Learning-Based Mean-Payoff Optimization in an Unknown MDP under Omega-Regular Constraints

Jan Křetínský, Guillermo A. Pérez, and Jean-François Raskin

Technische Universität München
Université libre de Bruxelles

MF&V Seminar
May, 2018

Parity games

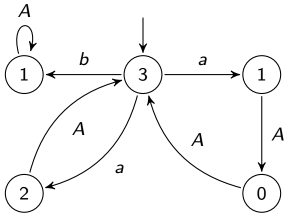
Playing on an automaton

A strategy σ in a parity automaton (Q, A, T, p) is a function $(Q \cdot A)^* Q \to \mathcal{D}(A)$. It is winning from q_0 if the min priority seen infinitely often is even, along all runs $q_0 a_0 \cdots \in (Q \cdot A)^\omega$ consistent with it.

Parity games

Playing on an automaton

A strategy σ in a parity automaton (Q, A, T, p) is a function $(Q \cdot A)^* Q \to \mathcal{D}(A)$. It is winning from q_0 if the min priority seen infinitely often is even, along all runs $q_0 a_0 \cdots \in (Q \cdot A)^\omega$ consistent with it.

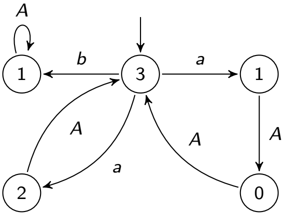


Strategies: (i) play b, (ii) always play a.

Parity games

Playing on an automaton

A strategy σ in a parity automaton (Q, A, T, p) is a function $(Q \cdot A)^* Q \to \mathcal{D}(A)$. It is winning from q_0 if the min priority seen infinitely often is even, along all runs $q_0 a_0 \cdots \in (Q \cdot A)^\omega$ consistent with it.



Strategies: (i) play b, (ii) always play a.

Expected mean-payoff optimization in MDPs

Reward MDPs

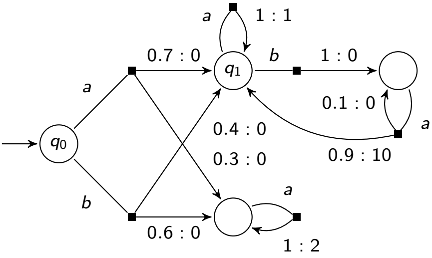
For a strategy σ in an MDP $\mathcal{M} = (Q, A, \alpha, \delta, r)$, we denote $\mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}]$ the expected (lim inf) mean payoff[1] from q_0 under σ.

[1] $\mathbf{MP}(x_0 x_1 \dots) := \liminf_{n \in \mathbb{N}} \frac{1}{n+1} \sum_{i=0}^{n} x_i$

Expected mean-payoff optimization in MDPs

Reward MDPs

For a strategy σ in an MDP $\mathcal{M} = (Q, A, \alpha, \delta, r)$, we denote $\mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}]$ the expected (lim inf) mean payoff[1] from q_0 under σ.

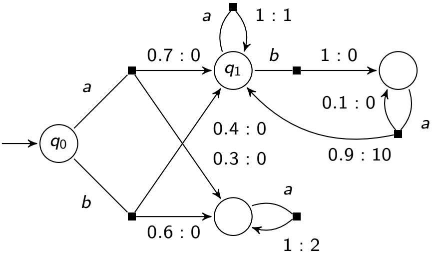


[1] $\mathbf{MP}(x_0 x_1 \ldots) := \liminf_{n\in\mathbb{N}} \frac{1}{n+1} \sum_{i=0}^{n} x_i$

Expected mean-payoff optimization in MDPs

Reward MDPs

For a strategy σ in an MDP $\mathcal{M} = (Q, A, \alpha, \delta, r)$, we denote $\mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}]$ the expected (lim inf) mean payoff[1] from q_0 under σ.



