Experiment design, Markov Decision Processes and Reinforcement Learning

Optimal decisions, Part VII

Christos Dimitrakakis

Chalmers

November 10, 2013

Clinical trials

- We have a number of treatments of unknown efficacy.
- When a new patient arrives, we must choose one of them.
- Some, slightly different, goals:
  1. Maximise the number of cured patients.
  2. Discover the best treatment.
- The optimal design is better than randomly assigning patients to treatments.

Experimental design and Markov decision processes

The following problems

- Shortest path problems.
- Optimal stopping problems.
- Reinforcement learning problems.
- Experiment design problems.
- Multi-armed bandit problems.
- Advertising.

can be all formalised as Markov decision processes.

The stochastic n-armed bandit problem

- Actions $\mathcal{A} = \{1, \ldots, n\}$.
- Expected reward $\mathbb{E}(r_t \mid a_t = i) = \omega_i$.
- Select actions to maximise

$$\sum_{t=0}^{T} \gamma^t r_t,$$

with discount factor $\gamma \in [0, 1]$, horizon $T \geq 0$.

The stochastic n-armed bandit problem

- Actions $\mathcal{A} = \{1, \ldots, n\}$.
- Expected reward $\mathbb{E}(r_t \mid a_t = i) = \omega_i$.
- Select actions to maximise

$$\sum_{t=0}^{T} \gamma^t r_t,$$

with discount factor $\gamma \in [0, 1]$, horizon $T \geq 0$.

Decision-theoretic approach

- Assume $r_t \mid a_t = i \sim \psi(\omega_i)$, with $\omega_i \in \Omega_i$, $\omega \in \Omega \triangleq \prod_i \Omega_i$ unknown parameters.
- Define prior $\xi(\omega_1, \ldots, \omega_n)$.
- Select actions to maximise $\mathbb{E}_\xi U_t = \mathbb{E}_\xi \sum_{k=1}^{T-t} \gamma^k r_{t+k}$.

Bernoulli example.

Consider n Bernoulli distributions with unknown parameters ω_i, $i = 1, \ldots, n$ such that

$$r_t \mid a_t = i \sim \mathcal{B}ern(\omega_i), \qquad \mathbb{E}(r_t \mid a_t = i) = \omega_i. \tag{1.1}$$

We model our belief for each bandit's parameter ω_i as a Beta distribution $\mathcal{B}eta(\alpha_i, \beta_i)$, with density $f(\omega \mid \alpha_i, \beta_i)$ so that

$$\xi(\omega_1, \ldots, \omega_n) = \prod_{i=1}^{n} f(\omega_i \mid \alpha_i, \beta_i).$$

$$N_{t.i} \triangleq \sum_{k=1}^{t} \mathbb{I}\{a_k = i\}$$

$$\hat{r}_{t.i} \triangleq \frac{1}{N_{t,i}} \sum_{k=1}^{t} r_t \, \mathbb{I}\{a_k = i\}$$

Then, the posterior distribution for the parameter of arm i is

$$\xi_t = \mathcal{B}eta(\alpha_i + N_{t,i}\hat{r}_{t,i} \ , \ \beta_i + N_{t,i}(1 - \hat{r}_{t,i}))$$

Since $r_t \in \{0, 1\}$ the possible states of our belief given some prior are $\mathbb{N}^{2n}$.

Belief states

- The state of the bandit problem is the state of our belief.
- A sufficient statistic is the number of plays and total rewards.
- Our state ξ_t is described by the priors α, β and the vectors

$$N_t = (N_{t,1}, \dots, N_{t,i}) \tag{1.2}$$

$$\hat{r}_t = (\hat{r}_{t,1}, \dots, \hat{r}_{t,i}). \tag{1.3}$$

- The next-state probabilities are defined as:

$$\xi_t(r_t = 1 \mid a_t = i) = \frac{\alpha_i + N_{t,i}\hat{r}_{t,i}}{\alpha_i + \beta_i + N_{t,i}}$$

- Thus decision-theoretic n-armed bandit problem can be formalised as a Markov decision process.

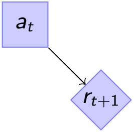


Figure: The basic bandit process

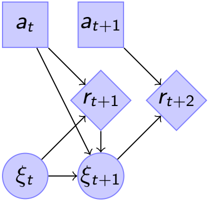


Figure: The decision-theoretic bandit process

Reinforcement learning

The reinforcement learning problem.

Learning to act in an unknown environment, by interaction and reinforcement.

- The environment has a changing state s_t.
- The agent obtains observations x_t.
- The agent takes actions a_t based on our observations.
- It receives rewards r_t.

