Lecture 5 - Reinforcement Learning in Markov Decision Processes

Pratik Gajane

September 21, 2022

2AMM20 Research Topics in Data Mining
Eindhoven University of Technology

A Quick Recap of Lecture 1, 2, 3 and 4

- Lecture 1 : Introduction to reinforcement learning and its basic elements.
- Lecture 2 : Upper confidence bound (UCB) for stationary stochastic bandits and its regret bound. Frequentist perspective.
- Lecture 3 : Thompson sampling for stationary stochastic bandits and its regret bound. Bayseian perspective.
- Lecture 4 : Non-stationary stochastic bandits, adversarial bandits, dueling bandits and contextual bandits.

A Quick Recap of Lecture 1, 2, 3 and 4

- Lecture 1 : Introduction to reinforcement learning and its basic elements.
- Lecture 2 : Upper confidence bound (UCB) for stationary stochastic bandits and its regret bound. Frequentist perspective.
- Lecture 3 : Thompson sampling for stationary stochastic bandits and its regret bound. Bayseian perspective.
- Lecture 4 : Non-stationary stochastic bandits, adversarial bandits, dueling bandits and contextual bandits.

A Quick Recap of Lecture 1, 2, 3 and 4

- Lecture 1 : Introduction to reinforcement learning and its basic elements.
- Lecture 2 : Upper confidence bound (UCB) for stationary stochastic bandits and its regret bound. Frequentist perspective.
- Lecture 3 : Thompson sampling for stationary stochastic bandits and its regret bound. Bayseian perspective.
- Lecture 4 : Non-stationary stochastic bandits, adversarial bandits, dueling bandits and contextual bandits.

A Quick Recap of Lecture 1, 2, 3 and 4

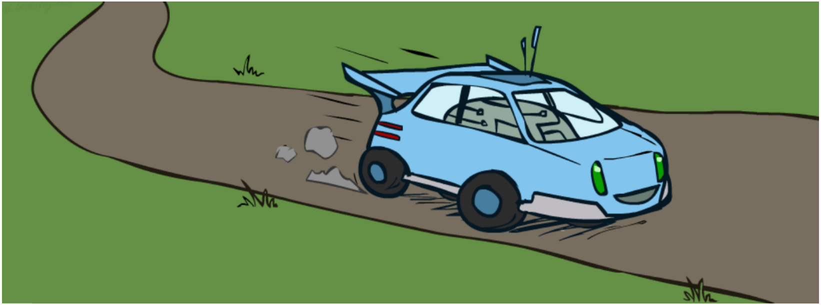
- Lecture 1 : Introduction to reinforcement learning and its basic elements.
- Lecture 2 : Upper confidence bound (UCB) for stationary stochastic bandits and its regret bound. Frequentist perspective.
- Lecture 3 : Thompson sampling for stationary stochastic bandits and its regret bound. Bayseian perspective.
- Lecture 4 : Non-stationary stochastic bandits, adversarial bandits, dueling bandits and contextual bandits.

Lecture 5 : Outline

- Markov decision processes.
- Mathematical setting and a lower bound on regret.
- A near-optimal algorithm $\mathrm{UCRL2}$.
- Regret analysis for $\mathrm{UCRL2}$.

Introduction to Markov Decision Processes

Markov Decision Process: Simple Example I



A race of robot cars

Image source: *UC Berkeley AI course, lecture 10*

Markov Decision Process: Simple Example II

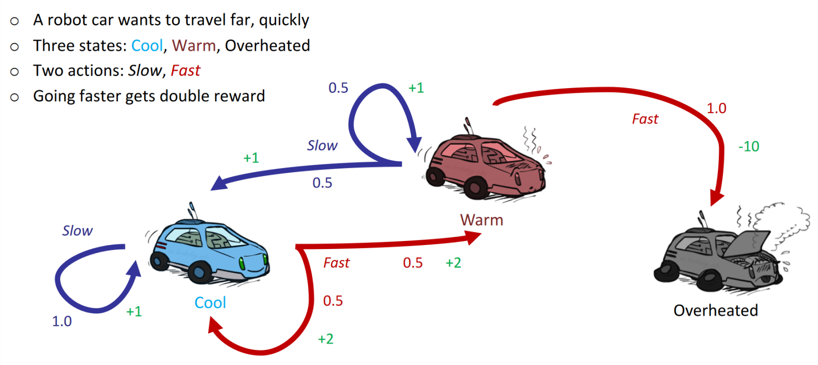


Image source: *UC Berkeley AI course, lecture 10*

- Going faster earns more rewards (usually), but runs the risk of overheating and not finishing the race.
  "To finish first, you must first finish".

Markov Decision Process : Simple Example III

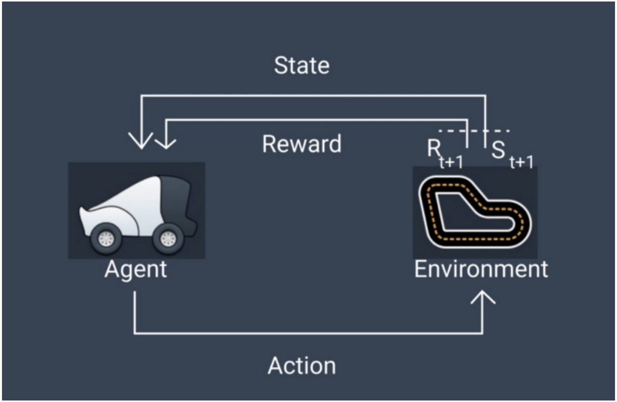


Image source: *Data science blog*

Mathematical Setting and a Lower bound on Regret

Mathematical Setting I

- Finite set of states $\mathcal{S}$ with $S = |\mathcal{S}|$.

Mathematical Setting I

- Finite set of states $\mathcal{S}$ with $S = |\mathcal{S}|$.
- Finite set of actions $\mathcal{A}$ with $A = |\mathcal{A}|$.

