Reinforcement Learning in a Birth and Death Process: Breaking the Dependence on the State Space

Louis-Sébastien Rebuffi Jonatha Anselmi, Bruno Gaujal

December 2nd, 2022

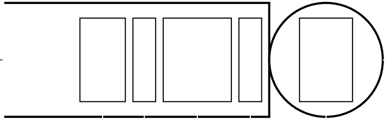
Outline

Example of a Simple Queue

Consider a processor with:

- Poisson arrivals with rate $\lambda\left(1-\frac{s_t}{S-1}\right)$.

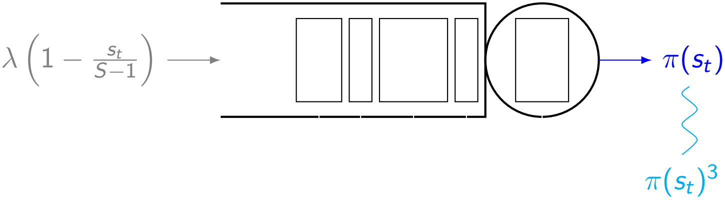
$\lambda\left(1-\frac{s_t}{S-1}\right) \longrightarrow$



Example of a Simple Queue

Consider a processor with:

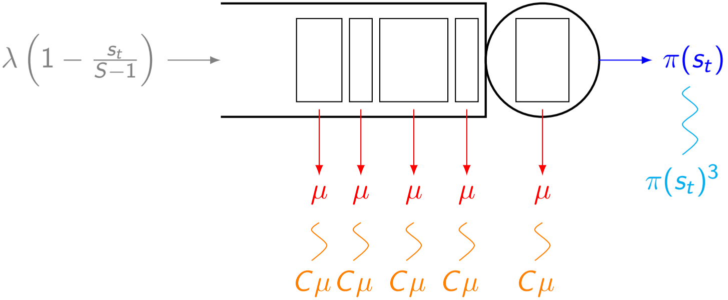
- Poisson arrivals with rate $\lambda \left(1 - \frac{s_t}{S-1}\right)$,
- Service rate $\pi(s_t)$ and power dissipation $\pi(s_t)^3$.



Example of a Simple Queue

Consider a processor with:

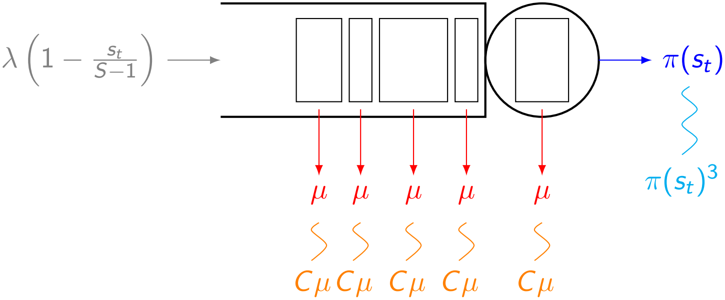
- Poisson arrivals with rate $\lambda\left(1-\frac{s_t}{S-1}\right)$,
- Service rate $\pi(s_t)$ and power dissipation $\pi(s_t)^3$,
- Jobs with Markovian deadlines that induce a cost C when dropped from the queue.



Example of a Simple Queue

Consider a processor with:

- Poisson arrivals with rate $\lambda\left(1-\frac{s_t}{S-1}\right)$,
- Service rate $\pi(s_t)$ and power dissipation $\pi(s_t)^3$,
- Jobs with Markovian deadlines that induce a cost C when dropped from the queue.



Objective: Find the optimal speeds that minimize the long term energy spent by the processor and the cost of missed deadlines.

Model as an MDP

Define the MDP $\mathcal{M} := (\mathcal{S} = \{0, \cdots, S-1\}, \mathcal{A} = \{0, \ldots, A-1\}, Q, c)$.

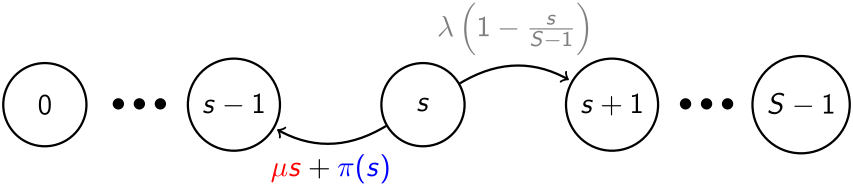


Figure: Transition diagram of the Markov chain induced by a policy $\pi : \mathcal{S} \to \mathcal{A}$.

