Reward Shaping for Reinforcement Learning with Omega-Regular Objectives

Ernst Moritz Hahn[1,2], Mateo Perez[3], Sven Schewe[4],
Fabio Somenzi[3], Ashutosh Trivedi[3], and Dominik Wojtczak[4]

[1] School of EEECS, Queens University Belfast, UK
[2] State Key Laboratory of Computer Science, Institute of Software, CAS, PRC
[3] University of Colorado Boulder, USA
[4] University of Liverpool, UK

Abstract. Recently, successful approaches have been made to exploit good-for-MDPs automata—Büchi automata with a restricted form of nondeterminism—for model free reinforcement learning, a class of automata that subsumes good for games automata and the most widespread class of limit deterministic automata [3]. The foundation of using these Büchi automata is that the Büchi condition can, for good-for-MDP automata, be translated to reachability [2].
The drawback of this translation is that the rewards are, on average, reaped very late, which requires long episodes during the learning process. We devise a new reward shaping approach that overcomes this issue. We show that the resulting a model is equivalent to a discounted payoff objective with a biased discount that simplifies and improves on [1].

1 Preliminaries

A *nondeterministic Büchi automaton* is a tuple $\mathcal{A} = \langle \Sigma, Q, q_0, \Delta, \Gamma \rangle$, where Σ is a finite *alphabet*, Q is a finite set of *states*, $q_0 \in Q$ is the *initial state*, $\Delta \subseteq Q \times \Sigma \times Q$ are transitions, and $\Gamma \subseteq Q \times \Sigma \times Q$ is the transition-based *acceptance condition*.

A *run* r of $\mathcal{A}$ on $w \in \Sigma^\omega$ is an ω-word $r_0, w_0, r_1, w_1, \ldots$ in $(Q \times \Sigma)^\omega$ such that $r_0 = q_0$ and, for $i > 0$, it is $(r_{i-1}, w_{i-1}, r_i) \in \Delta$. We write $\inf(r)$ for the set of transitions that appear infinitely often in the run r. A run r of $\mathcal{A}$ is *accepting* if $\inf(r) \cap \Gamma \neq \emptyset$.

The *language*, $L_\mathcal{A}$, of $\mathcal{A}$ (or, *recognized* by $\mathcal{A}$) is the subset of words in Σ^ω that have accepting runs in $\mathcal{A}$. A language is *ω-regular* if it is accepted by a Büchi automaton. An automaton $\mathcal{A} = \langle \Sigma, Q, Q_0, \Delta, \Gamma \rangle$ is *deterministic* if $(q, \sigma, q'), (q, \sigma, q'') \in \Delta$ implies $q' = q''$. $\mathcal{A}$ is *complete* if, for all $\sigma \in \Sigma$ and all $q \in Q$, there is a transition $(q, \sigma, q') \in \Delta$. A word in Σ^ω has exactly one run in a deterministic, complete automaton.

A *Markov decision process (MDP)* $\mathcal{M}$ is a tuple (S, A, T, Σ, L) where S is a finite set of states, A is a finite set of *actions*, $T : S \times A \to \mathcal{D}(S)$, where $\mathcal{D}(S)$ is the set of probability distributions over S, is the *probabilistic transition function*, Σ is an alphabet, and $L : S \times A \times S \to \Sigma$ is the *labelling function* of the set of transitions. For a state $s \in S$, $A(s)$ denotes the set of actions available in s. For states $s, s' \in S$ and $a \in A(s)$, we have that $T(s, a)(s')$ equals $\Pr(s'|s, a)$.

A *run* of $\mathcal{M}$ is an ω-word $s_0, a_1, \ldots \in S \times (A \times S)^\omega$ such that $\Pr(s_{i+1}|s_i, a_{i+1}) > 0$ for all $i \geq 0$. A finite run is a finite such sequence. For a *run* $r = s_0, a_1, s_1, \ldots$

we define the corresponding labelled run as $L(r) = L(s_0, a_1, s_1), L(s_1, a_2, s_2), \ldots \in \Sigma^\omega$. We write $\Omega(\mathcal{M})$ ($\mathrm{Paths}(\mathcal{M})$) for the set of runs (finite runs) of $\mathcal{M}$ and $\Omega_s(\mathcal{M})$ ($\mathrm{Paths}_s(\mathcal{M})$) for the set of runs (finite runs) of $\mathcal{M}$ starting from state s. When the MDP is clear from the context we drop the argument $\mathcal{M}$.