$\mathbb{E}^{q_1}_{\mathcal{M}^\sigma}[\mathbf{MP}] \geq 4.5$ for $\sigma : q_1 \mapsto (b \mapsto 1)$

[1] $\mathbf{MP}(x_0 x_1 \ldots) := \liminf_{n \in \mathbb{N}} \frac{1}{n+1} \sum_{i=0}^{n} x_i$

State of the art

Parity games

It is known that

- they are uniformly, deterministically, and memoryless determined
- they can be decided in $\mathrm{UP} \cap \mathrm{coUP}$ [Jurdziński 98]
- and in QP [Calude et al. 17]

State of the art

Parity games

It is known that

- they are uniformly, deterministically, and memoryless determined
- they can be decided in $\mathrm{UP} \cap \mathrm{coUP}$ [Jurdziński 98]
- and in QP [Calude et al. 17]

Mean-payoff MDPs

It is known that

- memoryless deterministic strategies suffice [Gimbert 07]
- uniformly optimal (mem-less and det.) unichain strategies
- the above can be computed in polynomial time [Puterman 05]

Mean-payoff parity MDPs

A.s. parity satisfaction and mean-payoff optimality

The existence of strategies σ s.t.

$$\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\textsc{Parity}] = 1 \text{ and } \mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}] \geq \nu$$

for given $\mathcal{M}$, q_0, and ν, has been studied before [CD11].

Mean-payoff parity MDPs

A.s. parity satisfaction and mean-payoff optimality

The existence of strategies σ s.t.

$$\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\textsc{Parity}] = 1 \text{ and } \mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}] \geq \nu$$

for given $\mathcal{M}$, q_0, and ν, has been studied before [CD11]. It is known that

- Infinite memory strategies are necessary in general.
- It can be decided in polynomial time.

Mean-payoff parity MDPs

A.s. parity satisfaction and mean-payoff optimality

The existence of strategies σ s.t.

$$\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\textsc{Parity}] = 1 \text{ and } \mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\textbf{MP}] \geq \nu$$

for given $\mathcal{M}$, q_0, and ν, has been studied before [CD11]. It is known that

- Infinite memory strategies are necessary in general.
- It can be decided in polynomial time.

A.s. MP optimality under sure parity constraints

The existence of strategies σ s.t.

$$\varrho \models \textsc{Parity} \text{ for all } \varrho \text{ consistent with } \sigma, \text{ and } \varepsilon\text{-``optimal'' } \mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\textbf{MP}]$$

has also been studied [AKV16].

Mean-payoff parity MDPs

A.s. parity satisfaction and mean-payoff optimality

The existence of strategies σ s.t.

$$\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\textsc{Parity}] = 1 \text{ and } \mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}] \geq \nu$$

for given $\mathcal{M}$, q_0, and ν, has been studied before [CD11]. It is known that

- Infinite memory strategies are necessary in general.
- It can be decided in polynomial time.

A.s. MP optimality under sure parity constraints

The existence of strategies σ s.t.

$$\varrho \models \textsc{Parity} \text{ for all } \varrho \text{ consistent with } \sigma, \text{ and } \varepsilon\text{-“optimal” } \mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}]$$

has also been studied [AKV16].

- Infinite memory strategies are again necessary.
- It is in NP $\cap$ coNP and parity-game hard.

Partially-specified MDPs

Verification problems for partially-spec'd models

There has been an increased interest in models with unknown parameters and the use of learning techniques.[2]

[2] and in using AI techniques to deal with them

Partially-specified MDPs

Verification problems for partially-spec'd models

There has been an increased interest in models with unknown parameters and the use of learning techniques.[2]

- Safe reinforcement learning via shielding [Alshiekh et al. 17]
- Verification of MDPs using learning algorithms [Brázdil et al. 14]
- Safety-constrained reinforcement learning for MDPs [Junges et al. 16]
- Correct-by-synthesis reinforcement learning with temporal logic constraints [WET15]
- Probably approximately correct learning in stochastic games with temporal logic specifications [WT16]

[2] and in using AI techniques to deal with them

MP optimization in unknown ergodic MDPs

End component: $\exists\sigma,\ \mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[q_0 \rightsquigarrow q_1] = 1$ for all q_0, q_1

MP optimization in unknown ergodic MDPs

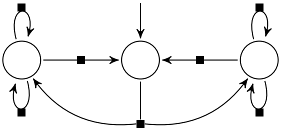
End component: $\exists\sigma$, $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[q_0 \rightsquigarrow q_1] = 1$ for all q_0, q_1

- So we have a uniform random exploration strategy λ that will almost-surely visit all states in the EC infinitely often.

