Regret minimization in Average Reward Reinforcement Learning
I. BASICS

Odalric-Ambrym MAILLARD
Centre Inria de l'Université de Lille
*Equipe **SCOOL***
(Sequential, Continual and Online Learning)

Master MVA

GLOBAL OUTLINE

❖ I. Basics of Average-reward RL
Regret definitions.
Gain, Bias, Diameter, Span.
Average-optimal Bellman operators, and Value iteration.

❖ II. Regret minimization algorithms
UCB for MDPs, TS for MDPs.
Concentration inequalities.
Instance-dependent performance bounds for bandits and MDPs.

❖ II. Extensions
The most confusing instance paradigm.
IMED for MDPs.
Linear quadratic and beyond.

References


Martin L Puterman.
Markov decision processes: discrete stochastic dynamic programming.
John Wiley & Sons, 2014.


Thomas Jaksch, Ronald Ortner, and Peter Auer.
Near-optimal regret bounds for reinforcement learning.
Journal of Machine Learning Research, 11:1563–1600, 2010.


Odalric-Ambrym Maillard.
Mathematics of statistical sequential decision making.
PhD thesis, Habilitation, Université de Lille, Sciences et Technologies, 2019.

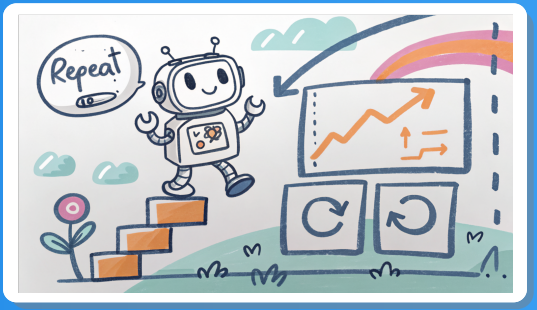

Fabien Pesquerel and Odalric-Ambrym Maillard.
Imed-rl: Regret optimal learning of ergodic markov decision processes.
Advances in Neural Information Processing Systems, 35:26363–26374, 2022.


Xi-Ren Cao.
Stochastic learning and optimization-a sensitivity-based approach.
IFAC Proceedings Volumes, 41(2):3480–3492, 2008.

REINFORCEMENT LEARNING

Learning **to act** in an **unknown/uncertain** environment

Learning by **Trials and errors** (automated)

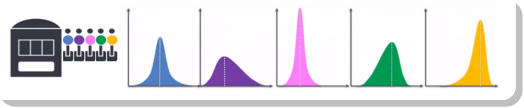


Decisional AI : **choose an action** at each decision step.

TWO FRAMEWORKS

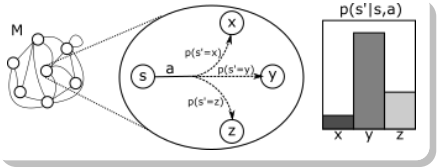
Learning by **Trials and Errors**
Sequential (adaptive) **Decision making**

❖ **Multi-armed Bandits** (MAB)



Uncertainty

❖ **Markov Decision Process** (MDP)



Dynamics

Take home message

- ☑ Regret-minimization in MDPs.
- ☑ Communicating vs Ergodic MDPs.
- ☑ Invariant probability measure .
- ☑ Poisson equation , gain $\mathbf{g}_\pi$, bias $\mathbf{b}_\pi$.
- ☑ Diameter D, Span semi-norm $\mathbb{S}(.)$
- ☑ Average-value iteration
- ☑ Intrinsic contraction , coalescence: No discount is ok.

WHY REGRET MINIMIZATION?

In a simulated world, mistakes have marginal cost (you can die and restart).
In the real-world, mistakes are costly, you cannot restart.

To address Sim-to-real gap, **regret minimization** captures the cost of mistakes while learning.

Example: Agronomy trials

- Similar fields (yellow cohort, orange cohort) = similar MDPs.

- A few practices you are uncertain about.
- Perform many trials, repeatedly in space and time.
- Each trial is costly: minimize cumulative errors.

OUTLINE

1] Value and regret

2] Complexity measures

3] ...

Notations

Markov Decision Process $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{r}, \mathbf{p})$
Reward distribution function $\mathbf{r} : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R})$, and mean $\mathbf{m} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.
State transition distribution function $\mathbf{p} : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$.