A strategy in $\mathcal{M}$ is a function $\mu : \mathrm{Paths} \to \mathcal{D}(A)$ that for all finite runs r we have $\mathrm{supp}(\mu(r)) \subseteq A(\mathrm{last}(r))$, where $\mathrm{supp}(d)$ is the support of d and $\mathrm{last}(r)$ is the last state of r. Let $\Omega_s^\mu(\mathcal{M})$ denote the subset of runs $\Omega_s(\mathcal{M})$ that correspond to strategy μ and initial state s. Let $\Sigma_\mathcal{M}$ be the set of all strategies. We say that a strategy μ is *pure* if $\mu(r)$ is a point distribution for all runs $r \in \mathrm{Paths}$ and we say that μ is *positional* if $\mathrm{last}(r) = \mathrm{last}(r')$ implies $\mu(r) = \mu(r')$ for all runs $r, r' \in \mathrm{Paths}$.

The behaviour of an MDP $\mathcal{M}$ under a strategy μ with starting state s is defined on a probability space $(\Omega_s^\mu, \mathcal{F}_s^\mu, \mathrm{Pr}_s^\mu)$ over the set of infinite runs of μ from s. Given a random variable over the set of infinite runs $f : \Omega \to \mathbb{R}$, we write $\mathbb{E}_s^\mu\{f\}$ for the expectation of f over the runs of $\mathcal{M}$ from state s that follow strategy μ.

Given an MDP $\mathcal{M}$ and an automaton $\mathcal{A} = \langle \Sigma, Q, q_0, \Delta, \Gamma \rangle$, we want to compute an optimal strategy satisfying the objective that the run of $\mathcal{M}$ is in the language of $\mathcal{A}$. We define the semantic satisfaction probability for $\mathcal{A}$ and a strategy μ from state s as:

$$\mathsf{PSem}_\mathcal{A}^\mathcal{M}(s, \mu) = \mathrm{Pr}_s^\mu\{r \in \Omega_s^\mu : L(r) \in L_\mathcal{A}\} \text{ and } \quad \mathsf{PSem}_\mathcal{A}^\mathcal{M}(s) = \sup_\mu \big(\mathsf{PSem}_\mathcal{A}^\mathcal{M}(s, \mu) \big) .$$

When using automata for the analysis of MDPs, we need a syntactic variant of the acceptance condition. Given an MDP $\mathcal{M} = (S, A, T, \Sigma, L)$ with initial state $s_0 \in S$ and an automaton $\mathcal{A} = \langle \Sigma, Q, q_0, \Delta, \Gamma \rangle$, the *product* $\mathcal{M} \times \mathcal{A} = (S \times Q, (s_0, q_0), A \times Q, T^\times, \Gamma^\times)$ is an MDP augmented with an initial state (s_0, q_0) and accepting transitions $\Gamma^\times$. The function $T^\times : (S \times Q) \times (A \times Q) \rightharpoondown \mathcal{D}(S \times Q)$ is defined by

$$T^\times((s, q), (a, q'))((s', q')) = \begin{cases} T(s, a)(s') & \text{if } (q, L(s, a, s'), q') \in \Delta \\ 0 & \text{otherwise.} \end{cases}$$

Finally, $\Gamma^\times \subseteq (S \times Q) \times (A \times Q) \times (S \times Q)$ is defined by $((s, q), (a, q'), (s', q')) \in \Gamma^\times$ if, and only if, $(q, L(s, a, s'), q') \in \Gamma$ and $T(s, a)(s') > 0$. A strategy μ on the MDP defines a strategy $\mu^\times$ on the product, and vice versa. We define the syntactic satisfaction probabilities as

$$\mathsf{PSyn}_\mathcal{A}^\mathcal{M}((s, q), \mu^\times) = \mathrm{Pr}_s^\mu\{r \in \Omega_{(s,q)}^{\mu^\times}(\mathcal{M} \times \mathcal{A}) : \inf(r) \cap \Gamma^\times \neq \emptyset\} \ , \quad \text{and}$$
$$\mathsf{PSyn}_\mathcal{A}^\mathcal{M}(s) = \sup_{\mu^\times} \big(\mathsf{PSyn}_\mathcal{A}^\mathcal{M}((s, q_0), \mu^\times) \big) \ .$$