With the expected instant cost:

$$c(s, \pi) := C\mu s + \pi(s)^3.$$

Definition of Rewards

Classically, different types of reward are considered, like the total discounted reward, for a discount $\gamma < 1$:

$$\rho_\gamma(\pi) := \lim_{T\to\infty} \sum_{t=1}^{T} \gamma^t \mathbb{E}[r(s_t, \pi(s_t))].$$

or the long-run average reward:

$$\rho(\pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r(s_t, \pi(s_t))].$$

While using discounted rewards would guarantee the stability of the system, the average reward is more suitable forqueuing systems.

Definition of the Regret

Let $\rho^* := \sup_\pi \rho(\pi)$ be the optimal average reward.

Definition (Regret)

For an MDP M, the *regret* at time T of a learning algorithm $\mathbb{L}$ is:

$$\mathrm{Reg}(M, \mathbb{L}, T) := T\rho^* - \sum_{t=1}^{T} r_t.$$

We place ourself in the context of tabular MDPs:

$\rightarrow$ Exploration: Learn the transitions probabilities and rewards for each state-action pair.

$\rightarrow$ Exploitation: Use the best policies to minimize the regret.

Definition of the Bias

Define the bias of any policy π:

$$h_\pi(s) := \mathbb{E}_\pi\left[\sum_{t=1}^{\infty}\left(r\left(s_t^\pi\right)-\rho(\pi)\right) \mid s_1^\pi = s\right], \quad \forall 0 \leq s \leq S-1,$$

In the computations of the regret, we use the Bellman equation:

$$\rho(\pi) + h_\pi(s) = r_\pi(s) + \sum_{s'} P_\pi(s' \mid s) h_\pi(s').$$

Relating the Diameter to the Bias

Usually, to control the bias, we need to introduce the diameter:

Definition

Letting $\tau\left(s'|\pi, s\right)$ be the time to go from s to s', the the diameter D of the MDP is:

$$D := \max_{s \neq s'} \min_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}\left[\tau\left(s'|\pi, s\right)\right].$$

Note that the bias and the diameter are related:

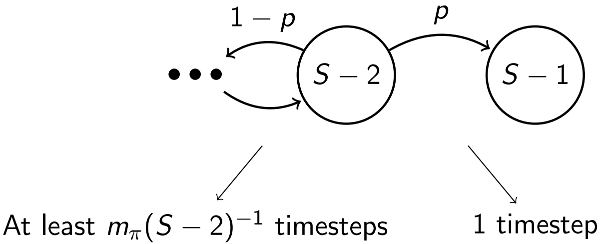
$$h_\pi(s) - h_\pi(s') \leq r_{\max} D_\pi$$

The diameter is an important quantity in the average reward case.

Diameter in our Model

Queues have a large diameter.

$\rightarrow$ We indeed have that: $D_\pi \geq \mathbb{E}\left[\tau(S-1 \mid \pi, S-2)\right]$. Letting $p := \mathbb{P}(S-1 \mid S-2, \pi(S-2))$.



In our example, the stationary measure m_π decreases exponentially, the diameter itself is therefore exponential in S.

Infinite Horizon in the Literature

In the average reward case:

- UCRL2-Jaksch et al. (2010), upper bound of the regret in $\tilde{\mathcal{O}}(r_{\max} DS\sqrt{AT})$.
- UCRL2B Fruit et al. (2019), regret bounded in $\tilde{\mathcal{O}}(r_{\max}\sqrt{D\Gamma SAT})$ with Γ the highest number of neighbours of any state.
- Using additional information in the algorithm, such as a upper bound on the bias H, in Zhang et al. (2019) the regret is bounded in $\tilde{\mathcal{O}}\left(r_{max}\right)\sqrt{HSAT}$.

In these algorithms, the parameters D and H still depend on S.

Using the MDP Structure in the Literature

Examples of algorithms using the structure of MDPs:

- We could think to use linear mixture models with d parameters. With discount γ, the regret is upper bounded by $r_{\text{max}} d\sqrt{T}/(1-\gamma)^2$ [Zhou et al. 2021], but it does not deal with the long-run average reward case.
- Model free algorithms with Q learning: Wei et al. (2020), regret bound close to $\mathcal{O}\left(r_{max}\sqrt{t_{mix}^3 SAT}\right)$ for ergodic MDPs, where t_{mix} is the mixing time of the MDP, which depends on S. This algorithm also requires additional information, such as a bound on t_{mix} and also on the worst hitting time.

Bounds for Generic MDPs

Theorem (Universal Lower Bound - Jaksch et al. (2010))

For any learning algorithm $\mathbb{L}$, any $S, A \geq 10$, $D \geq 20 \log_A S$, and $T \geq DSA$, there is an MDP M, such that:

$$\mathbb{E}[\mathrm{Reg}(M, \mathbb{L}, T)] \geq 0.015 r_{\max} \sqrt{DSAT}.$$

Existing learning algorithms have upper bounds in the regret that almost match this lower bound.

