CMPT 419/983: Theoretical Foundations of Reinforcement Learning

Lecture 3

Sharan Vaswani

September 22, 2023

Recap

- **Stochastic Linear Bandits**: For arm $a \in [K]$, $\mu_a = \langle X_a, \theta^* \rangle$.
- On pulling arm a, we observe reward $R_t = \mu_{a_t} + \eta_t$, $\mathbb{E}[\eta_t] = 0$ and η_t is conditional 1 sub-Gaussian, i.e. for $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda \eta_t) | \mathcal{H}_{t-1}] \leq \exp(\lambda^2/2)$.

Algorithm Linear Upper Confidence Bound

```
**Input**: $\{\beta_t\}_{t=1}^{T}$, $V_0 = \lambda I_d \in \mathbb{R}^{d \times d}$
For each arm $a \in [K]$, initialize $U_a(0, \delta) := \infty$.
**for** $t = 1 \to T$ **do**
    Select arm $a_t = \arg\max_{a \in [K]} U_a(t-1, \delta) = \arg\max_{a \in \mathcal{A}} U_a(t-1, \delta)$
    Observe reward $R_t$ and update:
        $V_t = V_{t-1} + X_t X_t^T \quad ; \quad b_t = b_{t-1} + R_t X_t \quad ; \quad \hat{\theta}_t = V_t^{-1} b_t$
        $U_a(t) = \langle X_a, \hat{\theta}_t \rangle + \sqrt{\beta_t} \, \|X_a\|_{V_t^{-1}} = \max_{\theta \in \mathcal{C}_t} \langle \theta, X_a \rangle \quad (\text{where } \mathcal{C}_t = \left\{ \theta \,|\, \left\| \theta - \hat{\theta}_t \right\|_{V_t}^2 \leq \beta_t \right\})$
**end for**
```

Recap

Claim: Assuming (i) $\|\theta^*\| \leq 1$, (ii) $\|X_a\| \leq 1$ for all a and (iii) $R_t \in [0,1]$, UCB with $\sqrt{\beta_t} = \sqrt{d \log\left(\frac{\lambda d + t}{\lambda\, d}\right) + 2\log(1/\delta)} + \sqrt{\lambda}$ achieves the following worst-case bound on the regret,

$$\text{Regret}(\text{LinUCB}, T) \leq O\left(d\sqrt{T}\log(T)\right)$$

Last time we showed the following results: if
$G := \{\forall t \in [T] | \theta^* \in \mathcal{C}_t := \left\{\theta \mid \left\|\theta - \hat{\theta}_t\right\|^2_{V_t} \leq \beta_t\right\}$,

$$\textbf{(1):}\ \text{Regret}(\text{LinUCB}, T) \leq 2\sqrt{T\,\beta_T\,\mathbb{E}\left[\sum_{t=1}^{T} \|X_t\|^2_{V_t^{-1}} | G\right]} + T\ \Pr[G^c]$$

$$\textbf{(2):}\ \sum_{t=1}^{T} \|X_t\|^2_{V_t^{-1}} \leq 2d\ \log\left(\frac{\lambda d + T}{\lambda d}\right)$$

Today, we will prove: **(3)**: For $\sqrt{\beta_t} = \sqrt{d \log\left(\frac{\lambda d + t}{\lambda d}\right) + 2\log(T)} + \sqrt{\lambda}$, $\Pr[G^c] \leq \frac{1}{T}$, and thus finish the proof.

Digression – (Super)-Martingales

Martingale: Sequence of random variables for which, at a particular time, the conditional expectation of the next value in the sequence is equal to the present value, regardless of all prior values.

A sequence of random variables – $M_1, M_2, \dots$ is a discrete-time martingale if for all t,

$$\mathbb{E}[|M_t|] \leq \infty \quad ; \quad \mathbb{E}[M_t | M_1, M_2, \dots M_{t-1}] = M_{t-1}$$

Example 1: An unbiased random walk
Example 2: Gambler's fortune: Suppose M_t is a gambler's fortune after t tosses of a fair coin, where the gambler wins \$1 if the coin comes up heads and loses \$1 if it comes up tails.

Super-Martingale: A sequence of random variables – $M_1, M_2, \dots$ is a discrete-time super-martingale if for all t,

$$\mathbb{E}[|M_t|] \leq \infty \quad ; \quad \mathbb{E}[M_t | M_1, M_2, \dots M_{t-1}] \leq M_{t-1}$$

Linear UCB – Regret Analysis

Claim: If (i) $\|\theta^*\| \leq 1$ and (ii) $\|X_a\| \leq 1$ for all a, for $\sqrt{\beta_t} = \sqrt{d \log\left(\frac{\lambda d + t}{\lambda d}\right) + 2\log(T)} + \sqrt{\lambda}$ and $G := \{\forall t \in [T] | \theta^* \in \mathcal{C}_t := \left\{\theta \,|\, \left\|\theta - \hat{\theta}_t\right\|_{V_t}^2 \leq \beta_t\right\}$, $\Pr[G^c] \leq \frac{1}{T}$.

Proof: Define $S_t := \sum_{s=1}^{t} \eta_s X_s$ and $K_t := \sum_{s=1}^{t} X_s X_s^{\intercal}$. We will prove the claim in 4 steps:

(i) $\left\|\theta - \hat{\theta}_t\right\|_{V_t} \leq \|S_t\|_{V_t^{-1}} + \sqrt{\lambda}$.

(ii) $M_t(z) = \exp\left(\langle z, S_t \rangle - \frac{1}{2}\|z\|_{K_t}^2\right)$ is a non-negative super-martingale with $M_0(z) = 1$.

(iii) Use the fact that a mixture of super-martingales given by $\bar{M}_t = \int_z M_t(z) h(z)\, dz$ is also a non-negative super-martingale for any probability density function $h(z)$.