MP optimization in unknown ergodic MDPs

End component: $\exists\sigma$, $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[q_0 \rightsquigarrow q_1] = 1$ for all q_0, q_1

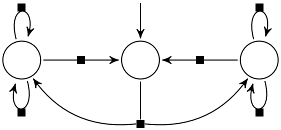
- So we have a uniform random exploration strategy λ that will almost-surely visit all states in the EC infinitely often.



MP optimization in unknown ergodic MDPs

End component: $\exists\sigma, \mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[q_0 \rightsquigarrow q_1] = 1$ for all q_0, q_1

- So we have a uniform random exploration strategy λ that will almost-surely visit all states in the EC infinitely often.



- General recipe: during episode i we first explore for L_i steps then exploit for O_i steps.

Almost-sure optimality (under weak assumptions)

Input

We are given

- an automaton $\mathcal{A}$ whose transition relation T is the exact support of the MDP's unknown probabilistic transition function
- and a transition-probability lower bound $\pi_{\min}$.

Almost-sure optimality (under weak assumptions)

Input

We are given

- an automaton $\mathcal{A}$ whose transition relation T is the exact support of the MDP's unknown probabilistic transition function
- and a transition-probability lower bound $\pi_{\min}$.

Assumptions

We suppose that

- the MDP is ergodic, i.e. it is an EC,
- and that the unknown reward function r instantaneously assigns transitions rewards from $[0, 1]$.

Almost-sure optimality (under weak assumptions)

Useful facts

- Since the MDP is ergodic, all hitting probabilities can be under-approx'd as biased coins X with success probability $\mu := (\pi_{\min}/|A|)^{|Q|}$, and we can use Hoeffding's inequality:

$$\mathbb{P}\left[\left|\frac{1}{k}\sum_{j=1}^{k} X_j - \mu\right| \geq \varepsilon\right] \leq 2\exp(-2k\varepsilon^2).$$

Almost-sure optimality (under weak assumptions)

Useful facts

- Since the MDP is ergodic, all hitting probabilities can be under-approx'd as biased coins X with success probability $\mu := (\pi_{\min}/|A|)^{|Q|}$, and we can use Hoeffding's inequality:

$$\mathbb{P}\left[\left|\frac{1}{k}\sum_{j=1}^{k} X_j - \mu\right| \geq \varepsilon\right] \leq 2\exp(-2k\varepsilon^2).$$

- Expectation-optimal strategies τ for any MDP $\mathcal{N}$ with the same support as $\mathcal{M}$ and s.t. $|\delta_{\mathcal{N}}(q,a,q') - \delta_{\mathcal{M}}(q,a,q')| \leq \frac{\pi_{\min}}{2}\left(\left(1+\frac{\varepsilon}{2}\right)^{\frac{1}{2|Q|}} - 1\right)$ give us

$$\left|\mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}] - \sup_{\sigma}\mathbb{E}^{q_0}_{\mathcal{M}^\sigma}[\mathbf{MP}]\right| \leq \varepsilon$$

  for all q_0 [Solan 03; Chatterjee 12].

Almost-sure optimality (under weak assumptions)

One last useful fact:

Almost-sure optimality (under weak assumptions)

One last useful fact:

Lemma (Convergence & ergodicitiy)

For all ergodic MDPs $\mathcal{M}$, for all q_0, for all unichain deterministic memoryless strategies μ, we have

- $\mathbb{P}^{q_0}_{\mathcal{M}^\mu}[\varrho : \mathsf{MP}(\varrho) \geq \mathbb{E}^{q_0}_{\mathcal{M}^\mu}[\mathbf{MP}]] = 1$; *and*
- *for all* $\varepsilon \in (0, 1)$, *one can compute* $M(\varepsilon) \in \mathbb{N}$ *s.t.* $\mathbb{P}^{q_0}_{\mathcal{M}^\mu}[\varrho : \forall k \geq M(\varepsilon), \mathbf{FinAvg}(\varrho(..k)) \geq \mathbb{E}^{q_0}_{\mathcal{M}^\mu}[\mathbf{MP}] - \varepsilon] \geq 1 - \varepsilon$.

In words

- Almost all runs have as their mean-payoff value the expected mean-payoff of the strategy.
- The finite averages eventually stay ever closer to the expected mean-payoff of the strategy with ever higher probability.