However this regret bound is unsatisfactory for queueing systems, where the diameter D is exponential in the number of states S.

MDPS for queuing systems have the following challenging characteristics:

- We consider the long run average reward rather than the total discounted reward.
- The considered MDPs have a large diameter D, *i.e.* a large expected time to cross the MDP.
- The transition matrices for queues are sparse and structured.

Our Contribution

We have seen that the previous bounds depend on the diameter, meaning that they are inaccurate for birth and death processes.

Question

When the underlying MDP has the structure of a queueing system, do the diameter D or the number of states S actually play a role in the regret?

To answer this question, we study the algorithm UCRL2 on our previous example.

UCRL2 Algorithm

Algorithm 1: The UCRL2 algorithm.

Set $s_1 = 0$
for *episodes* $k = 1, 2, \ldots$ **do**
  Initialize episode k with current reward and transition estimates $\hat{r}_k$ and $\hat{p}_k$.
  Find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$.
  Execute policy $\tilde{\pi}_k$ on the true MDP M until the end of the episode.
end

The optimistic MDP $\tilde{M}_k$ with policy $\tilde{\pi}_k$ is the queue of largest gain with rewards $\tilde{r}$ and transitions $\tilde{p}$ such that:

$$\forall (s,a), \quad |\tilde{r}(s,a) - \hat{r}_k(s,a)| \le r_{\max} \sqrt{\frac{2\log(At_k)}{\max\{1, N_{t_k}(s,a)\}}}$$

$$\forall (s,a), \quad \|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \le \sqrt{\frac{8\log(2At_k)}{\max\{1, N_{t_k}(s,a)\}}}$$

Monotonicity

For the class of MDPs $\mathcal{M}$, we assume the following assumptions hold:

Monotonicity

Denoting by π^0 a reference policy and π any other policy, if $s_0^{\pi} \leq_{st} s_0^{\pi^0}$, then for all t, $s_t^{\pi} \leq_{st} s_t^{\pi^0}$.

The reference policy controls the number of visits of a given state regardless of the chosen policy for the current episode:

Lemma

For $f : \mathcal{S} \to \mathbb{R}^+$ non-decreasing non-negative, we obtain

$$\mathbb{E}\left[\sum_{s \geq 0} f(s) N_t(s)\right] \leq t \sum_{s \geq 0} f(s) m^{\pi^0}(s).$$

Bound on the Bias

Reminder: define the optimal bias:

$$h_{\pi^*}(s) := \mathbb{E}_{\pi^*}\left[\sum_{t=1}^{\infty}\left(r\left(s_t^{\pi^*}\right) - \rho(\pi)\right) \mid s_1^{\pi^*} = s\right], \quad \forall 0 \le s \le S-1,$$

Bias Bound

There is a positive, bounded function Δ such that:

$$-\Delta(s) \le H(s) - H(s-1) \le 0.$$

Here, in our example $\Delta(s) = C$ is constant.

Main Result

Independently of S and D, let $E_2 := \left(\sum_{s \in \mathcal{S}} f(s)^{-1}\right) \mathbb{E}_{m^{\pi^0}}\left[(\Delta + r_{\max})^2 f\right]$, where $f : s \mapsto \frac{\max\{1, s(s-1)\}}{(\Delta(s) + r_{\max})^2}$.

Theorem

The expected regret achieved by UCRL2 *is upper bounded as follows:*

$$\mathbb{E}\left[\mathrm{Reg}(M, \mathrm{UCRL2}, T)\right] \leq 19\sqrt{E_2 A T \log(2AT)} + \mathcal{O}\left(T^{1/4}\right),$$

where the lower order term contains terms polynomial in D and S.

In the example, $E_2 \leq 12 r_{\max}^2 \left(1 + \frac{\lambda^2}{\mu^2}\right)$, so that the regret satisfies

$$\mathbb{E}\left[\mathrm{Reg}(M, \mathrm{UCRL2}, T)\right] = \mathcal{O}\left(r_{\max}\sqrt{AT\left(1 + \frac{\lambda^2}{\mu^2}\right)\log(AT)}\right).$$

Conclusion

- Despite using a basic and non-specific reinforcement learning algorithm, the analysis of the regret can be greatly improved when studying queueing systems.
- The regret bounds should not involve D nor S. Our bound relies instead on the stationary measure of a reference policy.
- This type of regret bound could be generalized to other queueing systems, such as optimal routing or admission control for example.