(iv) Use the maximal inequality for super-martingales to bound $\Pr\left[\sup_{t \in [T]} \log(\bar{M}_t(z)) \geq \log(1/\delta)\right]$ and hence bound $\left\|\theta - \hat{\theta}_t\right\|_{V_t}$.

Linear UCB – Regret Analysis

Part (i): If $S_t := \sum_{s=1}^{t} \eta_s X_s$ and $K_t := \sum_{s=1}^{t} X_s X_s^{\top}$, then $\left\|\theta^* - \hat{\theta}_t\right\|_{V_t} \leq \|S_t\|_{V_t^{-1}} + \sqrt{\lambda}$.

Proof:

$$b_t = \sum_{s=1}^{t} X_s R_s = \sum_{s=1}^{t} X_s \left[\langle X_s, \theta^* \rangle + \eta_s\right]$$

$$= \sum_{s=1}^{t} X_s^{\top} X_s \theta^* + \sum_{s=1}^{t} X_s \eta_s = S_t + \sum_{s=1}^{t} X_s^{\top} X_s \theta^*.$$

$$\implies \hat{\theta}_t = V_t^{-1} b_t = V_t^{-1} S_t + V_t^{-1} \left[\sum_{s=1}^{t} X_s^{\top} X_s\right] \theta^* = V_t^{-1} S_t + V_t^{-1} K_t \theta^*$$

$$\left\|\theta^* - \hat{\theta}_t\right\|_{V_t} = \left\|V_t^{-1} S_t + \left(V_t^{-1} K_t - I_d\right) \theta^*\right\|_{V_t} = \|S_t\|_{V_t^{-1}} + \sqrt{\theta^{*\top} \left(V_t^{-1} K_t - I_d\right) \underbrace{\left(K_t - V_t\right)}_{= -\lambda I_d} \theta^*}$$

$$= \|S_t\|_{V_t^{-1}} + \sqrt{\lambda} \sqrt{\theta^{*\top} \left(I_d - V_t{}^{-1} K_t\right) \theta^*} \qquad \left(\text{Since } \theta^{*\top} \left[V_t{}^{-1} K_t\right] \theta^* \geq 0\right)$$

$$\implies \left\|\theta^* - \hat{\theta}_t\right\|_{V_t} \leq \|S_t\|_{V_t^{-1}} + \sqrt{\lambda}\, \|\theta^*\| \leq \|S_t\|_{V_t^{-1}} + \sqrt{\lambda} \quad \square$$

Linear UCB – Regret Analysis

Part (ii): If $S_t := \sum_{s=1}^{t} \eta_s X_s$ and $K_t := \sum_{s=1}^{t} X_s X_s^{\mathsf{T}}$, $M_t(z) = \exp\left(\langle z, S_t \rangle - \frac{1}{2} \|z\|_{K_t}^2\right)$ is a non-negative super-martingale with $M_0(z) = 1$.

Proof: It is clear that $M_t(z) = \exp\left(\langle z, S_t \rangle - \frac{1}{2} \|z\|_{K_t}^2\right)$ is non-negative and $M_0(z) = 1$. By our assumption on the noise, $\mathbb{E}[\exp(\lambda \eta_t) | \mathcal{H}_{t-1}] \leq \exp\left(\frac{\lambda^2}{2}\right)$. Setting $\lambda = \langle z, X_t \rangle$, implies that $\mathbb{E}[\exp(\langle z, X_t \rangle \eta_t) | \mathcal{H}_{t-1}] \leq \exp\left(\frac{\|z\|_{X_t X_t^{\mathsf{T}}}^2}{2}\right) \implies \mathbb{E}\left[\exp\left(\langle z, X_t \rangle \eta_t - \frac{\|z\|_{X_t X_t^{\mathsf{T}}}^2}{2}\right) | \mathcal{H}_{t-1}\right] \leq 1$ (*).

$$\mathbb{E}[M_t(z) | \mathcal{H}_{t-1}] = \mathbb{E}\left[\exp\left(\langle z, S_{t-1} + \eta_t X_t \rangle - \frac{1}{2} \|z\|_{K_{t-1} + X_t X_t^{\mathsf{T}}}^2\right) | \mathcal{H}_{t-1}\right]$$

$$= \mathbb{E}\left[\exp\left(\langle z, \eta_t X_t \rangle - \frac{1}{2} \|z\|_{X_t X_t^{\mathsf{T}}}^2\right) | \mathcal{H}_{t-1}\right] \mathbb{E}\left[\exp\left(\langle z, S_{t-1} \rangle - \frac{1}{2} \|z\|_{K_{t-1}}^2\right) | \mathcal{H}_{t-1}\right]$$

$$= M_{t-1}(z)\, \mathbb{E}\left[\exp\left(\langle z, \eta_t X_t \rangle - \frac{1}{2} \|z\|_{X_t X_t^{\mathsf{T}}}^2\right) | \mathcal{H}_{t-1}\right]$$

$$\implies \mathbb{E}[M_t(z) | \mathcal{H}_{t-1}] \leq M_{t-1}(z) \qquad \text{(Using (*))}$$

Linear UCB – Regret Analysis

Fact 1: For a probability density h, if $M_t(z)$ is a non-negative super-martingale with $M_0(z) = 1$, the "mixture" $\bar{M}_t := \int_z M_t(z)\, h(z)\, dz$ is also a non-negative super-martingale with $\bar{M}_0 = 1$.

Fact 2: For a non-negative super-martingale $\bar{M}_t$ s.t. $\bar{M}_0 = 1$, for any $\epsilon > 0$, $\Pr[\sup_{t \in [T]} \bar{M}_t \geq \epsilon] \leq \frac{1}{\epsilon}$.