Mathematical Setting I

- Finite set of states $\mathcal{S}$ with $S = |\mathcal{S}|$.
- Finite set of actions $\mathcal{A}$ with $A = |\mathcal{A}|$.
- An initial state s_0.

Mathematical Setting I

- Finite set of states $\mathcal{S}$ with $S = |\mathcal{S}|$.
- Finite set of actions $\mathcal{A}$ with $A = |\mathcal{A}|$.
- An initial state s_0.
- When action a is executed in state s,
  - the learner receives a random reward drawn from an unknown distribution on $[0, 1]$ with mean reward $\bar{r}(s, a)$, and
  - a random transition to s' occurs according to unknown transition probabilities $p(s' \mid s, a)$.

Mathematical Setting I

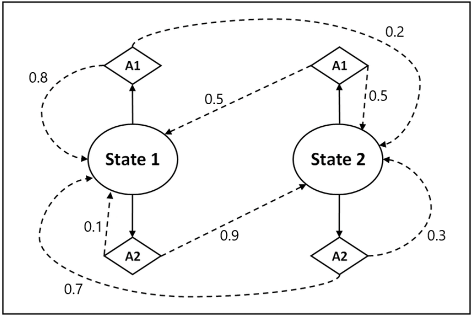
- Finite set of states $\mathcal{S}$ with $S = |\mathcal{S}|$.
- Finite set of actions $\mathcal{A}$ with $A = |\mathcal{A}|$.
- An initial state s_0.
- When action a is executed in state s,
  - the learner receives a random reward drawn from an unknown distribution on $[0, 1]$ with mean reward $\bar{r}(s, a)$, and
  - a random transition to s' occurs according to unknown transition probabilities $p(s' \mid s, a)$.
- Why Markov? "The future is independent of the past given the present". (For further understanding, consult slides 38,39 from lecture 1 in this track.)

Mathematical Setting II

- $\mathcal{S} = \{\text{State 1, State 2}\}$.
- $\mathcal{A} = \{A1, A2\}$.
- Initial state = State 1.
- Binary rewards = $\{0, 1\}$.
- Mean rewards $\bar{r}(s, a)$

	A1	A2
State 1	1	1
State 2	0.5	0.5

- Is A1 better than A2?
- Being in State 1 is more beneficial than being in State 2.



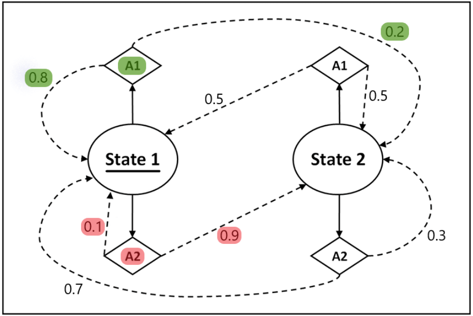
Transition probabilities

Mathematical Setting II

- $\mathcal{S} = \{\text{State 1, State 2}\}$.
- $\mathcal{A} = \{A1, A2\}$.
- Initial state = State 1.
- Binary rewards = $\{0, 1\}$.
- Mean rewards $\bar{r}(s, a)$

	A1	A2
State 1	1	1
State 2	0.5	0.5

- Is A1 better than A2?
- Being in State 1 is more beneficial than being in State 2.



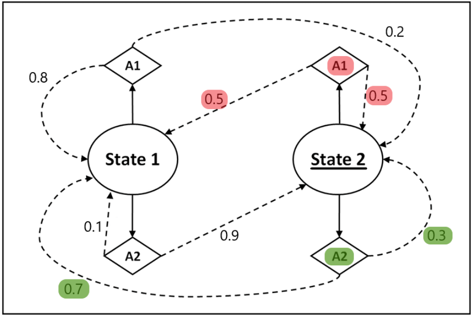
Transition probabilities

Mathematical Setting II

- $\mathcal{S} = \{\text{State 1, State 2}\}$.
- $\mathcal{A} = \{A1, A2\}$.
- Initial state = State 1.
- Binary rewards = $\{0, 1\}$.
- Mean rewards $\bar{r}(s, a)$

	A1	A2
State 1	1	1
State 2	0.5	0.5

- Is A1 better than A2?
- Being in State 1 is more beneficial than being in State 2.



Transition probabilities

Diameter

Definition (Diameter)

The *diameter* D of an MDP is the maximal expected time it takes to reach any state from any other state (using an appropriate policy).

Diameter

Definition (Diameter)

The *diameter* D of an MDP is the maximal expected time it takes to reach any state from any other state (using an appropriate policy).

Note: Typically we consider *communicating* MDPs i.e. with finite D.

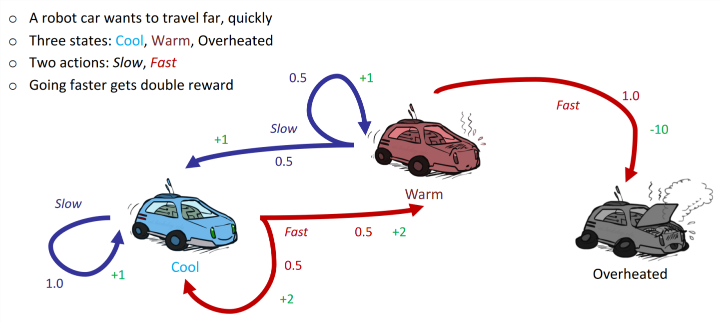
Diameter

Definition (Diameter)

The *diameter* D of an MDP is the maximal expected time it takes to reach any state from any other state (using an appropriate policy).

Note: Typically we consider *communicating* MDPs i.e. with finite D.

Does this MDP have a finite diameter?