Almost-sure optimality (under weak assumptions)

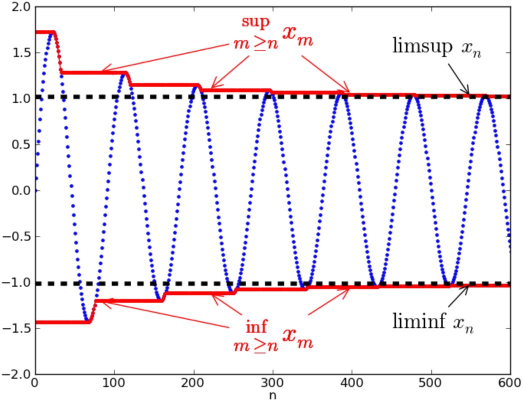
Theorem

One can compute a sequence $(L_i, O_i)_{i \in \mathbb{N}}$ *s.t. for the resulting strategy* σ_∞*, we have* $\mathbb{P}^{q_0}_{\mathcal{M}^{\sigma_\infty}}[\varrho : \mathsf{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}]] = 1$ *for all* q_0*.*

Almost-sure optimality (under weak assumptions)

Theorem

One can compute a sequence $(L_i, O_i)_{i \in \mathbb{N}}$ *s.t. for the resulting strategy* σ_∞*, we have* $\mathbb{P}^{q_0}_{\mathcal{M}^{\sigma_\infty}}[\varrho : \mathrm{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}]] = 1$ *for all* q_0*.*



Almost-sure optimality (under weak assumptions)

Theorem

One can compute a sequence $(L_i, O_i)_{i \in \mathbb{N}}$ *s.t. for the resulting strategy* σ_∞*, we have* $\mathbb{P}^{q_0}_{\mathcal{M}^{\sigma_\infty}}\left[\varrho : \mathrm{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}]\right] = 1$ *for all* q_0*.*

A more detailed recipe

Take any $(\varepsilon_i)_{i \in \mathbb{N}}$ s.t. $0 < \varepsilon_k < \varepsilon_j < 1$ for all $j < k$. We define σ_∞ as operating in episodes $i \in \mathbb{N}$:

1. It first explores during L_i steps so that with high probability we will be able to compute δ_i and r_i s.t. expectation-opt. strategies for $\mathcal{A}_{\delta_i, r_i}$ are ε_i-opt. for $\mathcal{M}$.
2. Then, σ_∞ follows an expectation-opt. strategy $\sigma^{\delta_i}_{\mathrm{MP}}$ for $\mathcal{A}_{\delta_i, r_i}$ during O_i steps that account for
   - previous average drops,
   - convergence speed of the finite averages, and
   - future average drops during the next L_{i+1} exploration steps,

   with high probability.

Limit-sure optimality under sure-parity constraints

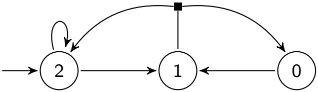
Let's add a parity constraint

- INPUT: a parity automaton and $\pi_{\min}$
- ASSUMPTIONS: the MDP is ergodic, its minimal priority is even, and all states are parity-winning, i.e. there is a winning strategy from them,
- SYNTH: a parity-winning strategy that achieves an optimal mean payoff with high probability

Limit-sure optimality under sure-parity constraints

Let's add a parity constraint

- INPUT: a parity automaton and $\pi_{\min}$
- ASSUMPTIONS: the MDP is ergodic, its minimal priority is even, and all states are parity-winning, i.e. there is a winning strategy from them,
- SYNTH: a parity-winning strategy that achieves an optimal mean payoff with high probability

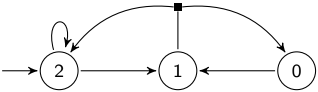


Why not probability 1?

Limit-sure optimality under sure-parity constraints

Let's add a parity constraint

- INPUT: a parity automaton and $\pi_{\min}$
- ASSUMPTIONS: the MDP is ergodic, its minimal priority is even, and all states are parity-winning, i.e. there is a winning strategy from them,
- SYNTH: a parity-winning strategy that achieves an optimal mean payoff with high probability



Why not probability 1? Parity-bad exploration must be made finite, lest we violate the parity constraint.