- Each policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ induces a Markov chain $\mathbf{M}_\pi = (\mathcal{S}, \mathbf{r}_\pi, \mathbf{p}_\pi)$:

$$\text{(Policy mean)} \quad \mathbf{m}_\pi(s) = \sum_{a \in \mathcal{A}_s} \mathbf{m}(s,a)\pi(a|s),$$

$$\text{(Policy transition)} \quad \mathbf{p}_\pi(s'|s) = \sum_{a \in \mathcal{A}_s} \mathbf{p}(s'|s,a)\pi(a|s),$$

$$(t\text{-step transition}) \quad \mathbf{p}_\pi^t(s'|s) = \mathbb{P}_\pi(s_t = s'|s_1 = s)\,.$$

Total/Discounted/Average Value

The cumulated value of policy π run for T steps from initial state s_1 is

$$\mathbf{V}_{T,\pi}(s_1) = \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t)\right] = \mathbf{m}_\pi(s_1) + (\mathbf{p}_\pi \mathbf{m}_\pi)(s_1) + \ldots = \sum_{t=1}^{T} (\mathbf{p}_\pi^{t-1} \mathbf{m}_\pi)(s_1)\,.$$

where $a_t \sim \pi(s_t)$, $s_{t+1} \sim \mathbf{p}(\cdot|s_t, a_t)$, $r(s,a) \sim \mathbf{r}(s,a)$ with mean $\mathbf{m}(s,a)$.

For a policy π, the total-average/average/discounted values are:

$$\overline{\mathbf{v}}_{T,\pi} = \overline{\mathbf{p}}_{T,\pi}\mathbf{m}_\pi \quad \overline{\mathbf{v}}_\pi = \overline{\mathbf{p}}_\pi \mathbf{m}_\pi \quad \mathbf{v}_{\gamma,\pi} = \mathbf{p}_{\gamma,\pi}\mathbf{m}_\pi, \text{where}$$

$$\overline{\mathbf{p}}_{T,\pi} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{p}_\pi^{t-1}, \quad \overline{\mathbf{p}}_\pi = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbf{p}_\pi^{t-1}, \quad \mathbf{p}_{\gamma,\pi} = (1-\gamma)\sum_{t=1}^{\infty} \gamma^{t-1}\mathbf{p}_\pi^{t-1}$$

Markov chain terminology

- π is Irreducible if all states are communicating under π:

$$\forall s, s' \in \mathcal{S}, \exists t < \infty : \mathbf{p}_\pi^t(s'|s) > 0.$$

- **M** is Communicating if every pair of states commnunicate under **some** policy

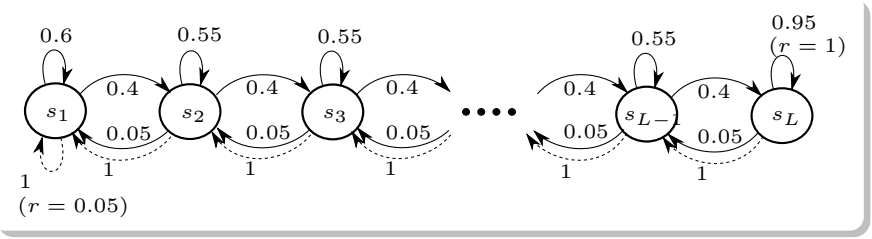
$$\forall s, s' \in \mathcal{S}, \exists \pi, \exists t < \infty : \mathbf{p}_\pi^t(s'|s) > 0.$$

M is "Ergodic" if every pair of states commnunicate under **all** policies

$$\forall \pi \forall s, s' \in \mathcal{S}, , \exists t < \infty : \mathbf{p}_\pi^t(s'|s) > 0.$$

Example: Riverswim MDP

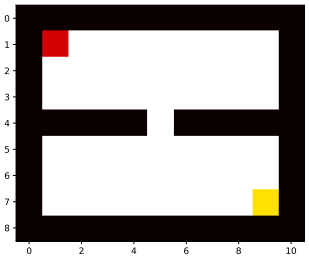
$\mathcal{S} = \{1, 2, \ldots, L\}$, $\mathcal{A} = \{\text{left,right}\}$, $p = 0.4$ to reach right when go right.



Communicating, but not ergodic: policy going left is not irreducible (e.g. not possible to reach state 4 starting from state 3)

Example: Gridworld MDP

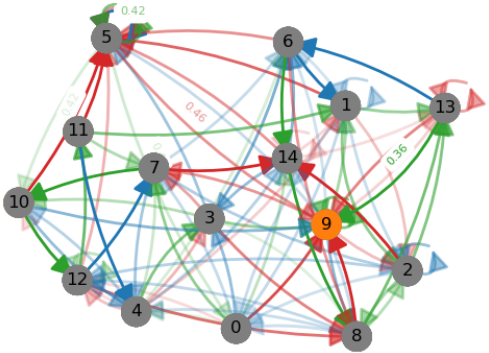
$\mathcal{S} = \{$ positions in the maze $\}$, $\mathcal{A} = \{$up,down,left,right$\}$, reward 1 in yellow state, 0 else. Proba $p = 0.9$ to reach desired state.



Communicating, but not ergodic.

Example: Randomly Generated MDP

Generic Commmunicating MDP with stochastic transitions.