In order to construct $\bar{M}_t$, we will choose $h = \mathcal{N}(0, H^{-1})$ and $H = \lambda I_d$.

$$\bar{M}_t = \int_z M_t(z)\, h(z)\, dz = \frac{1}{\sqrt{(2\pi)^d \det[H^{-1}]}} \int_z \exp\left(\langle z, S_t \rangle - \frac{1}{2} \|z\|_{K_t}^2 - \frac{1}{2} \|z\|_H^2 \right) dz$$

From **Fact 1**, $\bar{M}_t$ is a non-negative super-martingale, and hence using **Fact 2** with $\epsilon = {}^1\!/\!_\delta$

$$\Pr\left[\sup_{t \in [T]} \bar{M}_t \geq \epsilon\right] = \Pr\left[\sup_{t \in [T]} \log(\bar{M}_t) \geq \log(\epsilon)\right] = \Pr\left[\sup_{t \in [T]} \log(\bar{M}_t) \geq \log(1/\delta)\right] \leq \delta$$

In the last part of the proof, we will relate $\bar{M}_t$ to $\|S_t\|_{V_t^{-1}}$.

Linear UCB – Regret Analysis

Recall that $\bar{M}_t = \int_z M_t(z)\, h(z)\, dz = \frac{1}{\sqrt{(2\pi)^d \ \det[H^{-1}]}} \int_z \exp\left(\langle z, S_t \rangle - \frac{1}{2} \ \|z\|^2_{K_t} - \frac{1}{2} \ \|z\|^2_H \right) \ dz.$

Simplifying the term inside exp,

$$\langle z, S_t \rangle - \frac{1}{2} \ \|z\|^2_{K_t} - \frac{1}{2} \ \|z\|^2_H = \frac{1}{2} \ \|S_t\|^2_{(K_t+H)^{-1}} - \frac{1}{2} \ \left\| z - (K_t + H)^{-1} S_t \right\|^2_{(K_t+H)}$$

$$\implies \int_z M_t(z)\, h(z)\, dz = \frac{\exp\left(\frac{1}{2} \ \|S_t\|^2_{V_t^{-1}} \right)}{\sqrt{(2\pi)^d \ \det[H^{-1}]}} \int_z \exp\left(-\frac{1}{2} \ \left\| z - V_t^{-1} S_t \right\|^2_{V_t} \right) \ dz$$

The integral corresponds to the integral of the PDF for a multivariate Gaussian with mean $V_t^{-1} S_t$ and covariance V_t^{-1}. For a Gaussian with mean μ and covariance Σ^{-1}, $\frac{1}{\sqrt{(2\pi)^d \det[\Sigma^{-1}]}} \int_z \exp\left(-\frac{1}{2} \ \|z - \mu\|^2_\Sigma \right) \ dz = 1.$ Hence,

$$\bar{M}_t = \frac{\exp\left(\frac{1}{2} \ \|S_t\|^2_{V_t^{-1}} \right)}{\sqrt{(2\pi)^d \ \det[H^{-1}]}} \sqrt{(2\pi)^d \ \det[V_t^{-1}]} = \sqrt{\frac{\det[H]}{\det[V_t]}} \ \exp\left(\frac{1}{2} \ \|S_t\|^2_{V_t^{-1}} \right)$$

Linear UCB – Regret Analysis

Putting everything together, we know that for all $t \in [T]$, w.p $1-\delta$, $\log(\bar{M}_t) \leq \log(1/\delta)$. Using the result from the previous slide, w.p $1-\delta$, for all $t \in [T]$

$$\frac{1}{2}\|S_t\|^2_{V_t^{-1}} + \frac{1}{2}\log\left(\frac{\det[H]}{\det[V_t]}\right) \leq \log(1/\delta) \implies \|S_t\|_{V_t^{-1}} \leq \sqrt{\log\left(\frac{\det[V_t]}{\lambda^d}\right) + 2\log(1/\delta)}$$

$$\implies \|S_t\|_{V_t^{-1}} \leq \sqrt{d\log\left(\frac{\lambda d + t}{\lambda d}\right) + 2\log(1/\delta)}$$

From **Part (i)**, we know that,

$$\left\|\theta^* - \hat{\theta}_t\right\|_{V_t} \leq \|S_t\|_{V_t^{-1}} + \sqrt{\lambda} \leq \underbrace{\sqrt{d\log\left(\frac{\lambda d + t}{\lambda d}\right) + 2\log(1/\delta)} + \sqrt{\lambda}}_{:=\sqrt{\beta_t}}$$

Hence, we have shown that w.p. $1-\frac{1}{T}$, $\left\|\theta^* - \hat{\theta}_t\right\|^2_{V_t} \leq \beta_t$, and hence $\Pr[G^c] \leq \frac{1}{T}$ □

Improvements to LinUCB

- LinUCB results in $O(d\sqrt{T}\log(T))$ regret. Importantly, the same regret analysis works for infinitely many arms and even for a potentially changing set of actions $\mathcal{A}_t$.
- When the number of arms is finite, fixed and equal to K, a phase-based elimination algorithm can achieve $O(\sqrt{dT\log(KT)})$ regret (see [LS20, Chapter 22]).
- **Lower Bound**: For any bandit algorithm, there exists a linear bandit instance (with the set of actions $\mathcal{A}$ equal to a unit hyper-cube or a unit sphere) such that $\text{Regret}(T) = \Omega(d\sqrt{T})$ (see [LS20, Chapter 24]).
- LinUCB maintain confidence intervals, and ensures optimism. An alternative set of strategies that work better in practice is *Posterior Sampling* of which Thompson Sampling is the most common (see [LS20, Chapter 36]).

