Reducing Blackwell and Average Optimality to Discounted MDPs

Julien Grand-Clément[1], Marek Petrik[2]

Department of Information Systems and Operations Management, HEC Paris[1], Department of Computer Science, University of New Hampshire[2],

Markov decision processes (MDPs)

An MDP instance is characterized by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{r}, \boldsymbol{P})$:

- $\mathcal{S}$ is a finite set of states and $\mathcal{A}$ is a finite set of actions
- $\boldsymbol{r} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the instantaneous rewards
- $\boldsymbol{P} \in (\Delta(\mathcal{S}))^{\mathcal{S}\times\mathcal{A}}$ is the transition probabilities

The *discounted value function* $\boldsymbol{v}_\gamma^\pi \in \mathbb{R}^{\mathcal{S}}$ of a policy π is

$$v_{\gamma,s}^\pi = \mathbb{E}^{\pi,\boldsymbol{P}}\left[\sum_{t=0}^{+\infty} \gamma^t r_{s_t,a_t} \;\middle|\; s_0 = s\right], \forall\, s \in \mathcal{S}.$$

Discount optimality. Given $\gamma \in [0,1)$, a policy π is γ-discount-optimal if $v_{\gamma,s}^\pi \geq v_{\gamma,s}^{\pi'}, \forall\, \pi', \forall\, s \in \mathcal{S}$.

The average reward $\boldsymbol{g}^\pi \in \mathbb{R}^{\mathcal{S}}$ of a policy π is

$$g_s^\pi = \lim_{T\to+\infty} \frac{1}{T+1}\mathbb{E}^{\pi,\boldsymbol{P}}\left[\sum_{t=0}^{T} r_{s_t,a_t} \;\middle|\; s_0 = s\right], \forall\, s \in \mathcal{S}.$$

Average optimality. A policy π is average optimal if $g_s^\pi \geq g_s^{\pi'}, \forall\, \pi', \forall\, s \in \mathcal{S}$.

Blackwell optimality. A policy π is Blackwell optimal if it remains γ-discount optimal for all γ sufficiently close to 1.

For a Blackwell optimal policy π, we define

$$\gamma(\pi) = \min\{\gamma \in [0,1) \mid \pi \text{ is } \gamma'\text{-discount optimal}, \forall\, \gamma' \in [\gamma, 1)\}. \tag{1}$$

Our main contributions

1. We introduce the **Blackwell discount factor** $\gamma_{\mathsf{bw}} \in [0,1)$.
   Any policy that is γ-discount optimal for $\gamma > \gamma_{\mathsf{bw}}$ is average optimal and Blackwell optimal.
2. We obtain a closed-form **upper-bound on the Blackwell discount factor** $\hat{\gamma} \in (0,1)$.
   $\Rightarrow$ This reduces finding average and Blackwell optimal policies to solving discounted MDPs.
   $\Rightarrow$ This gives a weakly-polynomial time algorithm to compute Blackwell/average optimal policies, *without any structural assumptions*.
3. We extend our results to **robust MDPs** with polyhedral uncertainty.

Limitations of existing approaches.

For computing average optimal policies. Most approaches require structural assumptions: unichain, irreducibility, weakly-communicating MDPs, deterministic transitions, on mixing-times [JS21, WWY22].

For computing Blackwell optimal policies. Two existing algorithms, based on solving linear programs over power (Laurent) series [HDK85] or on solving $|\mathcal{S}|$ nested non-linear equations via linear programs [OVJ17].

Main limitation 1. Existing algorithms for average or Blackwell optimality are significantly more involved than those for discount optimality.

Main limitation 2. There exists a unichain MDP instance $\mathcal{M}$, a Blackwell-optimal policy π, and discount factors $\gamma_1, \gamma_2 \in [0,1)$ with $\gamma_1 < \gamma(\pi) < \gamma_2$ such that:

1. the policy π is γ_1-discount-optimal, and
2. there exists $\pi' \neq \pi$ that is γ_2-discount-optimal and *not* Blackwell-optimal.

A pathological example

In the MDP below:

$$v_\gamma^{a_1} = 1, v_\gamma^{a_2} = r_1\gamma + r_2\gamma^2, v_\gamma^{a_3} = r_4\gamma + r_5\gamma^2$$

$\Rightarrow$ We can choose the parameters r_1, r_2, r_3, r_4 to induce pathological behaviors.

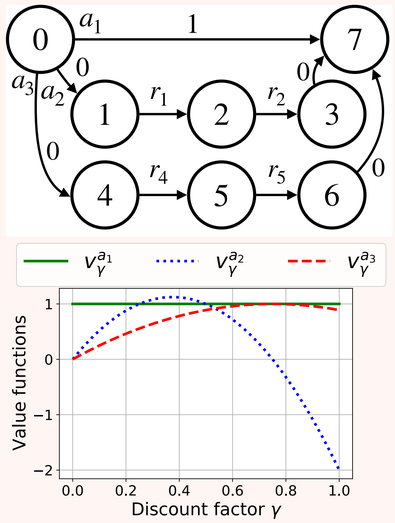

Figure 1. MDP instance (top) and value functions (bottom).