Note that $\mathsf{PSyn}_\mathcal{A}^\mathcal{M}(s) = \mathsf{PSem}_\mathcal{A}^\mathcal{M}(s)$ holds for a deterministic $\mathcal{A}$. In general, $\mathsf{PSyn}_\mathcal{A}^\mathcal{M}(s) \leq \mathsf{PSem}_\mathcal{A}^\mathcal{M}(s)$ holds, but equality is not guaranteed because the optimal resolution of nondeterministic choices may require access to future events.

Definition 1 (GFM automata [3]). *An automaton $\mathcal{A}$ is* good for MDPs *if, for all MDPs $\mathcal{M}$, $\mathsf{PSyn}_\mathcal{A}^\mathcal{M}(s_0) = \mathsf{PSem}_\mathcal{A}^\mathcal{M}(s_0)$ holds, where s_0 is the initial state of $\mathcal{M}$.*

For an automaton to match $\mathsf{PSem}_\mathcal{A}^\mathcal{M}(s_0)$, its nondeterminism is restricted not to rely heavily on the future; rather, it must be possible to resolve the nondeterminism on-the-fly.

2 Undiscounted Reward Shaping

We build on the reduction from [2,3] that reduces maximising the chance to realise an ω-regular objective given by a good-for-MDPs Büchi automaton $\mathcal{A}$ for an MDP $\mathcal{M}$ to maximising the chance to meet the reachability objective in the augmented MDP $\mathcal{M}^\zeta$ (for $\zeta \in]0,1[$) obtained from $\mathcal{M} \times \mathcal{A}$ by

- adding a new target state t (either as a sink with a self-loop or as a point where the computation stops; we choose here the latter view) and
- by making the target t a destination of each accepting transition τ of $\mathcal{M} \times \mathcal{A}$ with probability $1-\zeta$ and
  multiplying the original probabilities of all other destinations of an accepting transition τ by ζ.

Let

$$\mathsf{PSyn}_t^{\mathcal{M}^\zeta}((s,q),\mu) = \Pr{}_s^\mu\{r \in \Omega^\mu_{(s,q)}(\mathcal{M}^\zeta) : r \text{ reaches } t\} \ , \quad \text{and}$$
$$\mathsf{PSyn}_t^{\mathcal{M}^\zeta}(s) = \sup_\mu \big(\mathsf{PSyn}_t^{\mathcal{M}^\zeta}((s,q_0),\mu)\big) \ .$$

Theorem 1 ([2,3]). *The following holds:*

1. *$\mathcal{M}^\zeta$ (for $\zeta \in]0,1[$) and $\mathcal{M} \times \mathcal{A}$ have the same set of strategies.*
2. *For a strategy μ, the chance of reaching the target t in $\mathcal{M}^\zeta_\mu$ is 1 if, and only if, the chance of satisfying the Büchi objective in $(\mathcal{M} \times \mathcal{A})_\mu$ is 1:*
   $\mathsf{PSyn}_t^{\mathcal{M}^\zeta}((s_0,q_0),\mu) = 1 \ \Leftrightarrow\ \mathsf{PSyn}_{\mathcal{A}}^{\mathcal{M}}(s_0,q_0),\mu) = 1$
3. *There is a $\zeta_0 \in]0,1[$ such that, for all $\zeta \in [\zeta_0,1[$, an optimal reachability strategy μ for $\mathcal{M}^\zeta$ is an optimal strategy for satisfying the Büchi objective in $\mathcal{M} \times \mathcal{A}$:*
   $\mathsf{PSyn}_t^{\mathcal{M}^\zeta}((s_0,q_0),\mu) = \mathsf{PSyn}_t^{\mathcal{M}^\zeta}(s_0) \ \Rightarrow\ \mathsf{PSyn}_{\mathcal{A}}^{\mathcal{M}}(s_0,q_0),\mu) = \mathsf{PSyn}_{\mathcal{A}}^{\mathcal{M}}(s_0))$.