Markov Decision Processes

Markov Decision Processes (MDPs)

- In bandit problems, the “state” of the environment does not change *as a result of an action*.
- Applications in robotics, operations research or conversational agents require explicitly modelling the current information available in a round.
- *Example 1*: A robot needs to model what is its position, velocity in order to take an action at the next round. This information is summarized as the “state” of the environment. The robot’s action changes its velocity, position and hence the “state”.
- *Example 2*: A conversational agent requires context (the past conversation, who it is speaking to) in order to decide what to respond to a particular user. The agent’s action can change the context of the conversation, and hence the “state”.
- Markov Decision Processes (MDPs) is the standard approach to sequential decision-making in such applications.

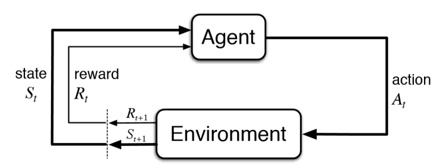
Markov Decision Processes (MDPs)

An MDP can be described by 5 elements: the state space ($\mathcal{S}$), action space ($\mathcal{A}$), starting state distribution (ρ), transition probabilities ($\mathcal{P}$) and rewards (r).

- State space $\mathcal{S}$
    - A state summarizes all the relevant information available to the agent. We will assume that the states are fully observable.
    - *Example*: Position of the rover on Mars, Inventory level of products.
    - States are mutually exclusive and exhaustive.
    - We will assume that the state space is discrete and finite, and $|\mathcal{S}| = S$.
- Starting state distribution $\rho \in \Delta_S$:
    - $\rho(s)$ corresponds to the probability that the agent starts in state s. $\sum_{s \in \mathcal{S}} \rho(s) = 1$.
- Action space $\mathcal{A}$:
    - Consists of the actions an agent can take. The action space can be different in each state.
    - *Example*: Move north for the Mars rover, buy more stock of a particular product.
    - We will assume that $\mathcal{A}$ is fixed, discrete and finite, and $|\mathcal{A}| = A$.

Markov Decision Processes (MDPs)

- Transition probabilities $\mathcal{P}$:
  - Model the inherent stochasticity in the system.
  - $\mathcal{P}(s'|s,a)$ is the probability of moving to a state s' when taking action a in state s.
  - **Markov property**: $\mathcal{P}(s'|s,a)$ only depends on the current state s and action a.
  - In some examples, such as robotics, transitions can be deterministic.
  - $\sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) = 1$; $\mathcal{P}(s'|s,a) \geq 0$.
  - If $\mathcal{P}$ does not change with t, the transition probabilities are referred to as *stationary*.
- Rewards r: Model how much the agent has moved towards achieving its goal.
  - $r(s,a)$ is the reward obtained on taking action a in state s.
  - The reward can depend on s', the state to which the agent transitioned to i.e. it is denoted as $r(s',a,s)$. In this case, $r(s,a) = \sum_{s' \in \mathcal{S}} r(s',a,s)\,\mathcal{P}(s'|s,a)$.



Protocol: At round (epoch) t, the agent observes state s_t and takes action a_t, transitions to state s_{t+1} and receives reward r_t.

Markov Decision Processes (MDPs)

Decision Rule: Describes the information and mechanism an agent uses to select an action in a given state and round. Can be classified as follows:

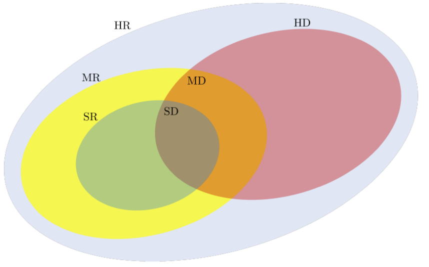
- *Information*: History dependent vs Markovian
  - A *history-dependent* decision rule uses some or all of the previous states and actions up to and including the current state when choosing an action.
  - A *Markovian* decision rule uses only the current state to select actions.
- *Mechanism*: Randomized vs Deterministic
  - A *randomized* decision rule maintains a probability distribution over the actions that can be taken in each state.
  - A *deterministic* decision rule corresponds to a degenerate distribution and consists of a deterministic mapping from states to actions.

- We define π_t to be the decision-rule at round t.
- A **policy** π is a sequence of decision rules, one for each round t, i.e. $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$.

Q: Why are history dependent policies computationally expensive to implement in general?

Markov Decision Processes (MDPs)

- The **policy class** depends on the decision rule it uses.
- A policy can be in Π_{HR}, Π_{HD}, Π_{MR}, Π_{MD} depending on whether the decision rule is history-dependent (H) or Markovian (M); randomized (R) or deterministic (D).
- *Example*: If $\mathcal{H}_t = \{S_0, A_0, S_1, \ldots, S_t\}$ is the history of interactions until round t, then, $\Pi_{\text{HR}} = \{\pi_0, \pi_1, \pi_2, \ldots\}$ where $\pi_t : \mathcal{H}_t \to \Delta_A$,
- *Example*: $\Pi_{\text{MD}} = \{\pi_0, \pi_1, \pi_2, \ldots\}$ where $\pi_t : S_t \to \mathcal{A}$.
- A policy is *stationary* if it uses the same decision rule in every round, i.e. $\pi = \{\pi_0, \pi_0, \ldots\}$.
- We will only consider stationary policies that are Markovian, and define $\Pi_{\text{SR}} \subset \Pi_{\text{MR}} \subset \Pi_{\text{HR}}$ and $\Pi_{\text{SD}} \subset \Pi_{\text{MD}} \subset \Pi_{\text{HD}}$.



Markov Decision Processes (MDPs)

- Specifying ρ and choosing a policy π results in a stochastic process over the state and action space. We will denote this *trajectory* as $(S_0, A_0, S_1, \ldots)$.
- When $\pi \in \Pi_{\mathrm{MR}}$, the stochastic process is a discrete-time Markov chain.

Q: For a policy $\pi \in \Pi_{\mathrm{MR}}$, calculate the probability of the trajectory $(s_0, a_0, s_1, a_1, \ldots)$

- The trajectory over states and actions generates a *reward process*: $\{r_0, r_1, \ldots,\} = (r(S_0, A_0), r(S_1, A_1), \ldots)$.
- When the stochastic process over states-actions is a Markov chain, the corresponding reward process is a *Markov reward process*.