Communicating, usually not ergodic.

OUTLINE

1] Value and regret

2] Complexity measures

3] Fundamental quantities

4] ...

Regret of learning agent

Set of (stationary) average-optimal policies:

$$\overline{\mathcal{O}}(\mathbf{M}, s_1) \;=\; \{\pi \in \Pi : \overline{\mathbf{v}}_\pi(s_1) \geqslant \max_\pi \overline{\mathbf{v}}_\pi(s_1)\}, \qquad \overline{\mathcal{O}}(\mathbf{M}) = \bigcap_{s \in \mathcal{S}} \overline{\mathcal{O}}(\mathbf{M}, s)$$

Learning while optimizing in a single stream of interaction.

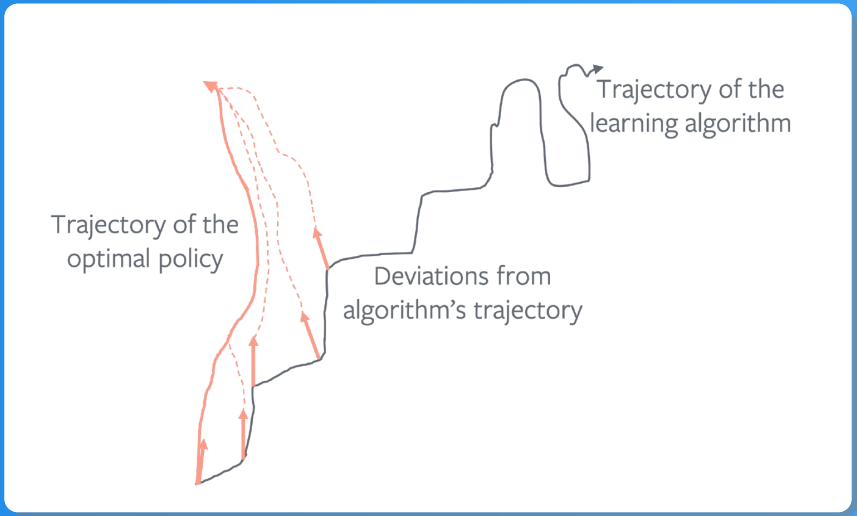
The cumulated value of $\boldsymbol{\pi} = (\pi_t)_t$ (that plays policy π_t at time $t \in [T]$) is

$$\mathbf{V}_{T,\pi}(s_1) = \mathbb{E}\Big[\sum_{t=1}^{T} r(s_t, a_t)\Big] = \sum_{t=1}^{T} \left(\left(\prod_{t'=1}^{t-1} \mathbf{p}_{\pi_{t'}}\right) \mathbf{m}_{\pi_t}\right)(s_1)$$

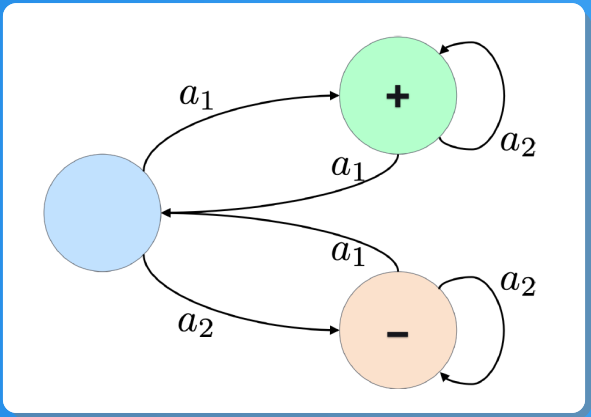
The cumulated regret of $\boldsymbol{\pi} = (\pi_t)_t$ with respect to an optimal policy $\pi^\star$ is

$$\mathfrak{R}_T(\boldsymbol{\pi}) \;=\; \mathbf{V}_{T,\pi^\star}(s_1) - \mathbf{V}_{T,\pi}(s_1) \text{ where } \pi^\star \in \overline{\mathcal{O}}(\mathbf{M})$$

Sample-complexity vs Regret minimization

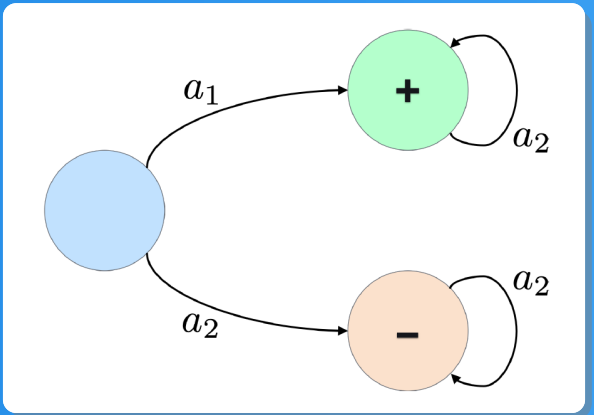


Sample-complexity vs Regret minimization



Sample-complexity: **Easy**. Regret minimization: **Easy**.