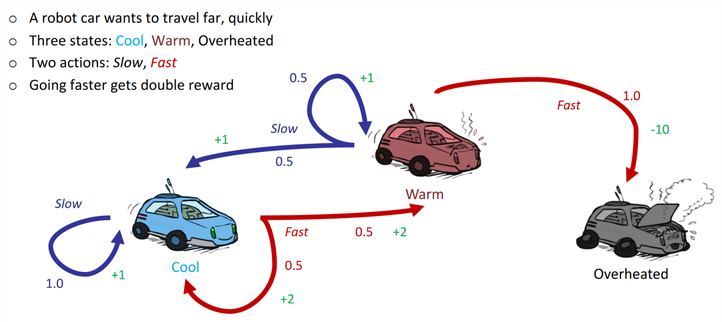
Diameter

Definition (Diameter)

The *diameter* D of an MDP is the maximal expected time it takes to reach any state from any other state (using an appropriate policy).

Note: Typically we consider *communicating* MDPs i.e. with finite D.

Does this MDP have a finite diameter? No. Cannot go from Overheated to Cool or Warm!



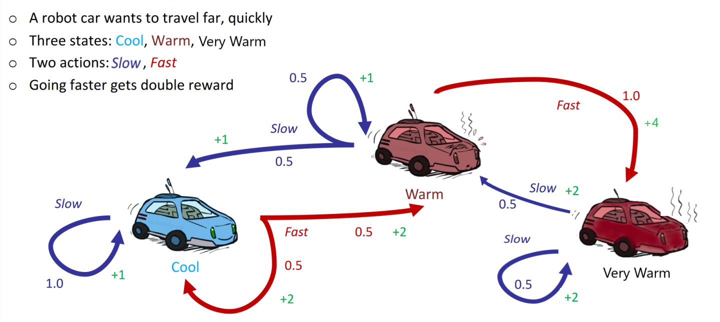
Diameter

Definition (Diameter)

The *diameter* D of an MDP is the maximal expected time it takes to reach any state from any other state (using an appropriate policy).

Note: Typically we consider *communicating* MDPs i.e. with finite D.

Does this MDP have a finite diameter?



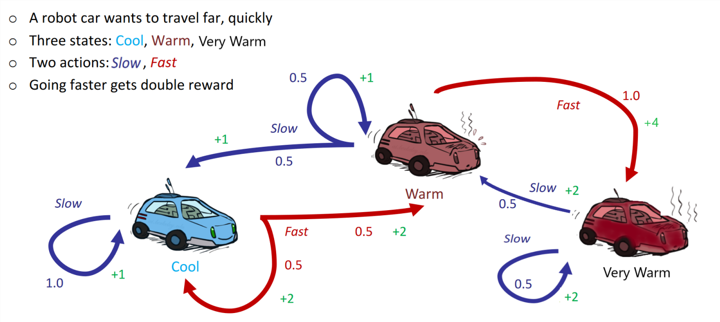
Diameter

Definition (Diameter)

The *diameter* *D* of an MDP is the maximal expected time it takes to reach any state from any other state (using an appropriate policy).

Note: Typically we consider *communicating* MDPs i.e. with finite D.

Does this MDP have a finite diameter? Yes. Diameter is 4 as the expected time from Cool to Very warm (and vice versa) is 4.



Accumulated Reward

An algorithm $\mathfrak{A}$ operating on MDP M with initial state s_0.

- At each time step t, algorithm $\mathfrak{A}$
    - is in state $s(t)$,
    - performs action $a(t)$, and
    - receives reward $r(t)$.

Accumulated Reward

An algorithm $\mathfrak{A}$ operating on MDP M with initial state s_0.

- At each time step t, algorithm $\mathfrak{A}$
    - is in state $s(t)$,
    - performs action $a(t)$, and
    - receives reward $r(t)$.
- Undiscounted accumulated reward,

$$R(M, \mathfrak{A}, s_0, T) = \sum_{t=1}^{T} r(t).$$

Accumulated Reward

An algorithm $\mathfrak{A}$ operating on MDP M with initial state s_0.

- At each time step t, algorithm $\mathfrak{A}$
  - is in state $s(t)$,
  - performs action $a(t)$, and
  - receives reward $r(t)$.
- Undiscounted accumulated reward,

$$R(M, \mathfrak{A}, s_0, T) = \sum_{t=1}^{T} r(t).$$

- Discounted accumulated reward with $0 < \gamma < 1$,

$$R(M, \mathfrak{A}, s_0, T) = \sum_{t=1}^{T} \gamma^{t-1} r(t).$$

Accumulated Reward

An algorithm $\mathfrak{A}$ operating on MDP M with initial state s_0.

- At each time step t, algorithm $\mathfrak{A}$
  - is in state $s(t)$,
  - performs action $a(t)$, and
  - receives reward $r(t)$.
- Undiscounted accumulated reward,

$$R(M, \mathfrak{A}, s_0, T) = \sum_{t=1}^{T} r(t).$$

- Discounted accumulated reward with $0 < \gamma < 1$,

$$R(M, \mathfrak{A}, s_0, T) = \sum_{t=1}^{T} \gamma^{t-1} r(t).$$

- In this lecture, we consider undiscounted accumulated reward.

Policy

- The learner uses a policy to choose actions.

Policy

- The learner uses a policy to choose actions.
- Policies can be stationary or non-stationary.

Definition (Stationary Policy)

A stationary policy is a mapping from $\pi : \mathcal{S} \to \mathcal{A}$.

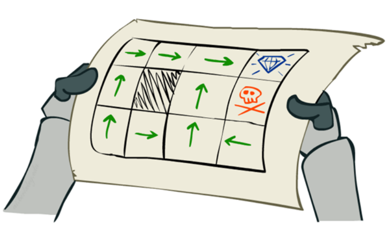


Image source: *UC Berkeley AI course, lecture 11*

Average Reward

- The average reward of policy π starting in state s_0 in MDP M is

$$\rho(M, \pi, s_0) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right]$$

Average Reward

- The average reward of policy π starting in state s_0 in MDP M is

$$\rho(M, \pi, s_0) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right]$$

- The optimal average reward ρ^* is

$$\rho^*(M) = \rho^*(M, s_0) := \max_{\pi} \rho(M, \pi, s_0)$$

- Why the first $=$ in the above? For MDPs with finite diameter, ρ^* does not depend on the initial state [Puterman, 1994, Section 8.3.3].