Q: How do we judge whether one reward process is "better" than the other?

We need some notion of *utility*. Common choice of utility functions is *additive*, i.e. the utility of a reward process is $(r_0, r_1, \ldots,)$ is given by: $\mathbb{E}\left[\sum_{i=0} U_i(r_i)\right]$ where $U_i : \mathbb{R} \to \mathbb{R}$ and the expectation is over the different trajectories produced by the policy.

Markov Decision Processes (MDPs)

Choosing a policy gives rise to a reward process, and we can design additive utility functions to compare different reward processes. This gives different optimality criterion w.r.t to policies:

(a) For a finite *horizon* H, $\max_{\pi \in \Pi_{\mathbf{HR}}} \mathbb{E}\left[\sum_{t=1}^{H} r_t^{\pi}\right]$. [**Finite Horizon Total Reward**]

(b) For an infinite horizon, $\max_{\pi \in \Pi_{\mathbf{HR}}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t^{\pi}\right]$ where $\gamma \in (0, 1)$ is the discount factor. [**Infinite Horizon Discounted Reward**]

(c) For an infinite horizon, $\max_{\pi \in \Pi_{\mathbf{HR}}} \lim_{T \to \infty} \frac{\mathbb{E}\left[\sum_{t=\mathbf{0}}^{T} r_t^{\pi}\right]}{T}$. [**Infinite Horizon Average Reward**]

We will focus mainly on (b) infinite horizon discounted reward setting and towards the end, consider (a) finite horizon total reward setting.

Infinite Horizon Discounted Reward: The discount factor γ models the fact that near-term rewards are preferable to future rewards. For example, it models inflation meaning that a penny today is worth 10 in the future.

Infinite-horizon Discounted Setting

Objective: For a starting state s, find policy $\pi \in \Pi_{\text{HR}}$ that maximizes the **value function** $v^\pi(s)$

$$v^\pi(s_0) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s_0\right] ,$$

where the expectation is over the randomness in the reward process induced by policy π. For a starting state distribution ρ, the related objective is to maximizes $v^\pi(\rho) := \mathbb{E}_{s\sim\rho} v^\pi(s)$.

Assumptions

- The reward function does not change across rounds.
- The rewards are bounded in $[0, 1]$.

Q: What are the upper and lower-bounds on the value function?

Infinite-horizon Discounted Setting

Claim: For each $s \in \mathcal{S}$, for a given policy $\pi = (\pi_0, \pi_1, \ldots) \in \Pi_{\text{HR}}$, there exists a policy $\pi' = (\pi'_0, \pi'_1, \ldots) \in \Pi_{\text{MR}}$ with the same value, conditioned on $S_0 = s_0$.

- Since there exists a Markov policy that has the same value as every history-dependent policy, we only need to consider Π_{MR} when we optimize for the optimal policy.
- Markov policies only need to maintain the knowledge of the current state, and are hence computationally tractable.

Proof: Using the definition of the value function,

$$v^\pi(s_0) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s_0\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) | S_0 = s_0\right]$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \Pr^\pi[S_t = s, A_t = a | S_0 = s_0]$$

Here, $\Pr^\pi$ corresponds to the probability distribution induced by policy π.

Infinite-horizon Discounted Setting

Recall that $v^{\pi}(s_0) = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a) \Pr^{\pi}[S_t = s, A_t = a | S_0 = s_0]$ Construct $\pi' \in \Pi_{\text{MR}}$ as follows: $\pi'_t(A_t = a | S_t = s) = \pi_t[A_t = a | \mathcal{H}_t]$. We will prove (by induction) that $\Pr^{\pi}[S_t = s, A_t = a | S_0 = s_0] = \Pr^{\pi'}[S_t = s, A_t = a | S_0 = s_0]$, and hence, $v^{\pi}(s_0) = v^{\pi'}(s_0)$.

$$v^{\pi}(s_0) = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a) \Pr^{\pi}[A_t = a | S_t = s, S_0 = s_0] \Pr^{\pi}[S_t = s | S_0 = s_0]$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a)\, \pi_t[A_t = a | \mathcal{H}_t] \Pr^{\pi}[S_t = s | S_0 = s_0]$$

Base Case: For $t = 0$, RHS $= \sum_a \pi_0[A_0 = a | S_0 = s_0] = \sum_a \pi'_0(A_0 = a | S_0 = s)$ by def. of π'.

Inductive Hypothesis: For $t \geq 1$, assume that $\Pr^{\pi}[S_t = s, A_t = a | S_0 = s_0] = \Pr^{\pi'}[S_t = s, A_t = a | S_0 = s_0]$. Let us now prove it for $t + 1$.

For a fixed (s, a), using the definition of π',

$\pi_{t+1}[A_{t+1} = a | \mathcal{H}_{t+1}] \Pr^{\pi}[S_{t+1} = s | S_0 = s_0] = \pi'_{t+1}(A_{t+1} = a | S_{t+1} = s) \Pr^{\pi}[S_{t+1} = s | S_0 = s_0]$.

Hence, we need to show that $\Pr^{\pi}[S_{t+1} = s | S_0 = s_0] = \Pr^{\pi'}[S_{t+1} = s | S_0 = s_0]$.