Sample-complexity vs Regret minimization



Sample-complexity: **Easy**. Regret minimization: **Impossible**

Diameter

❖ The Diameter of an MDP is the largest shortest path (path with minimal expected length) between any two states:

$$D(\mathbf{M}) = \max_{s,s' \in \mathcal{S}} \min_{\pi} \mathbb{E}[\min\{t > 1 : s_t = s'\} | s_1 = s]$$

❖ A finite, communicating MDP has finite diameter

❖ In grid-world type MDPs:

$$D(\mathbf{M}) \simeq \frac{\text{length of shortest path between maximally distant states}}{\text{probability of reaching desired next state}}.$$

E.g. River-swim $D \simeq L/p$. Two-room MDP: $D \simeq 15/p$

OUTLINE

1] Value and regret

2] Complexity measures

3] Fundamental quantities

4] Average value iteration

5] ...

Gain

❖ Definition (gain)
The average value $\overline{\mathbf{v}}_\pi = \overline{\mathbf{p}}_\pi \mathbf{m}_\pi$ is also called the gain, denoted $\mathbf{g}_\pi$.
A gain optimal policy $\star$ satisfies $\mathbf{g}_\star = \max_\pi \mathbf{g}_\pi$.

❖ Invariance $\mathbf{p}_\pi \mathbf{g}_\pi = \mathbf{g}_\pi$, thanks to the property:

$$\overline{\mathbf{p}}_\pi \mathbf{p}_\pi = \mathbf{p}_\pi \overline{\mathbf{p}}_\pi = \overline{\mathbf{p}}_\pi \overline{\mathbf{p}}_\pi = \overline{\mathbf{p}}_\pi \,.$$

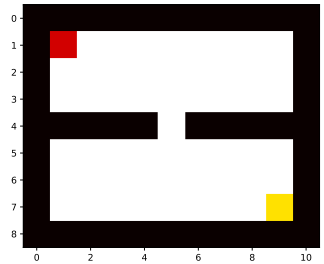
❖ The gain is a constant function of the initial state in a communicating MDP: $\forall s, \mathbf{g}_\pi(s) = g$

Bias

❖ Definition (Bias function)

$$\mathbf{b}_\pi \;=\; \sum_{t=1}^{\infty}(\mathbf{p}_\pi^{t-1} - \overline{\mathbf{p}}_\pi)\mathbf{m}_\pi = [I - \mathbf{p}_\pi + \overline{\mathbf{p}_\pi}]^{-1}[I - \overline{\mathbf{p}}_\pi]\mathbf{m}_\pi \,,$$

❖ Selectivity of $\mathbf{g}_\pi$ vs $\mathbf{b}_\pi$



Red state absorbing with reward 1 (0 elsewhere).

All policies reaching red state satisfy $\overline{\mathbf{p}}_\pi\mathbf{m}_\pi = \mathbf{g}_\pi = 1$ and are **optimal**.

$\mathbf{b}_\pi$ tells how long π takes to reach it.

(Beyond: "Blackwell" optimality)

Bellman equation

Lemma (Bias and Gain)

(Poisson equation) $\mathbf{b}_\pi = \mathbf{m}_\pi - \mathbf{g}_\pi + \mathbf{p}_\pi \mathbf{b}_\pi$

- Similar to Bellman equation for $\mathbf{b}_\pi$, except for additional $\mathbf{g}_\pi$:

(Bellman equation) $\mathbf{V}_{\gamma,\pi} = \mathbf{m}_\pi + \gamma \mathbf{p}_\pi \mathbf{V}_{\gamma,\pi}$

(Discounted Bellman equation) $\mathbf{v}_{\gamma,\pi} = (1-\gamma)\mathbf{m}_\pi + \gamma \mathbf{p}_\pi \mathbf{v}_{\gamma,\pi}$

$\mathbf{m}_\pi - \mathbf{g}_\pi = (I - \overline{\mathbf{p}}_\pi)\mathbf{m}_\pi$ "similar to" $(1-\gamma)\mathbf{m}_\pi$

- Unlike Bellman equation, Poisson equation admits many solutions:

if $\mathbf{b}_\pi$ then $\mathbf{b}_\pi + c\mathbf{1}$ is also solution for all constant c.

The Span semi-norm

No contraction of the Bellman operator in the usual sense ($\|\cdot\|_\infty$) !

❖ The span operator is

$$\mathbb{S}(f) = \max_x f(x) - \min_x f(x)$$

- It satisfies $\mathbb{S}(f + c\mathbf{1}) = \mathbb{S}(f)$ for any constant c.
- It is a semi-norm.