Limit-sure optimality under sure-parity constraints

Theorem

For all $\gamma \in (0,1)$ *there exists a strategy* σ *s.t. for all* q_0

1. $\varrho \models \textsc{Parity}$ *for all* ϱ *consistent with* σ *and*
2. $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}\left[\varrho : \mathsf{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}\left[\mathbf{MP}\right]\right] = 1 - \gamma.$

Limit-sure optimality under sure-parity constraints

Theorem

For all $\gamma \in (0,1)$ *there exists a strategy* σ *s.t. for all* q_0

1. $\varrho \models \textsc{Parity}$ *for all* ϱ *consistent with* σ *and*
2. $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}\left[\varrho : \mathsf{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}]\right] = 1 - \gamma.$

We modify σ_∞ go "give up":

We give up and switch to a parity-winning strategy forever if for sufficiently many episodes we have not witnessed the minimal even priority.

Limit-sure optimality under sure-parity constraints

Theorem

For all $\gamma \in (0,1)$ there exists a strategy σ s.t. for all q_0

1. $\varrho \models \textsc{Parity}$ *for all ϱ consistent with σ and*
2. $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}\left[\varrho : \mathsf{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}\left[\mathbf{MP}\right]\right] = 1 - \gamma.$

We modify σ_∞ go "give up":

We give up and switch to a parity-winning strategy forever if for sufficiently many episodes we have not witnessed the minimal even priority.

- σ_∞ explores (for $\geq |Q|$ steps) infinitely often

Limit-sure optimality under sure-parity constraints

Theorem

For all $\gamma \in (0,1)$ there exists a strategy σ s.t. for all q_0

1. $\varrho \models \textsc{Parity}$ *for all ϱ consistent with σ and*
2. $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}\left[\varrho : \mathsf{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}\left[\mathbf{MP}\right]\right] = 1 - \gamma.$

We modify σ_∞ go "give up":

We give up and switch to a parity-winning strategy forever if for sufficiently many episodes we have not witnessed the minimal even priority.

- σ_∞ explores (for $\geq |Q|$ steps) infinitely often
- every $|Q|$ steps of exploration we visit every state with probability at least $(\pi_{\min}/|A|)^{|Q|}$

Limit-sure optimality under sure-parity constraints

Theorem

For all $\gamma \in (0,1)$ there exists a strategy σ s.t. for all q_0

1. $\varrho \models \textsc{Parity}$ *for all ϱ consistent with σ and*
2. $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}\left[\varrho : \mathsf{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}\left[\mathbf{MP}\right]\right] = 1 - \gamma.$

We modify σ_∞ go "give up":

We give up and switch to a parity-winning strategy forever if for sufficiently many episodes we have not witnessed the minimal even priority.

- σ_∞ explores (for $\geq |Q|$ steps) infinitely often
- every $|Q|$ steps of exploration we visit every state with probability at least $(\pi_{\min}/|A|)^{|Q|}$
- $\lim_{i \in \mathbb{N}} \prod_{j=i}^{\infty} (1 - 2^{-j}) = 1$

Limit-sure optimality under sure-parity constraints

Theorem

For all $\gamma \in (0,1)$ *there exists a strategy* σ *s.t. for all* q_0

1. $\varrho \models \textsc{Parity}$ *for all* ϱ *consistent with* σ *and*
2. $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}\left[\varrho : \mathsf{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}]\right] = 1 - \gamma.$

We modify σ_∞ go "give up":

We give up and switch to a parity-winning strategy forever if for sufficiently many episodes we have not witnessed the minimal even priority.

- σ_∞ explores (for $\geq |Q|$ steps) infinitely often
- every $|Q|$ steps of exploration we visit every state with probability at least $(\pi_{\min}/|A|)^{|Q|}$
- $\lim_{i \in \mathbb{N}} \prod_{j=i}^{\infty}(1 - 2^{-j}) = 1 \implies \exists K_0, \prod_{j=K_0}^{\infty}(1 - 2^{-j}) \geq (1 - \gamma)$

Limit-sure optimality under sure-parity constraints

Theorem

For all $\gamma \in (0,1)$ there exists a strategy σ s.t. for all q_0

1. $\varrho \models \textsc{Parity}$ *for all ϱ consistent with σ and*
2. $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\varrho : \mathsf{MP}(\varrho) \geq \sup_\tau \mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}]] = 1 - \gamma$.