Infinite-horizon Discounted Setting

Recall that $\pi'_t(A_t = a|S_t = s) = \pi_t[A_t = a|\mathcal{H}_t]$ by def. of π', and by the inductive hypothesis, $\Pr^{\pi}[S_t = s, A_t = a|S_0 = s_0] = \Pr^{\pi'}[S_t = s, A_t = a|S_0 = s_0]$.

$$\begin{aligned}\Pr^{\pi}[S_{t+1} = s|S_0 = s_0] &= \sum_s \sum_a \mathcal{P}[s'|s,a] \Pr^{\pi}[S_t = s', A_t = a|S_0 = s_0] \\ &= \sum_s \sum_a \mathcal{P}[s'|s,a] \Pr^{\pi'}[S_t = s', A_t = a|S_0 = s_0] \\ &\qquad \text{(Inductive Hypothesis)} \\ &= \Pr^{\pi'}[S_{t+1} = s|S_0 = s_0]\end{aligned}$$

Hence, $\Pr^{\pi}[S_{t+1} = s, A_{t+1} = a|S_0 = s_0] = \Pr^{\pi'}[S_{t+1} = s, A_{t+1} = a|S_0 = s_0]$. Using the definition of v^{π}, $v^{\pi}(s_0) = v^{\pi'}(s_0)$. □

Infinite-horizon Discounted Setting

Claim: For $\pi \in \Pi_{\text{MR}}$, if we define

$$r^{\pi_t} \in \mathbb{R}^S \quad \text{s.t.} \quad r^{\pi_t}(s) := \sum_{a \in \mathcal{A}} r(s,a)\, \pi_t[a|s]\,,$$

$$P^{\pi_t} \in \mathbb{R}^{S \times S} \quad \text{s.t.} \quad P^{\pi_t}[s,s'] = P^{\pi_t}(s \to s') := \sum_{a \in \mathcal{A}} \Pr[s'|s,a]\, \pi_t(a|s)\,,$$

then, $v^\pi \in \mathbb{R}^S$ can be expressed as:

$$v^\pi = \sum_{t=0}^{\infty} \gamma^t \left[\prod_{j=0}^{t-1} P^{\pi_j}\right] r^{\pi_t}\,.$$

Furthermore, for a policy $\pi \in \Pi_{\text{SR}}$, $v^\pi = r^\pi + \gamma\, P^\pi\, v^\pi$. Examining each component,

$$v^\pi(s) = r^\pi(s) + \gamma \sum_{s'} P^\pi[s,s']\, v^\pi(s') = \sum_{a \in \mathcal{A}} r(s,a)\, \pi[a|s] + \gamma \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{P}[s'|s,a]\, \pi[a|s]\, v^\pi(s')$$

This is the **Bellman equation** for a fixed policy $\pi \in \Pi_{\text{SR}}$.

Infinite-horizon Discounted Setting

Proof: Starting from the definition of $v^{\pi}(s_0)$,

$$v^{\pi}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s_0\right] = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \sum_{a \in cA} r(s, a) \Pr[S_t = s, A_t = a | S_0 = s_0]$$

Let us evaluate the first three terms in this sum,

For $t = 0:\ \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \Pr[S_0 = s, A_0 = a | S_0 = s_0] = \sum_{a \in \mathcal{A}} r(s_0, a)\, \pi_0(a|s_0) = r^{\pi_{\mathbf{0}}}(s_0)$

For $t = 1:\ \gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \Pr[A_1 = a | S_1 = s, S_0 = s_0] \Pr[S_1 = s_1 | S_0 = s_0]$

$$= \gamma \sum_{s \in \mathcal{S}} r^{\pi_{\mathbf{1}}}(s) \Pr[S_1 = s | S_0 = s_0] = \gamma \sum_{s \in \mathcal{S}} r^{\pi_{\mathbf{1}}}(s) \sum_{a \in \mathcal{A}} \mathcal{P}[s | s_0, a]\, \pi_0(a|s_0) = \gamma \sum_{s \in \mathcal{S}} r^{\pi_{\mathbf{1}}}(s)\, P^{\pi_{\mathbf{0}}}[s_0, s]$$

For $t = 2:\ \gamma^2 \sum_{s \in \mathcal{S}} r^{\pi_{\mathbf{2}}}(s) \Pr[S_2 = s | S_0 = s_0] = \sum_{s \in \mathcal{S}} r^{\pi_{\mathbf{1}}}(s) \sum_{s_{\mathbf{1}} \in \mathcal{S}} P^{\pi_{\mathbf{1}}}[s_1, s]\, P^{\pi_{\mathbf{0}}}[s_0, s_1]$

For a general $t:\ \gamma^t \sum_{s \in \mathcal{S}} r^{\pi_t}(s) \sum_{s_{t-\mathbf{1}} \in \mathcal{S}} \ldots \sum_{s_{\mathbf{1}} \in \mathcal{S}} P^{\pi_{t-\mathbf{1}}}[s_{t-1}, s]\, P^{\pi_{t-\mathbf{2}}}[s_{t-2}, s_{t-1}] \cdots P^{\pi_{\mathbf{0}}}[s_0, s_1]$

Infinite-horizon Discounted Setting

Recall that, $v^\pi(s_0) = \sum_{t=0}^{\infty} \gamma^t \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a)\, \Pr[S_t = s, A_t = a | S_0 = s_0]$, and that term t in the above sum is equal to $\gamma^t \sum_{s\in\mathcal{S}} r^{\pi_t}(s) \sum_{s_{t-1}\in\mathcal{S}} \cdots \sum_{s_1\in\mathcal{S}} P^{\pi_{t-1}}[s_{t-1}, s]\, P^{\pi_{t-2}}[s_{t-2}, s_{t-1}] \cdots P^{\pi_0}[s_0, s_1]$. Hence,

$$v^\pi(s_0) = \sum_{t=0}^{\infty} \gamma^t \sum_{s\in\mathcal{S}} r^{\pi_t}(s) \sum_{s_{t-1}\in\mathcal{S}} \cdots \sum_{s_1\in\mathcal{S}} P^{\pi_{t-1}}[s_{t-1}, s]\, P^{\pi_{t-2}}[s_{t-2}, s_{t-1}] \cdots$$

$$\implies v^\pi = \sum_{t=0}^{\infty} \gamma^t \left[\prod_{j=0}^{t-1} P^{\pi_j}\right] r^{\pi_t} \qquad (V^\pi(s_0) \text{ is the } s_0 \text{ component of the vector } v^\pi)$$

