Laurent Serie Expansion

$$\nu_\beta^\pi(x) = \frac{g_\pi(x)}{1-\beta} + h_\pi(x) + \sum_{n=1}^\infty h_\pi^{(n)}(x)(1-\beta)^n \tag{1}$$

when β is close enough to 1, see [1]. It is known that as $\beta \to 1$, the class of discounted optimal policies stabilises onto a single class $\Pi_\infty^*(M)$, called *Blackwell optimal policies*. They are policies which maximize the whole vector $(g_\pi(x), h_\pi(x), h_\pi^{(1)}(x), h_\pi^{(2)}(x), \ldots)$ for the lexicographic order. Namely, they maximize the asymptotical average reward or *gain* $g_\pi(x)$, but also the transient rewards i.e. the *bias* $h_\pi(x)$ and all higher order biases $h_\pi^{(n)}(x)$. Blackwell optimality is the last refinement of infinite horizon optimality that merges the discounted and undiscounted cases. When M is given, algorithms that compute Blackwell optimal policies are already known, see [1]. Our interest is to figure out if such policies can be *learned*.

Probably Correct Identification Algorithms. We are interested in the *identification* of Blackwell optimal policies in the *generative model* in a similar fashion as best-arm identification algorithms for stochastic bandits [2]. By generative model, we mean that at each time step, the algorithm is allowed to sample any edge in edge space. An identification algorithm $\mathcal{I}$ is made of three components:

- an *allocation rule* that chooses, according to past observations, the next edge (x_t, y_t) to be sampled;
- a *stopping rule* τ_δ to stop the learning phase of the algorithm;
- a *recommendation rule* to return a policy $\pi^{\mathcal{A}}_{\tau_\delta}$ at the end of the learning phase.

If $\mathcal{M}$ is a class of MDP, $\mathcal{I}$ is said to be δ-PC on $\mathcal{M}$ if when executed on any $M \in \mathcal{M}$, it returns a Blackwell optimal policy with probability at least $1-\delta$. Recent works [3] have designed identification algorithms for the discounted setting. The undiscounted setting remained open.

2 An Identification Algorithm for Blackwell Optimality

The learning of Blackwell optimal policies is limited to a specific class of d-MDP that we denote $\mathcal{M}$, defined as the set of d-MDPs M such that :

(H1) *M has a unique optimal cycle i.e. the cycle $\mathcal{C}_* \subseteq \mathcal{E}$ that maximize its average expected reward $g(\mathcal{C}_*) := \frac{1}{|C_*|}\sum_{e\in\mathcal{C}_*} r(e)$ is unique.*

(H2) Under H1, *writing $\mathcal{C}_*$ as the sequence of states $u_0, u_1, \ldots, u_{c-1}$, we define the bias of a state $u_i \in \mathcal{C}_*$ as $h_*(u_i) := \frac{1}{c}\sum_{\ell=1}^{c}\sum_{k=0}^{\ell}[r(u_{i+k}, u_{i+k+1}) - g(\mathcal{C}_*)]$ where indices are taken modulo c. Then for all state x_0, there exists a unique path $(x_0, x_1, \ldots, x_k)$ to $\mathcal{C}_*$ that maximizes $h_*(x_k) + \sum_{i=0}^{k-1}[r(x_i, x_{i+1}) - g(\mathcal{C}_*)]$, $x_k \in \mathcal{C}_*$.*

These two assumptions are minimal. Specifically, we can show that if $\mathcal{M}'$ is a space of d-MDPs such that $\mathcal{M}' \cap \mathcal{M}^{\complement} \neq \varnothing$, there is no $\frac{1}{4}S^{-A}$-PC identification algorithm on $\mathcal{M}'$. Finally, we propose a δ-PC identification algorithm on $\mathcal{M}$ for any $\delta > 0$.

Theorem 1 *Consider the algorithm $\mathcal{I}$ that samples edges uniformly with stopping time*

$$\tau_\delta := \inf\left\{ t \geq A \;:\; \left(\frac{\frac{1}{2}\log(\frac{2At^2}{\delta})}{\lfloor t/A \rfloor}\right)^{\frac{1}{2}} \leq \min\left\{\frac{\Delta_0(\hat{M}_t)}{4S}, \frac{\Delta_{-1}(\hat{M}_t)}{2}\right\}\right\} \quad (2)$$

where $\hat{M}_t$ is the MDP of empirical observations up to time t and $\Delta_0(\hat{M}_t)$, $\Delta_{-1}(\hat{M}_t)$ are MDP dependent parameters and that returns any $\pi^{\mathcal{I}}_{\tau_\delta} \in \Pi^*_\infty(\hat{M}_{\tau_\delta})$. Then $\mathcal{I}$ is δ-PC and stops almost surely. In addition, setting $\Delta(M) := \min\{\frac{1}{8S}\Delta_0(M), \frac{1}{4}\Delta_{-1}(M)\}$,

$$\mathbb{P}\left\{\tau_\delta \leq 3A\max\left\{1, \tfrac{1}{2}\log(\tfrac{2A}{\delta})\Delta(M)^{-2}, 3A\Delta(M)^{-4}\right\}\right\} \geq 1-\delta. \quad (3)$$

This uniform algorithm can be improved into a faster, non-uniform one. Moreover, these results can be generalized to MDPs with general probabilistic transitions.

References

[1] Puterman, M. L. (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* John Wiley and Sons, Inc., USA, 1st edition.

[2] Kaufmann, E., Cappé, O., and Garivier, A. (2016) On the complexity of best arm identification in multi-armed bandit models.

[3] Marjani, A. A. and Proutiere, A. (2021) Adaptive sampling for best policy identification in markov decision processes.