Average Reward

- The average reward of policy π starting in state s_0 in MDP M is

$$\rho(M, \pi, s_0) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right]$$

- The optimal average reward ρ^* is

$$\rho^*(M) = \rho^*(M, s_0) := \max_{\pi} \rho(M, \pi, s_0)$$

- Why the first $=$ in the above? For MDPs with finite diameter, ρ^* does not depend on the initial state [Puterman, 1994, Section 8.3.3].
- Optimal policy $\pi^* :=$ a policy that gives optimal average reward ρ^*.

Average Reward

- The average reward of policy π starting in state s_0 in MDP M is

$$\rho(M, \pi, s_0) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right]$$

- The optimal average reward ρ^* is

$$\rho^*(M) = \rho^*(M, s_0) := \max_{\pi} \rho(M, \pi, s_0)$$

- Why the first $=$ in the above? For MDPs with finite diameter, ρ^* does not depend on the initial state [Puterman, 1994, Section 8.3.3].
- Optimal policy $\pi^* :=$ a policy that gives optimal average reward ρ^*.
- When $\mathcal{S}$ and $\mathcal{A}$ are finite, the rewards are bounded and D is finite, it is sufficient to consider stationary policies as ρ^* can be achieved by a stationary policy [Puterman, 1994].

Value Iteration

How to compute an optimal policy π^* (for example):

Value iteration

- Set $v_0(s) := 0$ for all states $s \in \mathcal{S}$.
- For $n = 1, 2, \ldots$ and all $s \in \mathcal{S}$, set the iterated state values to be

$$v_{n+1}(s) \quad := \quad \max_{a \in A} \left\{ \bar{r}(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) v_n(s') \right\}.$$

Value Iteration

How to compute an optimal policy π^* (for example):

Value iteration

- Set $v_0(s) := 0$ for all states $s \in \mathcal{S}$.
- For $n = 1, 2, \ldots$ and all $s \in \mathcal{S}$, set the iterated state values to be

$$v_{n+1}(s) \quad := \quad \max_{a \in A} \left\{ \bar{r}(s,a) \; + \sum_{s' \in \mathcal{S}} p(s'|s,a) v_n(s') \right\}.$$

Value Iteration

How to compute an optimal policy π^* (for example):

Value iteration

- Set $v_0(s) := 0$ for all states $s \in \mathcal{S}$.
- For $n = 1, 2, \ldots$ and all $s \in \mathcal{S}$, set the iterated state values to be

$$v_{n+1}(s) \quad := \quad \max_{a \in A} \left\{ \bar{r}(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) v_n(s') \right\}.$$

Convergence (under certain conditions)

For $n \to \infty$, the $\arg\max$-actions converge to an optimal policy π^*.

Value Iteration

How to compute an optimal policy π^* (for example):

Value iteration

- Set $v_0(s) := 0$ for all states $s \in \mathcal{S}$.
- For $n = 1, 2, \ldots$ and all $s \in \mathcal{S}$, set the iterated state values to be

$$v_{n+1}(s) \quad := \quad \max_{a \in A} \left\{ \bar{r}(s, a) \; + \sum_{s' \in \mathcal{S}} p(s' | s, a) v_n(s') \right\}.$$

Convergence (under certain conditions)

For $n \to \infty$, the $\arg\max$-actions converge to an optimal policy π^*.

Another stopping criterion: Stop the value iteration when the maximum difference between two successive v's $\leq$ some threshold.
Then, the arg-max action policy is near-optimal.

For further information about value iteration and other stopping criteria, click here.

Performance Measure : Regret

- How do we define regret usually?
  Regret = optimal cumulative reward - learner's reward.

Performance Measure : Regret

- How do we define regret usually?
  Regret = optimal cumulative reward - learner's reward.

Definition (Regret)

The *regret* of an algorithm $\mathfrak{A}$ in MDP M with initial state s_0 after T steps is

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) := T\rho^* - R(M, \mathfrak{A}, s_o, T) = T\rho^* - \sum_{t=1}^{T} r(t),$$

where $r(t)$ is the random reward the algorithm receives at step t.

Note: $T\rho^*$ is a good proxy for the optimal T-step reward [Jaksch et al., 2010, Page 3].

$$\max_{\mathfrak{A}} \mathbb{E}[R(M, \mathfrak{A}, s_o, T)] = T\rho^* + O(D).$$

Lower Bound on Regret

Theorem (Jaksch et al. [2010])

For any algorithm and any natural numbers T, S, $A > 1$, and $D \geq \log_A S$, there is an MDP M with S states, A actions, and diameter D, such that for any initial state s the expected regret after T steps is of the order

$$\sqrt{DSAT}.$$

An Algorithm with Near-optimal Regret

Optimism Principle

"The learner should act as if it is in the **best plausible** world."

Optimism in MDPs: Estimates

- **For bandits:**
  Estimates $\hat{\mu}_a$ for mean reward of each arm a

$$\hat{\mu}_a := \frac{\text{sum of received rewards when playing arm } a}{\text{number of times arm } a \text{ was played}}.$$

Optimism in MDPs: Estimates

- **For bandits:**
  Estimates $\hat{\mu}_a$ for mean reward of each arm a

$$\hat{\mu}_a := \frac{\text{sum of received rewards when playing arm } a}{\text{number of times arm } a \text{ was played}}.$$

- **For MDPs:**
  Estimates for mean rewards and transition probabilities

$$\hat{r}(s,a) := \frac{\text{sum of received rewards when playing } a \text{ in } s}{\text{number of times } a \text{ was played in } s},$$

$$\hat{p}(s'|s,a) := \frac{\text{number of transitions to } s' \text{ when playing } a \text{ in } s}{\text{number of times } a \text{ was played in } s}.$$

Optimism in MDPs: Confidence Intervals

- **For bandits:**
  confidence intervals for reward of each arm

Optimism in MDPs: Confidence Intervals

- **For bandits:**
  confidence intervals for reward of each arm
- **For MDPs:**
  confidence intervals for rewards and transition probabilities

Set of plausible MDPs

The set $\mathbb{M}$ of plausible MDPs given the estimates $\hat{r}$ and $\hat{p}$ is the set of all MDPs with rewards $\tilde{r}$ and transition probabilities $\tilde{p}$ such that

$$\begin{aligned} \left|\hat{r}(s,a) - \tilde{r}(s,a)\right| &\leq \mathrm{conf}_r(s,a), \\ \left\|\hat{p}(\cdot|s,a) - \tilde{p}(\cdot|s,a)\right\|_1 &\leq \mathrm{conf}_p(s,a). \end{aligned}$$

where $\|\mathbf{x}\|_1 = \sum |x_i|$.