This allows for analysing the much simpler reachability objective in $\mathcal{M}^\zeta_\mu$ instead of the Büchi objective in $\mathcal{M} \times \mathcal{A}$, and is open to implementation in model free reinforcement learning.

However, it has the drawback that rewards occur late when ζ is close to 1. We amend that by the following observation:

We build, for a good-for-MDPs Büchi automaton $\mathcal{A}$ and an MDP $\mathcal{M}$, the augmented MDP $\overline{\mathcal{M}}^\zeta$ (for $\zeta \in]0,1[$) obtained from $\mathcal{M} \times \mathcal{A}$ in the same way as $\mathcal{M}^\zeta$, i.e. by

- adding a new sink state t (as a sink where the computation stops) and
- by making the sink t a destination of each accepting transition τ of $\mathcal{M} \times \mathcal{A}$ with probability $1-\zeta$ and
  multiplying the original probabilities of all other destinations of an accepting transition τ by ζ.

Different to $\mathcal{M}^\zeta$, $\overline{\mathcal{M}}^\zeta$ has an undiscounted reward objective, where taking an accepting (in $\mathcal{M} \times \mathcal{A}$) transition τ provides a reward of 1, regardless of whether it leads to the sink t or stays in the state-space of $\mathcal{M} \times \mathcal{A}$.

Let, for a run r of $\mathcal{M}^\zeta$ that contains $n \in \mathbb{N}_0 \cup \{\infty\}$ accepting transitions, the total reward be $\mathsf{Total}(r) = n$, and let

$$\mathsf{ETotal}^{\overline{\mathcal{M}}^\zeta}((s,q),\mu) = \mathbb{E}_s^\mu\{\mathsf{Total}(r) : r \in \Omega^\mu_{(s,q)}(\overline{\mathcal{M}}^\zeta)\} \ , \quad \text{and}$$
$$\mathsf{ETotal}^{\overline{\mathcal{M}}^\zeta}(s) = \sup_\mu \left(\mathsf{ETotal}^{\overline{\mathcal{M}}^\zeta}((s,q_0),\mu)\right) \ .$$

Note that the set of runs with $\mathsf{Total}(r) = \infty$ has probability 0 in $\Omega^\mu_{(s,q)}(\overline{\mathcal{M}}^\zeta)$: they are the runs that infinitely often do not move to t on an accepting transition, where the chance that this happens at least n times is $(1-\zeta)^n$ for all $n \in \mathbb{N}_0$.

Theorem 2. *The following holds:*

1. $\overline{\mathcal{M}}^\zeta$ *(for* $\zeta \in]0,1[$*),* $\mathcal{M}^\zeta$ *(for* $\zeta \in]0,1[$*), and* $\mathcal{M} \times \mathcal{A}$ *have the same set of strategies.*
2. *For a strategy* μ*, the expected reward for* $\overline{\mathcal{M}}^\zeta_\mu$ *is* r *if, and only if, the chance of reaching the target* t *in* $\mathcal{M}^\zeta_\mu$ *is* $\frac{r}{1-\zeta}$*:*
   $\mathsf{PSyn}_t^{\mathcal{M}^\zeta}((s_0,q_0),\mu) = (1-\zeta)\mathsf{ETotal}^{\overline{\mathcal{M}}^\zeta}((s_0,q_0),\mu)$.
3. *The expected reward for* $\overline{\mathcal{M}}^\zeta_\mu$ *is in* $[0, \frac{1}{1-\zeta}]$.
4. *The chance of satisfying the Büchi objective in* $(\mathcal{M} \times \mathcal{A})_\mu$ *is* 1 *if, and only if, the expected reward for* $\overline{\mathcal{M}}^\zeta_\mu$ *is* $\frac{1}{1-\zeta}$.
5. *There is a* $\zeta_0 \in]0,1[$ *such that, for all* $\zeta \in [\zeta_0, 1[$*, a strategy* μ *that maximises the reward for* $\overline{\mathcal{M}}^\zeta$ *is an optimal strategy for satisfying the Büchi objective in* $\mathcal{M} \times \mathcal{A}$.