We modify σ_∞ go "give up":

We give up and switch to a parity-winning strategy forever if for sufficiently many episodes we have not witnessed the minimal even priority.

- σ_∞ explores (for $\geq |Q|$ steps) infinitely often
- every $|Q|$ steps of exploration we visit every state with probability at least $(\pi_{\min}/|A|)^{|Q|}$
- $\lim_{i\in\mathbb{N}} \prod_{j=i}^{\infty}(1-2^{-j}) = 1 \implies \exists K_0, \prod_{j=K_0}^{\infty}(1-2^{-j}) \geq (1-\gamma)$
- So we wait for enough episodes so that the probability of seeing the minimal even priority is at least $1-2^{-K_0}$, then $1-2^{-K_0-1}$, ...

Limit-sure near-optimality under sure-parity constraints

Let's weaken the assumptions

- INPUT: a parity automaton and $\pi_{\min}$
- ASSUMPTIONS: the MDP is ergodic, and all states are parity-winning, i.e. there is a winning strategy from them,
- SYNTH: a parity-winning strategy that achieves a near-optimal mean payoff with high probability

Limit-sure near-optimality under sure-parity constraints

Let's weaken the assumptions

- INPUT: a parity automaton and $\pi_{\min}$
- ASSUMPTIONS: the MDP is ergodic, and all states are parity-winning, i.e. there is a winning strategy from them,
- SYNTH: a parity-winning strategy that achieves a near-optimal mean payoff with high probability

Optimality w.r.t. what?

$$\mathbf{sVal}(\mathcal{M}) := \max_{q_0} \sup\{\mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}] : \tau \text{ is a winning strategy}\}$$

Limit-sure near-optimality under sure-parity constraints

Let's weaken the assumptions

- INPUT: a parity automaton and $\pi_{\min}$
- ASSUMPTIONS: the MDP is ergodic, and all states are parity-winning, i.e. there is a winning strategy from them,
- SYNTH: a parity-winning strategy that achieves a near-optimal mean payoff with high probability

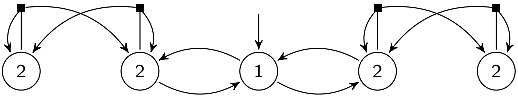
Optimality w.r.t. what?

$$\mathbf{sVal}(\mathcal{M}) := \max_{q_0} \sup \{\mathbb{E}^{q_0}_{\mathcal{M}^\tau}[\mathbf{MP}] : \tau \text{ is a winning strategy}\}$$

- Since all states are parity-winning, the MDP contains at least one EC with even min priority.
- From [AKV16] we know $\mathbf{sVal}(\mathcal{M})$ is equal to

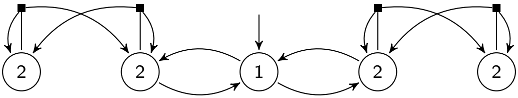
$$\max\left\{\sup_{\tau} \mathbb{E}^{q_0}_{S^\tau}[\mathbf{MP}] \mid S \text{ is an EC with even min priority in } \mathcal{M}\right\}$$

Limit-sure near-optimality under sure-parity constraints



Why near-optimality?

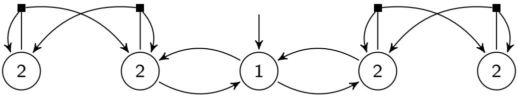
Limit-sure near-optimality under sure-parity constraints



Why near-optimality?

- Global exploration must be made finite since the minimal priority may be odd.

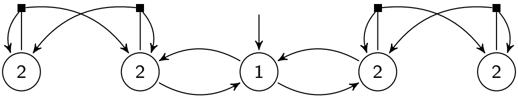
Limit-sure near-optimality under sure-parity constraints



Why near-optimality?

- Global exploration must be made finite since the minimal priority may be odd.
- So, we can only choose ε-maximal ECs with even minimal priority (with high probability). This can be done via an initial exploration phase.

Limit-sure near-optimality under sure-parity constraints



Why near-optimality?

- Global exploration must be made finite since the minimal priority may be odd.
- So, we can only choose ε-maximal ECs with even minimal priority (with high probability). This can be done via an initial exploration phase.