Optimism in MDPs : Policy

- **For bandits :**
  Choose arm with the highest upper confidence bound.

Optimism in MDPs : Policy

- **For bandits :**
  Choose arm with the highest upper confidence bound.
- **For MDPs :**
  Choose an optimistic MDP $\tilde{\mathcal{M}} \in \mathbb{M}$ that promises highest average reward under an optimal policy $\tilde{\pi}$,
  where $\mathbb{M}$ is the set of plausible MDPs built using confidence intervals.

$\rightsquigarrow$ Choose optimistic MDP $\tilde{\mathcal{M}} \in \mathbb{M}$ and optimal policy $\tilde{\pi}$ such that

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}} \rho(\mathcal{M}, \pi).$$

The Optimal Policy in the Optimistic MDP

$\rightsquigarrow$ Choose optimistic MDP $\tilde{\mathcal{M}} \in \mathbb{M}$ and optimal policy $\tilde{\pi}$ such that

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

The Optimal Policy in the Optimistic MDP

$\rightsquigarrow$ Choose optimistic MDP $\tilde{\mathcal{M}} \in \mathbb{M}$ and optimal policy $\tilde{\pi}$ such that

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

- Use an extension of value iteration.
  That is, for all states s, set $u_0(s) := 0$ and

$$u_{i+1}(s) \quad := \quad \max_a \left\{ \hat{r}(s,a) + \mathrm{conf}_r(s,a) \; + \max_{p \in \mathcal{P}(s,a)} \left\{ \sum_{s'} p(s') u_i(s') \right\} \right\},$$

where $\mathcal{P}(s,a)$ is the set of all plausible transition probabilities.

The Optimal Policy in the Optimistic MDP

$\rightsquigarrow$ Choose optimistic MDP $\tilde{\mathcal{M}} \in \mathbb{M}$ and optimal policy $\tilde{\pi}$ such that

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

- Use an extension of value iteration.
  That is, for all states s, set $u_0(s) := 0$ and

$$u_{i+1}(s) \quad := \quad \max_a \left\{ \hat{r}(s,a) + \mathrm{conf}_r(s,a) \; + \max_{p \in \mathcal{P}(s,a)} \left\{ \sum_{s'} p(s') u_i(s') \right\} \right\},$$

  where $\mathcal{P}(s,a)$ is the set of all plausible transition probabilities.
- $\max \mathbf{p} \cdot \mathbf{u}_i$ is a linear optimization problem over the convex polytope $\mathcal{P}(s,a)$. So it can be evaluated considering only the finite number of vertices of this polytope.

  (For further information, see [Jaksch et al., 2010, Section 3.1].)

Break

We start again after a break.

Recap

Before the break, we saw

- Introduction to Markov decision processes, definitions of average reward, diameter and regret.
- Lower bound for regret of the order $\sqrt{DSAT}$.
- Optimism principle in MDPs : use confidence intervals for rewards and transition probabilities to build a set of plausible MDPs and then choose an optimal policy in the optimistic MDP.
- How to compute the optimal policy in the optimistic MDP using extended value iteration.

Recap

Before the break, we saw

- Introduction to Markov decision processes, definitions of average reward, diameter and regret.
- Lower bound for regret of the order $\sqrt{DSAT}$.
- Optimism principle in MDPs : use confidence intervals for rewards and transition probabilities to build a set of plausible MDPs and then choose an optimal policy in the optimistic MDP.
- How to compute the optimal policy in the optimistic MDP using extended value iteration.
- Next, we shall see the algorithm $\mathrm{UCRL2}$ and its regret bound.

Algorithm : UCRL2

- Runs in episodes i.e., a series of time steps – these are used by the algorithm internally.

Algorithm UCRL2 [Jaksch et al., 2010]

---

for episode $k = 1, 2, \ldots$ **do**

    Compute the estimates for rewards and transition probabilities.

    Build the set $\mathbb{M}_k$ of plausible MDPs based on current estimates.

    Find an optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

    using extended value iteration.

    Execute $\tilde{\pi}_k$ until

end for

Algorithm : UCRL2

- Runs in episodes i.e., a series of time steps – these are used by the algorithm internally.

Algorithm UCRL2 [Jaksch et al., 2010]

for episode $k = 1, 2, \ldots$ **do**

Compute the estimates for rewards and transition probabilities.

Build the set $\mathbb{M}_k$ of plausible MDPs based on current estimates.

Find an optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

using extended value iteration.

Execute $\tilde{\pi}_k$ until

end for

Algorithm : UCRL2

- Runs in episodes i.e., a series of time steps – these are used by the algorithm internally.

Algorithm $\mathrm{UCRL2}$ [Jaksch et al., 2010]

---

for episode $k = 1, 2, \ldots$ **do**

  Compute the estimates for rewards and transition probabilities.

  Build the set $\mathbb{M}_k$ of plausible MDPs based on current estimates.

  Find an optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

  using extended value iteration.

  Execute $\tilde{\pi}_k$ until

end for

Algorithm : UCRL2

- Runs in episodes i.e., a series of time steps – these are used by the algorithm internally.