❖ In some cases (see Bonus), contraction in $\mathbb{S}(\cdot)$ (instead of $\|\cdot\|_\infty$)

Regret decomposition

❖ Sub-optimal gap (aka advantage function) is

$$\begin{aligned}\Delta(s,a) &= \mathbf{m}_\star(s) + (\mathbf{p}_\star \mathbf{b}_\star)(s) - \mathbf{m}_a(s) - (\mathbf{p}_a \mathbf{b}_\star)(s) \\ &= \mathbf{Q}(s,\pi^\star(s)) - \mathbf{Q}(s,a)\end{aligned}$$

Pseudo-regret

Consider a communicating MDP and $\boldsymbol{\pi} = (\pi_t)_t$ policy

$$\mathfrak{R}_T(\boldsymbol{\pi}) = \underbrace{\sum_{s,a} \mathbb{E}_\pi[N_T(s,a)]\Delta(s,a)}_{\text{pseudo-regret}} + \underbrace{\left(\left[\prod_{t=1}^{T} \mathbf{p}_{\pi_t} - \mathbf{p}_\star^T\right]\mathbf{b}_\star\right)(s_1)}_{\leqslant \mathbb{S}(\mathbf{b}_\star)}.$$

❖ Reminiscent of multi-armed bandits!

Proof outline

$$\mathbf{V}_{T,\pi}(s_1) \;=\; \sum_{t=1}^{T}\Big(\prod_{t'=1}^{t-1}\mathbf{p}_{\pi_{t'}}\mathbf{m}_{\pi_t}\Big)(s_1) = \sum_{t=1}^{T}\Big(\prod_{t'=1}^{t-1}\mathbf{p}_{\pi_{t'}}(\mathbf{g}_{\pi_t}+(I-\mathbf{p}_{\pi_t})\mathbf{b}_{\pi_t})\Big)(s_1)\,.$$

When $\mathbf{g}_{\star,s_1} \equiv g_\star$ (e.g. unichain optimal policy), then $\forall\pi, (\mathbf{p}_\pi \mathbf{g}_\star)(s_1) \equiv g_\star$:

$$\begin{aligned}
\mathfrak{R}_T(\pi) \;&=\; \sum_{t=1}^{T}\Big(\prod_{t'=1}^{t-1}\mathbf{p}_{\pi_{t'}}\underbrace{[(g_\star-\mathbf{g}_{\pi_t})+(I-\mathbf{p}_{\pi_t})(\mathbf{b}_\star-\mathbf{b}_{\pi_t})]}_{\mathbf{\Delta}\pi_t}\Big)(s_1)\\
&\quad+\sum_{t=1}^{T}\Big([\mathbf{p}_\star^{t-1}-\mathbf{p}_\star^{t}-\prod_{t'=1}^{t-1}\mathbf{p}_{\pi_{t'}}+\prod_{t'=1}^{t}\mathbf{p}_{\pi_{t'}}]\mathbf{b}_\star\Big)(s_1)\\
&=\; \Big(\sum_{t=1}^{T}\Big(\prod_{t'=1}^{t-1}\mathbf{p}_{\pi_{t'}}\Big)\mathbf{\Delta}_{\pi_t}\Big)(s_1)+\underbrace{\Big(\Big[\prod_{t'=1}^{T}\mathbf{p}_{\pi_{t'}}-\mathbf{p}_\star^{T}\Big]\mathbf{b}_\star\Big)(s_1)}_{\leqslant\mathbb{S}(\mathbf{b}_\star)}\,.
\end{aligned}$$

Proof outline II

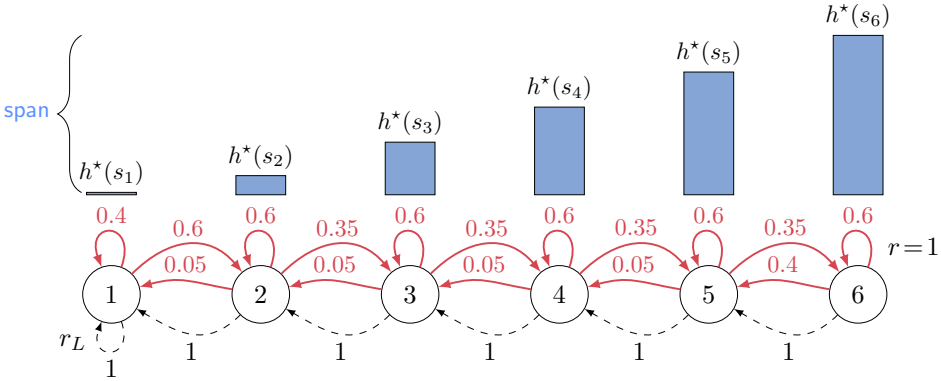
$$
\begin{aligned}
\boldsymbol{\Delta}_\pi(s) &= g_\star - \mathbf{g}_\pi(s) + (I - \mathbf{p}_\pi)(\mathbf{b}_\star - \mathbf{b}_\pi)(s) \\
&= \mathbf{m}_\star(s) - \mathbf{m}_\pi(s) + ([\mathbf{p}_\star - \mathbf{p}_\pi]\mathbf{b}_\star)(s) \\
&= \mathbb{E}_\pi[\Delta(s,a)]\,.
\end{aligned}
$$