For a policy $\pi \in \Pi_{\mathrm{SR}}$, $P^{\pi_t} = P^\pi$ for all t. Hence,

$$v^\pi = \sum_{t=0}^{\infty} \gamma^t\, [P^\pi]^t\, r^\pi = r^\pi + \gamma\, P^\pi\, r^\pi + \gamma^2\, [P^\pi]^2\, r^\pi + \ldots$$

$$= r^\pi + \gamma\, P^\pi\, \left[r^\pi + \gamma\, P^\pi\, r^\pi + \gamma^2\, [P^\pi]^2\, r^\pi + \ldots\right] = r^\pi + \gamma\, P^\pi\, v^\pi$$

$$\implies v^\pi = r^\pi + \gamma\, P^\pi\, v^\pi \quad \square$$

Infinite-horizon Discounted Setting

For $\pi \in \Pi_{\text{MR}}$, we have seen that $v^\pi = r^\pi + \gamma\, P^\pi\, v^\pi$. This corresponds to a system of linear equations, and can be solved in closed form. Since $\gamma < 1$, and P^π is a stochastic matrix (i.e. its elements correspond to probabilities, and sums and columns add up to one), the eigenvalues of $I_S - \gamma P^\pi$ are strictly positive and hence it is invertible.

$$v^\pi = r^\pi + \gamma\, P^\pi\, v^\pi \implies (I_S - \gamma P^\pi)\, v^\pi = r^\pi \implies v^\pi = (I_S - \gamma P^\pi)^{-1}\, r^\pi\,.$$

- By the Neumann series, $(I - A)^{-1} = \sum_{t=0}^{\infty} A^t$. Hence, $(I_S - \gamma P^\pi)^{-1}\, r^\pi = \sum_{t=0}^{\infty} \gamma^t\, [P^\pi]^t$ which recovers the expression for v^π from the previous slide.
- Q: For a vector $x \geq 0$, prove that $(I_S - \gamma P^\pi)^{-1}\, x \geq x \geq 0$
- Q: For vectors $u \geq v$, prove that $(I_S - \gamma P^\pi)^{-1}\, u \geq (I_S - \gamma P^\pi)^{-1}\, v$

Bellman policy evaluation operator for policy π: $\mathcal{T}^\pi : \mathbb{R}^S \to \mathbb{R}^S$ s.t. for vector $u \in \mathbb{R}^S$ $\mathcal{T}^\pi u = r^\pi + \gamma P^\pi u$ and $(\mathcal{T}^\pi u)(s) = r^\pi(s) + \gamma \sum_{s'} P^\pi[s, s']\, u(s')$.

Bellman Optimality Operator

Define the **Bellman optimality operator** $\mathcal{T} : \mathbb{R}^S \to \mathbb{R}^S$. For a vector $u \in R^S$,

$$(\mathcal{T}u)(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) u(s') \right\}$$

Consider $w := \max_{\pi \in \Pi_{\mathbf{SD}}} \{r^\pi + \gamma P^\pi u\}$, where max refers to the element-wise maximum.

$$\begin{aligned} w(s) &= \max_{\pi \in \Pi_{\mathbf{SD}}} \left\{ r^\pi(s) + \gamma \sum_{s'} P^\pi[s,s'] u(s') \right\} \\ &= \max_{\substack{\pi(\cdot|s) \\ \exists a^* \text{ s.t } \pi(a^*|s)=1}} \left\{ r(s,a)\pi(a|s) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\pi(a|s) u(s') \right\} \end{aligned}$$

(Optimization over degenerate distributions)

$$= \max_a \left\{ r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) u(s') \right\} = (\mathcal{T}u)(s)$$

$$\implies \mathcal{T}u = \max_{\pi \in \Pi_{\mathbf{SD}}} \{r^\pi + \gamma P^\pi u\}$$

Bellman Optimality Operator

Claim: $\mathcal{T}$ is a contraction mapping with modulus γ, i.e. for any 2 vectors $u, w \in \mathbb{R}^S$ $\|\mathcal{T}u - \mathcal{T}w\|_\infty \leq \gamma \|u - w\|_\infty$.

Proof: For a fixed s, without loss of generality, consider the case when $(\mathcal{T}w)(s) \geq (\mathcal{T}v)(s)$. By the definition of $\mathcal{T}$, if $a^*(s) = \arg\max\{r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)w(s')\}$, then,

$$(\mathcal{T}w)(s) = r(s, a^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s))w(s')$$

$$r(s, a^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s))u(s') \leq \max\{r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)u(s')\} = (\mathcal{T}u)(s)$$

$$\begin{aligned} \implies (\mathcal{T}w)(s) - (\mathcal{T}u)(s) &\leq \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s))\left[w(s') - u(s')\right] \\ &\leq \gamma \|\mathcal{P}(\cdot|s, a^*(s)\|_1 \|w - u\|_\infty = \gamma \|w - u\|_\infty \end{aligned}$$

Similarly, $(\mathcal{T}w)(s) - (\mathcal{T}u)(s) \leq \gamma \|w - u\|_\infty$. Since this result is true for an arbitrary s,

$$\|\mathcal{T}u - \mathcal{T}w\|_\infty \leq \gamma \|w - u\|_\infty \quad \square$$

Banach's Fixed Point Theorem

Fact: Under certain technical assumptions, if L is a contraction mapping, then,

- There exists a unique fixed point u^* such that $Lu^* = u^*$.
- For any vector u_0, $u_{n+1} = Lu_n = L^{n+1}u_0$ converges to u^* i.e $\|u_n - u^*\| \to 0$ as $n \to \infty$.