Algorithm UCRL2 [Jaksch et al., 2010]

for episode $k = 1, 2, \ldots$ **do**
  Compute the estimates for rewards and transition probabilities.
  Build the set $\mathbb{M}_k$ of plausible MDPs based on current estimates.
  Find an optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

  using extended value iteration.
  Execute $\tilde{\pi}_k$ until
end for

Algorithm : UCRL2

- Runs in episodes i.e., a series of time steps – these are used by the algorithm internally.

Algorithm UCRL2 [Jaksch et al., 2010]

---

for episode $k = 1, 2, \ldots$ **do**

  Compute the estimates for rewards and transition probabilities.

  Build the set $\mathbb{M}_k$ of plausible MDPs based on current estimates.

  Find an optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

  using extended value iteration.

  Execute $\tilde{\pi}_k$ until episode stopping criterion is satisfied.

end for

Algorithm : UCRL2

- Runs in episodes i.e., a series of time steps – these are used by the algorithm internally.

Algorithm UCRL2 [Jaksch et al., 2010]

```
for episode k = 1, 2, ... do
    Compute the estimates for rewards and transition probabilities.
    Build the set 𝕄_k of plausible MDPs based on current estimates.
    Find an optimal policy π̃_k in the optimistic MDP M̃ which satisfies
        ρ(M̃, π̃_k) = max_{π, M ∈ 𝕄_k} ρ(M, π).
    using extended value iteration.
    Execute π̃_k until the visits in some state-action pair have doubled .
end for
```

$$N_k(s,a) := \text{ visits to state action pair } (s,a) \text{ prior to episode } k.$$

$$v_k(s,a) := \text{ visits to state action pair } (s,a) \text{ in episode } k.$$

Episode stopping criterion : $v_k(s,a) = \max\{1, N_k(s,a)\}$ for some (s,a).

Regret Bound for UCRL2

Theorem (Jaksch et al. [2010])

In an MDP with S states, A actions, and diameter D, with probability of at least $1-\delta$ the regret of $\mathrm{UCRL2}$ after T steps is bounded by

$$34 \cdot DS\sqrt{AT \log\left(\frac{T}{\delta}\right)}.$$

Regret Bound for UCRL2

Theorem (Jaksch et al. [2010])

In an MDP with S states, A actions, and diameter D, with probability of at least $1 - \delta$ the regret of $\mathrm{UCRL2}$ after T steps is bounded by

$$34 \cdot DS\sqrt{AT \log\left(\frac{T}{\delta}\right)}.$$

- Gap of $\sqrt{DS}$ between the lower bound (i.e., $\sqrt{DSAT}$) and $\mathrm{UCRL2}$ upper bound. So $\mathrm{UCRL2}$ is near-optimal. ☺

Proving the Regret Bound

Proving the Regret Bound for UCB : Roadmap

- Reduce regret to the sum of *per episode-regret*.
- Bound the number of episodes.
- Bound per-episode regret.

Reduction to Per-Episode Regret

- Let us define regret in episode k to be

$$\Delta_k := \sum_{s,a} v_k(s,a)(\rho^* - \bar{r}(s,a)),$$

where $v_k(s,a) :=$ the number of times a was played in s in episode k.

Reduction to Per-Episode Regret

- Let us define regret in episode k to be

$$\Delta_k := \sum_{s,a} v_k(s,a)(\rho^* - \bar{r}(s,a)),$$

where $v_k(s,a) :=$ the number of times a was played in s in episode k.

- Then, the regret can be bounded as,

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_k \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)},$$

with high probability (see [Jaksch et al., 2010, Section 4.1]).

Reduction to Per-Episode Regret

- Let us define regret in episode k to be

$$\Delta_k := \sum_{s,a} v_k(s,a)(\rho^* - \bar{r}(s,a)),$$

where $v_k(s,a) :=$ the number of times a was played in s in episode k.

- Then, the regret can be bounded as,

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_k \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)},$$

with high probability (see [Jaksch et al., 2010, Section 4.1]).

- So to bound regret, we need to bound $\sum_k \Delta_k$.

Bound on the Number of Episodes

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_k \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

Bound on the Number of Episodes

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_k \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Algorithm UCRL2 [Jaksch et al., 2010]

1: **for** episode $k = 1, 2, \ldots$ **do**
2: Compute the estimates for rewards and transition probabilities.
3: Build the set $\mathbb{M}_k$ of plausible MDPs based on current estimates.
4: Find the optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

using extended value iteration.
5: Execute $\tilde{\pi}_k$ until the visits in some state-action pair have doubled.
6: **end for**

Bound on the Number of Episodes

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_k \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Algorithm UCRL2 [Jaksch et al., 2010]

for episode $k = 1, 2, \ldots$ **do**
    Compute the estimates for rewards and transition probabilities.
    Build the set $\mathbb{M}_k$ of plausible MDPs based on current estimates.
    Find the optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

    using extended value iteration.
    Execute $\tilde{\pi}_k$ until the visits in some state-action pair have doubled.
end for

Due to the episode stopping criterion, the number of episodes of UCRL2 up to step T are upper bounded as

$$m \leq O\left(SA \log_2\left(\frac{8T}{SA}\right)\right),$$

where $S = |states|$ and $A = |actions|$ [Jaksch et al., 2010, Appendix C.2].