$$
\begin{aligned}
\sum_{t=1}^{T} [\mathbf{p}_{\pi_1}\mathbf{p}_{\pi_2}\cdots\mathbf{p}_{\pi_{t-1}}\boldsymbol{\Delta}_{\pi_t}](s_1) &= \sum_{s,a}\sum_{t=1}^{T} \mathbb{E}_{\pi_1,\ldots\pi_t}[\Delta(s,a)\mathbb{I}\{S_t = s, A_t = a\}] \\
&= \sum_{s,a} \Delta(s,a)\mathbb{E}[N_T(s,a)]\,.
\end{aligned}
$$

Span and Diameter

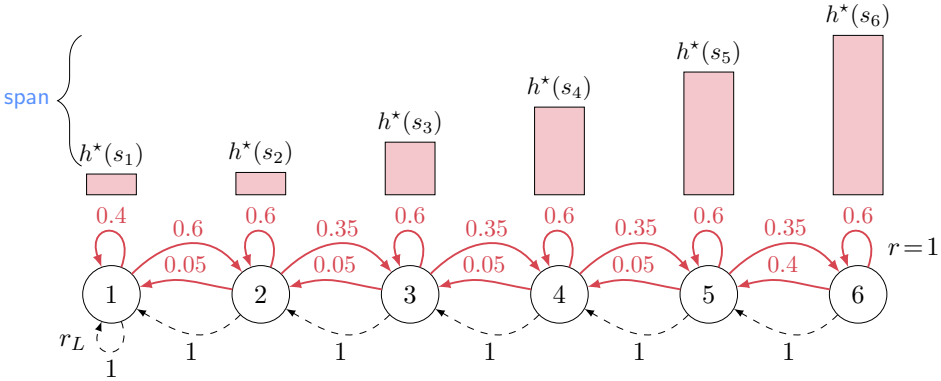
Assuming $\mathbf{r}$ is supported in $[-B, B]$, $\mathbb{S}(\mathbf{b}_\star) \leqslant B \times D(\mathbf{M})$

River Swim: Optimality



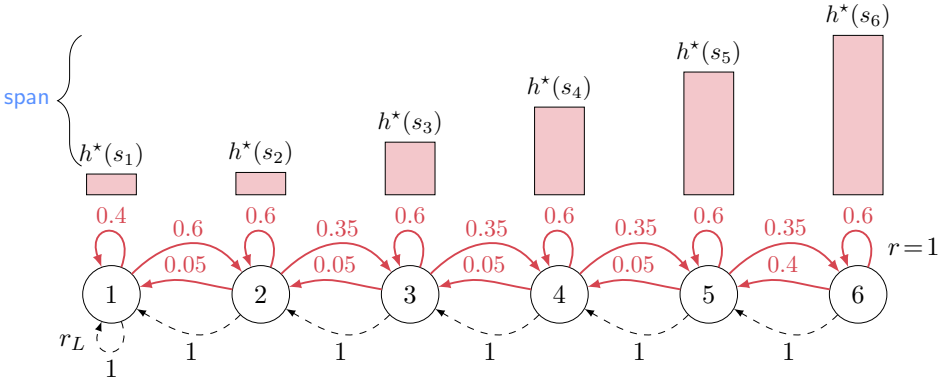
- $\pi^\star = \pi_R$
- If $r_L = 0.01$, $g^\star \approx 0.43$, $\mathsf{sp}(h^\star) \approx 6.4$

River Swim: Optimality



- $\pi^\star = \pi_R$
- If $r_L = 0.01$, $g^\star \approx 0.43$, $\mathsf{sp}(h^\star) \approx 6.4$
- If $r_L = 0.4$, $g^\star \approx 0.43$, $\mathsf{sp}(h^\star) \approx 5.5$

River Swim: Optimality



- $\pi^\star = \pi_R$
- If $r_L = 0.01$, $g^\star \approx 0.43$, $\mathsf{sp}(h^\star) \approx 6.4$
- If $r_L = 0.4$, $g^\star \approx 0.43$, $\mathsf{sp}(h^\star) \approx 5.5$

} D is constant

OUTLINE

1] Value and regret

2] Complexity measures

3] Fundamental quantities

4] Average value iteration

5] Bonus: Intrinsic contraction

Value iteration

Given $\mathbf{M}$, how do we compute π such that $\mathbf{g}_\star - \mathbf{g}_\pi \leqslant \varepsilon$?