The Blackwell discount factor

The **Blackwell discount factor** $\gamma_{\mathsf{bw}} \in [0,1)$ is the smallest discount factor $\gamma \in [0,1)$ such that if π is γ'-discount optimal for $\gamma' > \gamma$, then π is Blackwell optimal.

Theorem. The Blackwell discount factor γ_{bw} exists in any finite MDP.

Why is this useful? If we know γ_{bw} or an upper bound $\hat{\gamma}$, we can compute Blackwell and average optimal policies by solving discounted MDPs, a problem that has been studied extensively!

How to bound the Blackwell discount factor?

A bound must depend on the "granularity" of the instance data $\boldsymbol{r}$ and $\boldsymbol{P}$:

Proposition. For any $\eta > 0$, there exists an MDP instance $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{r}, \boldsymbol{P})$ with $|\mathcal{S}| = 2, |\mathcal{A}| = 2$ and deterministic transitions, such that $\gamma_{\mathsf{bw}} > 1 - \eta$.

Main theorem

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{r}, \boldsymbol{P})$ be an MDP instance with a maximum bit-size $m \in \mathbb{N}$. Then we have $\gamma_{\mathsf{bw}} < 1 - \eta_{\mathcal{M}}$, with $\eta_{\mathcal{M}} \in (0,1)$ defined as

$$\eta_{\mathcal{M}} = \frac{1}{2N^{N/2+2}(L+1)^N},$$
$$N = 2|\mathcal{S}| - 1,$$
$$L = 2 \cdot |\mathcal{S}| \cdot r_\infty \cdot m^{2|\mathcal{S}|} \cdot 4^{|\mathcal{S}|}.$$

Intuition for our main theorem

The proof of our main results rely on the following intuition:

1. For any two policies π, π' we want a bound on the largest $\gamma \in (0,1)$ such that
   $$v_{\gamma,s}^\pi = v_{\gamma,s}^{\pi'}.$$
2. For any policy π, the value function $\gamma \mapsto v_{\gamma,s}^\pi$ is a *rational* function:
   $$v_{\gamma,s}^\pi = \frac{N^{\pi,s}(\gamma)}{D^\pi(\gamma)}$$
   for two polynomials $N^{\pi,s}(X), D^\pi(X)$.
3. Therefore, if at $\gamma \in (0,1)$ we have $v_{\gamma,s}^\pi = v_{\gamma,s}^{\pi'}$, then γ is the root of a polynomial equation:
   $$N^{\pi,s}(\gamma)D^{\pi'}(\gamma) - N^{\pi',s}(\gamma)D^\pi(\gamma) = 0.$$
   Note that 1 is always a zero of the above equation.
4. We can separate the root of a polynomial equation:

   Theorem. ([Rum79]) Let p be a polynomial of degree N with integer coefficients. Let L be the sum of the absolute values of its coefficients. The distance between any two distinct roots of p is strictly larger than $\eta > 0$, with $\eta = 2N^{-N/2+2}(L+1)^{-N}$.
5. We apply this theorem to separate γ_{bw} and 1: $\gamma_{\mathsf{bw}} < 1 - \eta < 1$ for some $\eta > 0$.

Corollary: weakly-polynomial time algorithms for Blackwell optimality.

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{r}, \boldsymbol{P})$ be an MDP instance with total bit-size $Q(\boldsymbol{r}, \boldsymbol{P}) \in \mathbb{N}$.

We can compute a Blackwell-optimal policy in $O\left(|\mathcal{S}|^5|\mathcal{A}|^2Q(\boldsymbol{r}, \boldsymbol{P})\right)$ arithmetic operations.

In the paper we extend part of these results from nominal MDPs to robust MDPs.

References

[HDK85] Arie Hordijk, Rommert Dekker, and Lodewijk Cornelis Maria Kallenberg.
Sensitivity-analysis in discounted Markovian decision problems.
Operations-Research-Spektrum, 7(3):143–151, 1985.

[JS21] Yujia Jin and Aaron Sidford.
Towards tight bounds on the sample complexity of average-reward MDPs.
In *International Conference on Machine Learning*, pages 5055–5064. PMLR, 2021.

[OVJ17] Michael O'Sullivan and Arthur F Veinott Jr.
Polynomial-time computation of strong and n-present-value optimal policies in Markov decision chains.
Mathematics of Operations Research, 42(3):577–598, 2017.

[Rum79] Siegfried M Rump.
Polynomial minimum root separation.
Mathematics of Computation, 33(145):327–336, 1979.

[WWY22] Jinghan Wang, Mengdi Wang, and Lin F Yang.
Near sample-optimal reduction-based policy learning for average reward MDP.
arXiv preprint arXiv:2212.00603, 2022.