Decomposing the Sum of Per-Episode Regret

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k=1}^{m} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Decomposing the Sum of Per-Episode Regret

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k=1}^{m} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

Algorithm UCRL2 [Jaksch et al., 2010]

1: **for** episode $k = 1, 2, \ldots$ **do**
2: Compute the estimates for rewards and transition probabilities.
3: Build the set $\mathbb{M}_k$ of plausible MDPs.
4: Find the optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

using extended value iteration.
5: Execute $\tilde{\pi}_k$ until the visits in some state-action pair have doubled.
6: **end for**

Set of plausible MDPs

The set $\mathbb{M}$ of plausible MDPs given the estimates $\hat{r}$ and $\hat{p}$ is the set of all MDPs with rewards $\tilde{r}$ and transition probabilities $\tilde{p}$ such that

$$\begin{aligned} \left|\hat{r}(s,a) - \tilde{r}(s,a)\right| &\leq \mathrm{conf}_r(s,a), \\ \left\|\hat{p}(\cdot|s,a) - \tilde{p}(\cdot|s,a)\right\|_1 &\leq \mathrm{conf}_p(s,a). \end{aligned}$$

Decomposing the Sum of Per-Episode Regret

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k=1}^{m} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Algorithm UCRL2 [Jaksch et al., 2010]

1: **for** episode $k = 1, 2, \ldots$ **do**
2: Compute the estimates for rewards and transition probabilities.
3: Build the set $\mathbb{M}_k$ of plausible MDPs.
4: Find the optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

using extended value iteration.
5: Execute $\tilde{\pi}_k$ until the visits in some state-action pair have doubled.
6: **end for**

Set of plausible MDPs

The set $\mathbb{M}$ of plausible MDPs given the estimates $\hat{r}$ and $\hat{p}$ is the set of all MDPs with rewards $\tilde{r}$ and transition probabilities $\tilde{p}$ such that

$$\begin{aligned} \left|\hat{r}(s,a) - \tilde{r}(s,a)\right| &\leq \mathrm{conf}_r(s,a), \\ \left\|\hat{p}(\cdot|s,a) - \tilde{p}(\cdot|s,a)\right\|_1 &\leq \mathrm{conf}_p(s,a). \end{aligned}$$

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Decomposing the Sum of Per-Episode Regret

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k=1}^{m} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

Algorithm UCRL2 [Jaksch et al., 2010]

1: **for** episode $k = 1, 2, \ldots$ **do**
2: Compute the estimates for rewards and transition probabilities.
3: Build the set $\mathbb{M}_k$ of plausible MDPs.
4: Find the optimal policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{\mathcal{M}}$ which satisfies

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}_k) = \max_{\pi, \mathcal{M} \in \mathbb{M}_k} \rho(\mathcal{M}, \pi).$$

using extended value iteration.
5: Execute $\tilde{\pi}_k$ until the visits in some state-action pair have doubled.
6: **end for**

Set of plausible MDPs

The set $\mathbb{M}$ of plausible MDPs given the estimates $\hat{r}$ and $\hat{p}$ is the set of all MDPs with rewards $\tilde{r}$ and transition probabilities $\tilde{p}$ such that

$$\begin{aligned} \left|\hat{r}(s,a) - \tilde{r}(s,a)\right| &\leq \mathrm{conf}_r(s,a), \\ \left\|\hat{p}(\cdot|s,a) - \tilde{p}(\cdot|s,a)\right\|_1 &\leq \mathrm{conf}_p(s,a). \end{aligned}$$

$$\begin{aligned} \mathfrak{R}(M, \mathfrak{A}, s_0, T) &\leq \sum_{k} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} \\ &= \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} \end{aligned}$$

Dealing with Failing Confidence Regions

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Dealing with Failing Confidence Regions

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

- We need to bound $\mathbb{P}\{M \notin \mathbb{M}(t)\}$ i.e. the probability of mean rewards and transition probabilities in the true MDP M deviating far from their respective estimates. How do we do that?

Dealing with Failing Confidence Regions

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

- We need to bound $\mathbb{P}\{M \notin \mathbb{M}(t)\}$ i.e. the probability of mean rewards and transition probabilities in the true MDP M deviating far from their respective estimates. How do we do that?
- The confidence intervals $\mathrm{conf}_r(s, a)$ and $\mathrm{conf}_p(s, a)$ for the set $\mathbb{M}$ of plausible MDPs are chosen such that

$$\mathbb{P}\{M \notin \mathbb{M}(t)\} \leq \frac{\delta}{15t^6}.$$

where $\mathbb{M}(t) :=$ set of plausible MDPs using the estimates at time t.

Dealing with Failing Confidence Regions

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

- We need to bound $\mathbb{P}\{M \notin \mathbb{M}(t)\}$ i.e. the probability of mean rewards and transition probabilities in the true MDP M deviating far from their respective estimates. How do we do that?
- The confidence intervals $\mathrm{conf}_r(s, a)$ and $\mathrm{conf}_p(s, a)$ for the set $\mathbb{M}$ of plausible MDPs are chosen such that

$$\mathbb{P}\{M \notin \mathbb{M}(t)\} \leq \frac{\delta}{15t^6}.$$

  where $\mathbb{M}(t) :=$ set of plausible MDPs using the estimates at time t.
- Then, it can be shown with high probability,

$$\sum_{k, M \notin \mathbb{M}_k} \Delta_k \leq \sqrt{T}.$$

(For more details, see [Jaksch et al., 2010, Section 4.2])

Isolating the Dominating Term

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Recall $\tilde{\rho}_k$ = Average reward of policy $\tilde{\pi}_k$ and $\tilde{r}$ = reward function of a plausible MDP.

$$\sum_{k, M \in \mathbb{M}_k} \Delta_k = \sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\rho^* - \bar{r}(s,a))$$

Isolating the Dominating Term

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Recall $\tilde{\rho}_k$ = Average reward of policy $\tilde{\pi}_k$ and $\tilde{r}$ = reward function of a plausible MDP.

$$\sum_{k, M \in \mathbb{M}_k} \Delta_k = \sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s, a)(\rho^* - \bar{r}(s, a))$$

Isolating the Dominating Term

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

Recall $\tilde{\rho}_k$ = Average reward of policy $\tilde{\pi}_k$ and $\tilde{r}$ = reward function of a plausible MDP.

$$\begin{aligned}\sum_{k, M \in \mathbb{M}_k} \Delta_k &= \sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\rho^* - \bar{r}(s,a)) \\ &= \underbrace{\sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\tilde{\rho}_k - \tilde{r}(s,a))}_{\text{Dominating term}} + \dots\end{aligned}$$

Isolating the Dominating Term

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

Recall $\tilde{\rho}_k$ = Average reward of policy $\tilde{\pi}_k$ and $\tilde{r}$ = reward function of a plausible MDP.