Theorem

For all $\varepsilon, \gamma \in (0,1)$ there exists a strategy σ s.t. for all q_0

1. $\varrho \models \textsc{Parity}$ *for all ϱ consistent with σ and*
2. $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}\left[\varrho : \mathsf{MP}(\varrho) \geq \mathbf{sVal}(\mathcal{M}) - \varepsilon\right] \geq 1 - \gamma.$

General non-ergodic MDPs

Let's weaken the assumptions further

- INPUT: a parity automaton and $\pi_{\min}$
- ASSUMPTIONS: all states are parity-winning, i.e. there is a winning strategy from them,
- SYNTH: a parity-winning strategy that achieves a near-optimal mean payoff with high probability

General non-ergodic MDPs

Theorem

For all $\varepsilon, \gamma \in (0,1)$ *there exists a strategy* σ *s.t. for all* $q_0 \in Q$

- $\varrho \models \textsc{Parity}$ *for all* ϱ *consistent with* σ *and*
- $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\varrho : \mathsf{MP}(\varrho) \geq \mathbf{sVal}(S) - \varepsilon \mid \mathrm{Inf} \subseteq S] \geq 1 - \gamma$ *for all ECs containing an EC* S *with even min priority and s.t.* $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\mathrm{Inf} \subseteq S] > 0$.

General non-ergodic MDPs

Theorem

For all $\varepsilon, \gamma \in (0,1)$ *there exists a strategy* σ *s.t. for all* $q_0 \in Q$

- $\varrho \models \textsc{Parity}$ *for all* ϱ *consistent with* σ *and*
- $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\varrho : \mathsf{MP}(\varrho) \geq \mathbf{sVal}(S) - \varepsilon \mid \mathrm{Inf} \subseteq S] \geq 1 - \gamma$ *for all ECs containing an EC* S *with even min priority and s.t.* $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\mathrm{Inf} \subseteq S] > 0$.

Strategy recipe

- Follow a parity-winning strategy until a new (previously unvisited) EC containing an EC with even min priority is reached.
- Switch to our solution for such ECs.
- If we ever exit it, switch back to a parity-winning strategy and mark the EC as visited.

General non-ergodic MDPs

Theorem

For all $\varepsilon, \gamma \in (0,1)$ *there exists a strategy* σ *s.t. for all* $q_0 \in Q$

- $\varrho \models \textsc{Parity}$ *for all* ϱ *consistent with* σ *and*
- $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\varrho : \mathsf{MP}(\varrho) \geq \mathbf{sVal}(S) - \varepsilon \mid \mathrm{Inf} \subseteq S] \geq 1 - \gamma$ *for all ECs containing an EC* S *with even min priority and s.t.* $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\mathrm{Inf} \subseteq S] > 0$.

Theorem

For all $\varepsilon, \gamma \in (0,1)$ *one can construct a finite-memory strategy* σ *s.t. for all* $q_0 \in Q$

- $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\textsc{Parity}] = 1$ *and*
- $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\varrho : \mathsf{MP}(\varrho) \geq \mathbf{sVal}(S) - \varepsilon \mid \mathrm{Inf} \subseteq S] \geq 1 - \gamma$ *for all ECs containing an EC* S *with even min priority and s.t.* $\mathbb{P}^{q_0}_{\mathcal{M}^\sigma}[\mathrm{Inf} \subseteq S] > 0$.

A single exploration phase followed by alternating phases of exploitation and $|Q|$ random steps.

Fin

Conclusions

Given an unknown MDP, we have shown how to construct

- limit-sure near-optimal strategies that surely satisfy a parity constraint; and
- limit-sure near-optimal finite-memory strategies that almost-surely satisfy the parity constraint.

Fin

Conclusions

Given an unknown MDP, we have shown how to construct

- limit-sure near-optimal strategies that surely satisfy a parity constraint; and
- limit-sure near-optimal finite-memory strategies that almost-surely satisfy the parity constraint.

Future work

- Can we obtain model-free learning strategies that yield the same guarantees?
- How does one implement the finite-memory strategies efficiently? (Memory vs. processing power)
- Can we weaken the assumptions? Support, lower transition-probability bound, or bounded rewards?
- Can we obtain bounds on the sample complexity of these problems?