Value iteration

Given $\mathbf{M}$, how do we compute π such that $\mathbf{g}_\star - \mathbf{g}_\pi \leqslant \varepsilon$?

Value iteration computes a sequence of functions $(\mathbf{u}_n)_{n\in\mathbb{N}}$ and policies $(\pi_n)_{n\in\mathbb{N}}$ according to the following equations

$$\forall n \in \mathbb{N} \begin{cases} \mathbf{u}_{n+1}(s) = \max\limits_{a\in\mathcal{A}} \mathbf{m}(s,a) + (\mathbf{p}_a \mathbf{u}_n)(s)\,, & \text{where } \mathbf{u}_0 = 0 \\ \pi_{n+1}(s) \in \operatorname*{Argmax}\limits_{a\in\mathcal{A}} \mathbf{m}(s,a) + (\mathbf{p}_a \mathbf{u}_n)(s)\,. \end{cases}$$

π_{n+1} is called a $\mathbf{u}_n$-improving policy (Greedy w.r.t. $\mathbf{u}_n$).

$$\mathbf{u}_n \quad = \quad \sum_{i=1}^{n} \left(\mathbf{p}_{\pi_n} \dots \mathbf{p}_{\pi_{i+1}}\right) \mathbf{m}_{\pi_i}$$

Value iteration errors

Average-gain guarantee? When to stop?

Value error control

Let $\varepsilon > 0$ and n be such that $\mathbf{g}_\star - \mathbf{g}_{\pi_{n+1}} \leqslant \varepsilon$. Then

$$\mathbf{V}_{\star,T}(s_1) - \mathbf{V}_{\pi_{n+1},T}(s_1) \quad \leqslant \quad \mathbf{b}_{\star,T}(s_1) + \varepsilon T + \mathbb{S}(\mathbf{b}_{\pi_{n+1}}) \ .$$

(where $\mathbf{b}_{\star,T}$ is T-step truncation of $\mathbf{b}_\star$).

How to control $\mathbb{S}(b_{\pi_{n+1}})$? How to ensure $\mathbf{g}_\star - \mathbf{g}_{\pi_{n+1}} \leqslant \varepsilon$?

Span is controlled by diameter

Lemma)

For all n (starting from $u_0 = 0$), if rewards are bounded by 1,

$$\mathbb{S}(u_n) \leqslant D(\mathbf{M}), \qquad \mathbb{S}(b_{\pi_n}) \leqslant D(\mathbf{M})\,.$$

Sketch (informal):

Span is controlled by diameter

Lemma)

For all n (starting from $u_0 = 0$), if rewards are bounded by 1,

$$\mathbb{S}(u_n) \leqslant D(\mathbf{M}), \qquad \mathbb{S}(b_{\pi_n}) \leqslant D(\mathbf{M})\,.$$

Sketch (informal):

- Otherwise $\exists s_1, s_2$ s.t. $\mathbf{u}_n(s_1) - \mathbf{u}_n(s_2) > D$, that is $\mathbf{u}_n(s_2) < \mathbf{u}_n(s_1) - D$.

Span is controlled by diameter

Lemma)

For all n (starting from $u_0 = 0$), if rewards are bounded by 1,

$$\mathbb{S}(u_n) \leqslant D(\mathbf{M}), \qquad \mathbb{S}(b_{\pi_n}) \leqslant D(\mathbf{M})\,.$$

Sketch (informal):

- Otherwise $\exists s_1, s_2$ s.t. $\mathbf{u}_n(s_1) - \mathbf{u}_n(s_2) > D$, that is $\mathbf{u}_n(s_2) < \mathbf{u}_n(s_1) - D$.
- However, it takes at most D steps to reach s_2 from s_1, by some policy π_{Fast}. Since rewards are bounded by 1, we loose at most D rewards by following π_{Fast} from s_2 to reach s_1 then continue.
- The value of such a combined policy is at least $\mathbf{u}_n(s_1) - D$, and $\mathbf{u}_n(s_2)$ should be larger than this quantity by optimality.