$$\begin{aligned}\sum_{k, M \in \mathbb{M}_k} \Delta_k &= \sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\rho^* - \bar{r}(s,a)) \\ &= \underbrace{\sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\tilde{\rho}_k - \tilde{r}(s,a))}_{\text{Dominating term}} \\ &\quad + \underbrace{\sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\tilde{r}(s,a) - \bar{r}(s,a))}_{O(S\sqrt{AT \log(T/\delta)})} + \dots\end{aligned}$$

Isolating the Dominating Term

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

Recall $\tilde{\rho}_k$ = Average reward of policy $\tilde{\pi}_k$ and $\tilde{r}$ = reward function of a plausible MDP.

$$\begin{aligned}
\sum_{k, M \in \mathbb{M}_k} \Delta_k &= \sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\rho^* - \bar{r}(s,a)) \\
&= \underbrace{\sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\tilde{\rho}_k - \tilde{r}(s,a))}_{\text{Dominating term}} \\
&\quad + \underbrace{\sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\tilde{r}(s,a) - \bar{r}(s,a))}_{O(S\sqrt{AT \log(T/\delta)})} \\
&\quad + \underbrace{\sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\rho^* - \tilde{\rho}_k)}_{O(\sqrt{SAT})}
\end{aligned}$$

Bounding the Dominating Term

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Bounding the Dominating Term

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

With high probability,

$$\underbrace{\sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\tilde{\rho}_k - \tilde{r}(s,a))}_{\text{Dominating term}} \leq O(DS\sqrt{AT \log(T/\delta)})$$

Bounding the Dominating Term

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$

With high probability,

$$\underbrace{\sum_{k, M \in \mathbb{M}_k} \sum_{s,a} v_k(s,a)(\tilde{\rho}_k - \tilde{r}(s,a))}_{\text{Dominating term}} \leq O(DS\sqrt{AT \log(T/\delta)})$$

Therefore, with high probability,

$$\begin{aligned}\sum_{k, M \in \mathbb{M}_k} \Delta_k &\leq O(DS\sqrt{AT \log(T/\delta)}) + O(S\sqrt{AT \log(T/\delta)}) + O(\sqrt{SAT}) \\ &\leq O(DS\sqrt{AT \log(T/\delta)}).\end{aligned}$$

(For more details, see [Jaksch et al., 2010, Section 4.3])

Putting Everything Together

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Putting Everything Together

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq O(\sqrt{T}) + \cdots + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Putting Everything Together

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq O(\sqrt{T}) + O(DS\sqrt{AT\log(T/\delta)}) + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

Putting Everything Together

$$\mathfrak{R}(M, \mathfrak{A}, s_0, T) \leq \sum_{k, M \notin \mathbb{M}_k} \Delta_k + \sum_{k, M \in \mathbb{M}_k} \Delta_k + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)}$$

$$\begin{aligned}\mathfrak{R}(M, \mathfrak{A}, s_0, T) &\leq O(\sqrt{T}) + O(DS\sqrt{AT \log(T/\delta)}) + \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)} \\ &\leq 34DS\sqrt{AT \log(T/\delta)} \quad \square\end{aligned}$$

Summary

- Markov decision processes.
- Mathematical setting and a lower bound on regret.
- UCRL2.
- Sketch of the regret analysis.

Recall the Objectives from Lecture 1

- To gain an understanding of various reinforcement learning problems and formulate them mathematically. ✓
- To devise solution strategies for these problems. ✓
- To prove performance guarantees for these solutions. ✓

Recall the Objectives from Lecture 1

- To gain an understanding of various reinforcement learning problems and formulate them mathematically. ✓
- To devise solution strategies for these problems. ✓
- To prove performance guarantees for these solutions. ✓

Recall the Objectives from Lecture 1

- To gain an understanding of various reinforcement learning problems and formulate them mathematically. ✓
- To devise solution strategies for these problems. ✓
- To prove performance guarantees for these solutions. ✓

About Research Project Phase I

- Research project should be of mathematical nature.
- Basic criteria :
  - novelty in the proved results, and/or
  - novelty in the proof techniques.

About Research Project Phase II

- From week 4 (Sep 26-30) to week 8 (Oct 24-28), each group is entitled to a single half-hour meeting.
- On Mondays, at Metaforum 09,
    - Time-slot 1 : 14:15 - 14:45
    - Time-slot 2 : 14:50 - 15:20
    - Time-slot 3 : 15:25 - 15:55
    - Time-slot 4 : 16:00 - 16:30

  (Except on Oct 10. On Oct 10, the above time-slots shifted to 4 hours earlier i.e., Time-slot 1 from 10:15, Time-slot 2 from 10:50 and Time-slot 3 from 11:25 and Time-slot 4 from 12:00.)
- On Wednesdays, at Matrix 1.122,
    - Time-slot 5 : 11:00 - 11:30
    - Time-slot 6 : 11:35 - 12:05
    - Time-slot 7 : 12:10 - 12:40

About Research Project Phase II

- From week 4 (Sep 26-30) to week 8 (Oct 24-28), each group is entitled to a single half-hour meeting.
- On Mondays, at Metaforum 09,
  - Time-slot 1 : 14:15 - 14:45
  - Time-slot 2 : 14:50 - 15:20
  - Time-slot 3 : 15:25 - 15:55
  - Time-slot 4 : 16:00 - 16:30

  (Except on Oct 10. On Oct 10, the above time-slots shifted to 4 hours earlier i.e., Time-slot 1 from 10:15, Time-slot 2 from 10:50 and Time-slot 3 from 11:25 and Time-slot 4 from 12:00.)
- On Wednesdays, at Matrix 1.122,
  - Time-slot 5 : 11:00 - 11:30
  - Time-slot 6 : 11:35 - 12:05
  - Time-slot 7 : 12:10 - 12:40

In case of any change in the above schedule, I will inform the concerned groups in advance and we will come up with another time-slot.

References

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. URL `http://jmlr.org/papers/v11/jaksch10a.html`.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley amp; Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.