Since the Bellman optimality operator, $\mathcal{T}$ is a contraction mapping, using Banach's Fixed Point Theorem above, there exists a fixed point $v^* \in \mathbb{R}^S$ s.t. $\mathcal{T}v^* = v^*$.

Similarly, $\mathcal{T}^\pi$ is also a contraction mapping with modulus γ, and converges to a fixed point equal to v^π. Prove in Assignment 2!

Claim: $\|u^* - \mathcal{T}^n u_0\| \leq \gamma^n \|u^* - u_0\|$ i.e. u_n converges to u^* at a linear rate.

Q: *Proof?*

Fundamental Theorem

Claim: There exists a policy $\pi^* \in \Pi_{\mathrm{SD}}$ s.t. $v^{\pi^*}(s) = \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}(s)$ for all $s \in \mathcal{S}$.

- Hence, for MDPs, it is sufficient to only consider the class of stationary, deterministic policies in order to compute the optimal policy.

Proof: We know the following:

(a) From Slide 19, $\max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}(s) = \max_{\pi \in \Pi_{\mathbf{MR}}} v^{\pi}(s)$.

(b) If v^* is the fixed point of T and $\pi^* \in \Pi_{\mathrm{SD}}$ is the *greedy* policy s.t. $\pi^*(s) = \arg\max_a \{r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\, v^*(s')\}$, then,

$$v^* = \mathcal{T} v^* = \max_{\pi \in \Pi_{\mathbf{SD}}} \{r^{\pi} + \gamma P^{\pi} u\} = r^{\pi^*} + \gamma P^{\pi^*} u$$

(c) $v^* = \max_{\pi \in \Pi_{\mathbf{SR}}} \{r^{\pi} + \gamma P^{\pi} u\}$ i.e. randomized policies cannot increase the value. (Prove in Assignment 2!)

We will prove that for a v s.t. $v = \mathcal{T} v$, then $v = \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$. Together with (b), this implies that $v^* = \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$ and that this value function corresponds to the policy $\pi^* \in \Pi_{\mathrm{SD}}$.

Fundamental Theorem

We will now prove that:

(i) If $v \geq \mathcal{T}v$, then $v \geq \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$.
(ii) If $v \leq \mathcal{T}v$, then $v \leq \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$.

Hence, if $v = \mathcal{T}v$, then $v = \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$.

Let us first prove (i): if $v \geq \mathcal{T}v$, then $v \geq \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$. Define an arbitrary $\pi' := \{\pi'_1, \pi'_2, \ldots, \} \in \Pi_{\mathrm{MR}}$. For an arbitrary i, define $\pi_i := \{\pi'_i, \pi'_i, \ldots\} \in \Pi_{\mathrm{SR}}$.

$$v \geq \mathcal{T}v = \max_{\pi \in \Pi_{\mathbf{SD}}} \{r^{\pi} + \gamma P^{\pi} u\} = \max_{\pi \in \Pi_{\mathbf{SR}}} \{r^{\pi} + \gamma P^{\pi} u\} \geq r^{\pi_i} + \gamma P^{\pi_i} v \qquad \text{(Using (c))}$$

$$\implies v \geq r^{\pi_{\mathbf{0}}} + \gamma P^{\pi_{\mathbf{0}}} v \geq r^{\pi_{\mathbf{0}}} + \gamma P^{\pi_{\mathbf{0}}}[r^{\pi_{\mathbf{1}}} + \gamma P^{\pi_{\mathbf{1}}} v] \implies v \geq \sum_{t=0}^{\infty} \gamma^t \left[\prod_{j=0}^{t-1} P^{\pi_j}\right] r^{\pi_t}$$

$$\text{(Recursing)}$$

$$\implies v \geq v^{\pi'} = \max_{\pi \in \Pi_{\mathbf{MR}}} v^{\pi} = \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi} \qquad \text{(Using def of } v^{\pi'} \text{ for } \pi' \in \Pi_{\mathrm{MR}}\text{, and then (a))}$$

Fundamental Theorem

Now let us prove (ii): if $v \leq \mathcal{T}v$, then $v \leq \max_{\pi \in \Pi_{\mathbf{HR}}} v^\pi$. For a specific $\pi \in \Pi_{\text{SD}}$,

$$
\begin{aligned}
& v \leq \mathcal{T}v = r^\pi + \gamma P^\pi v \leq r^\pi + \gamma P^\pi \left[r^\pi + \gamma P^\pi v \right] \implies v \leq \sum_{t=0}^{\infty} \gamma^t \left[P^\pi \right]^t r^\pi && \text{(Recursing)} \\
\implies & v \leq v^\pi \leq \max_{\pi \in \Pi_{\mathbf{SD}}} v^\pi && \text{(By def of } v^\pi \text{ for } \pi \in \Pi_{\text{SD}}) \\
& = \max_{\pi \in \Pi_{\mathbf{SR}}} v^\pi \leq \max_{\pi \in \Pi_{\mathbf{MR}}} v^\pi && \text{(Using (c))} \\
\implies & v = \max_{\pi \in \Pi_{\mathbf{HR}}} v^\pi \quad \square && \text{(Using (a))}
\end{aligned}
$$

The fundamental theorem immediately suggests a way to calculate π^*:

- Starting from an arbitrary vector $u \in \mathbb{R}^S$, iterate $u = \mathcal{T}u$ to converge to a fixed point u^*. Since the fixed point of $\mathcal{T}$ is unique, $u^* = v^*$.
- Once we have computed v^*, compute the greedy policy in each state $s \in \mathcal{S}$: $\pi^*(s) = \arg\max_a \{ r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \, v^*(s') \}$.

This is value iteration!

References i

Tor Lattimore and Csaba Szepesvári, *Bandit algorithms*, Cambridge University Press, 2020.