Proof. (1) Obvious, because all the states and their actions are the same apart from the sink state t for which the strategy can be left undefined.

(2) The sink state t can only be visited once along any run, so the expected number of times a run starting at (s_0, q_0) is going to visit t while using strategy μ is the same as its probability of visiting t, i.e., $\mathsf{PSyn}_t^{\mathcal{M}^\zeta}((s_0,q_0),\mu)$. The only way t can be reached is by traversing an accepting transition and this always happens with the same probability $(1-\zeta)$. So the expected number of visits to t is the expected number of times an accepting transition is used, i.e., $\mathsf{ETotal}^{\overline{\mathcal{M}}^\zeta}((s_0,q_0),\mu)$, multiplied by $(1-\zeta)$.

(3) follows from (2), because $\mathsf{PSyn}_t^{\mathcal{M}^\zeta}((s_0,q_0),\mu)$ cannot be greater than 1.

(4) follows from (2) and Theorem 1 (2).

(5) follows from (2) and Theorem 1 (3).

3 Discounted Reward Shaping

The expected undiscounted reward for $\overline{\mathcal{M}}^\zeta_\mu$ can be viewed as a discounted reward for $(\mathcal{M} \times \mathcal{A})_\mu$, by giving a reward ζ^i to when passing through an accepting transition when i *accepting* transitions have been passed before. We call this reward ζ-*biased*.

Let, for a run r of $\mathcal{M} \times \mathcal{A}$ that contains $n \in \mathbb{N}_0 \cup \{\infty\}$ accepting transitions, the ζ-biased discounted reward be $\mathsf{Disct}_\zeta(r) = \sum_{i=0}^{n-1} \zeta^i$, and let

$$\mathsf{EDisct}_\zeta^{\mathcal{M}\times\mathcal{A}}((s,q),\mu) = \mathbb{E}_s^\mu\{r \in \Omega^\mu_{(s,q)}(\mathcal{M}\times\mathcal{A}) : \mathsf{Disct}_\zeta(r)\} \quad , \qquad \text{and}$$

$$\mathsf{EDisct}_\zeta^{\mathcal{M}\times\mathcal{A}}(s) = \sup_\mu \left(\mathsf{EDisct}_\zeta^{\mathcal{M}\times\mathcal{A}}((s,q_0),\mu)\right) \ .$$

Theorem 3. *For every strategy μ, the expected reward for $\overline{\mathcal{M}}_\mu^\zeta$ is equal to the expected ζ-biased reward for $(\mathcal{M}\times\mathcal{A})_\mu$:* $\mathsf{EDisct}_\zeta^{\mathcal{M}\times\mathcal{A}}((s,q),\mu) = \mathsf{ETotal}^{\overline{\mathcal{M}}^\zeta}((s,q),\mu)$.

This is simply because the discounted reward for each transition is equal to the chance of not having reached t before (and thus still seeing this transition) in $\overline{\mathcal{M}}_\mu^\zeta$.

This improves over [1] because it only uses one discount parameter, ζ, instead of two (called γ and γ_B in [1]) parameters (that are not independent). It is also simpler and more intuitive: discount whenever you have earned a reward.

References

1. Alper Kamil Bozkurt, Yu Wang, Michael M. Zavlanos, and Miroslav Pajic. Control synthesis from linear temporal logic specifications using model-free reinforcement learning. *CoRR*, abs/1909.07299, 2019.
2. E. M. Hahn, M. Perez, S. Schewe, F. Somenzi, A. Trivedi, and D. Wojtczak. Omega-regular objectives in model-free reinforcement learning. In *Tools and Algorithms for the Construction and Analysis of Systems*, pages 395–412, 2019. LNCS 11427.
3. E. M. Hahn, M. Perez, S. Schewe, F. Somenzi, A. Trivedi, and D. Wojtczak. Good-for-mdps automata for probabilistic analysis and reinforcement learning. In *Tools and Algorithms for the Construction and Analysis of Systems*, page to appear, 2020.