Gain optimality of VI

How to ensure $\mathbf{g}_\star - \mathbf{g}_{\pi_{n+1}} \leqslant \varepsilon$? Answer: stop when $\mathbb{S}(\mathbf{u}_{n+1} - \mathbf{u}_n) \leqslant \varepsilon$

Lemma [Sandwich bounds]

$$\forall n \in \mathbb{N}, \quad \overline{\mathbf{p}}_{\pi_{n+1}}[\mathbf{u}_{n+1} - \mathbf{u}_n] \leqslant \mathbf{g}_{\pi_{n+1}} \leqslant \mathbf{g}_\star \leqslant \overline{\mathbf{p}}_\star[\mathbf{u}_{n+1} - \mathbf{u}_n]\,.$$

Proof hint: $\mathbf{g}_\pi = \mathbf{m}_\pi + \mathbf{p}_\pi \mathbf{b}_\pi - \mathbf{b}_\pi = \mathcal{T}_\pi[\mathbf{b}_\pi] - \mathbf{b}_\pi$

Corollary [Value and gain]

For each n, it holds $\quad \mathbf{g}_\star - \mathbf{g}_{\pi_{n+1}} \leqslant \mathbb{S}(\mathbf{u}_{n+1} - \mathbf{u}_n)$.

OUTLINE

1] Value and regret

2] Complexity measures

3] Fundamental quantities

4] Average value iteration

5] Bonus: Intrinsic contraction

Contraction

Value iteration usually relies on existence of a fixed point ...

- But under average-gain, there is no discount ($\gamma = 1.$)
  - Where is the contraction property?
  - Is value iteration converging at all (no discount) ?
  - Where is contraction of Bellman operator $T_\pi[f] = \mathbf{m}_\pi + \mathbf{p}_\pi f$?

- It turns out, contraction property may hold w.r.t. to the Span operator $\mathbb{S}(f) = \max_x f(x) - \min_x f(x)$, instead of the $\|\cdot\|_\infty$ norm.

Contraction properties

Lemma [Contraction in span semi-norm]

The Bellman operator has the following contraction property:

$$\forall f, g \in \mathbb{R}^S, \quad \mathbb{S}(T_\pi[f] - T_\pi[g]) \leqslant \frac{1}{2}\|\mathbf{p}_\pi(\cdot|\overline{s}) - \mathbf{p}_\pi(\cdot|\underline{s})\|_1 \mathbb{S}(f-g)$$

where $\overline{s} = \operatorname{argmax}_{s\in\mathcal{S}} \mathbf{p}_\pi(f-g)(s)$, $\underline{s} = \operatorname{argmin}_{s\in\mathcal{S}} \mathbf{p}_\pi(f-g)(s)$.

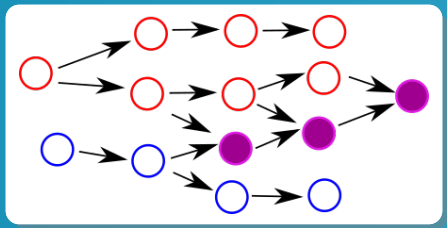
- Further,

$$\frac{1}{2}\|\mathbf{p}_\pi(\cdot|\overline{s}) - \mathbf{p}_\pi(\cdot|\underline{s})\|_1 = 1 - \sum_{s'\in\mathcal{S}} \min(\mathbf{p}_\pi(s'|\overline{s}), \mathbf{p}_\pi(s'|\underline{s}))$$

The higher the probability of reaching same states from $\overline{s}$ and $\underline{s}$, the more contraction.

Contraction as coalescence

❖ The higher the probability of coalescing from two different states, the more contraction.



Remark: In discounted MDP, $1-\gamma$ interprets as probability to reach same external state $\perp$ from any state: all policies coalesce in a single step.

Intrinsic contraction coefficient

Definition [Policy contraction coefficient]

We define the one step and multi-step contraction coefficients as:

$$\forall k, \gamma_{\pi,k} \;=\; \max_{s_1,s_2} \frac{1}{2}\|\mathbf{p}_\pi^k(\cdot|s_1) - \mathbf{p}_\pi^k(\cdot|s_2)\|_1 = 1 - \min_{s_1,s_2} \sum_{s' \in \mathcal{S}} \min(\mathbf{p}_\pi^k(s'|s_1), \mathbf{p}s_\pi^k(s'|s_2))$$

- In particular

$$\mathbb{S}(T_\pi^k[f] - T_\pi^k[g]) \leqslant \min_{\ell \in \{1,\dots,k\}} \gamma_{\pi,\ell}^{k/\ell} \mathbb{S}(f-g)$$

Suggests multi-step analyses $\min\{k : \gamma_{\pi,k} < 1\}$.

OUTLINE

1] Value and regret

2] Complexity measures

3] Fundamental quantities

4] Average value iteration

5] Bonus: Intrinsic contraction

Self-check

- ☑ Regret-minimization in MDPs.
- ☑ Communicating vs Ergodic MDPs.
- ☑ Invariant probability measure .
- ☑ Poisson equation , gain $\mathbf{g}_\pi$, bias $\mathbf{b}_\pi$.
- ☑ Diameter $D(\mathbf{M})$, Span semi-norm $\mathbb{S}(.)$
- ☑ Average-value iteration convergence and stopping criterion.
- ☑ Intrinsic contraction , coalescence: No discount is ok.

MERCI