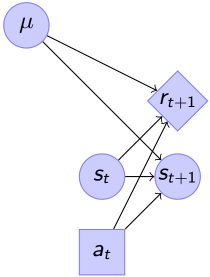
The goal (informally)

Maximise total reward $\sum_t r_t$

Types of environments

- Markov decision processes (MDPs).
- Partially observable MDPs (POMDPs).
- (Partially observable) Markov games.

Markov decision processes



Markov decision processes (MDP).

At each time step t:

- We observe state $s_t \in \mathcal{S}$.
- We take action $a_t \in \mathcal{A}$.
- We receive a reward $r_t \in \mathbb{R}$.

Markov property of the reward and state distribution

$$\mathbb{P}_\mu(s_{t+1} \in S \mid s_t, a_t) = \mathbb{P}_\mu(s_{t+1} \in S \mid s_1, a_1, \ldots, s_t, a_t) \qquad \text{(Transition distribution)}$$

$$\mathbb{P}_\mu(r_{t+1} \in R \mid s_t, a_t) = \mathbb{P}_\mu(r_{t+1} \in R \mid s_1, a_1, \ldots, s_t, a_t) \qquad \text{(Reward distribution)}$$

The agent

The agent's policy π

$$\mathbb{P}^\pi(a_t \mid s_t, \ldots, s_1, a_{t-1}, \ldots, a_1) \qquad \text{(history-dependent policy)}$$

$$\mathbb{P}^\pi(a_t \mid s_t) \qquad \text{(Markov policy)}$$

Definition 1 (Utility)

$$U_t \triangleq \sum_{k=0}^{T-t} r_{t+k}$$

We wish to find π maximising the expected total future reward

$$\mathbb{E}_\mu^\pi U_t = \mathbb{E}_\mu^\pi \sum_{k=0}^{T-t} r_{t+k} \qquad \text{(expected utility)}$$

to the horizon T .

The agent

The agent's policy π

$$\mathbb{P}^\pi(a_t \mid s_t, \ldots, s_1, a_{t-1}, \ldots, a_1) \qquad \text{(history-dependent policy)}$$

$$\mathbb{P}^\pi(a_t \mid s_t) \qquad \text{(Markov policy)}$$

Definition 1 (Utility)

$$U_t \triangleq \sum_{k=0}^{T-t} \gamma^k r_{t+k}$$

We wish to find π maximising the expected total future reward

$$\mathbb{E}^\pi_\mu U_t = \mathbb{E}^\pi_\mu \sum_{k=0}^{T-t} \gamma^k r_{t+k} \qquad \text{(expected utility)}$$

to the horizon T with discount factor $\gamma \in (0, 1]$.

State value function

$$V^{\pi}_{\mu,t}(s) \triangleq \mathbb{E}^{\pi}_{\mu}(U_t \mid s_t = s) \tag{2.1}$$

State-action value function

$$Q^{\pi}_{\mu,t}(s, a) \triangleq \mathbb{E}^{\pi}_{\mu}(U_t \mid s_t = s, a_t = a) \tag{2.2}$$

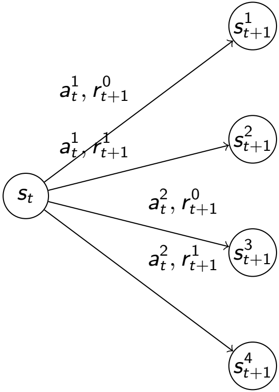
$$\pi^*(\mu) : V^{\pi^*(\mu)}_{t,\mu}(s) \geq V^{\pi}_{t,\mu}(s) \quad \forall \pi, t, s \tag{2.3}$$

The optimal policy π^* dominates all other policies π everywhere in $\mathcal{S}$.

$$V^*_{t,\mu}(s) \triangleq V^{\pi^*(\mu)}_{t,\mu}(s), \quad Q^*_{t,\mu}(s) \triangleq Q^{\pi^*(\mu)}_{t,\mu}(s, a). \tag{2.4}$$

The optimal value function V^* is the value function of the optimal policy π^*.

Finding the optimal policy when μ is known



Iterative/offline methods

- Estimate the optimal value function V^* (i.e. with backwards induction on all $\mathcal{S}$).
- Iteratively improve π (i.e. with policy iteration) to obtain π^*.

Online methods

- Forward search followed by backwards induction (on subset of $\mathcal{S}$).

Policy evaluation

An optimal policy

An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. – Bellman.

$$V^{\pi}_{\mu,t}(s) \triangleq \mathbb{E}^{\pi}_{\mu}(U_t \mid s_t = s) \tag{3.1}$$

$$\tag{3.2}$$

This derivation directly gives a number of policy evaluation algorithms.

Policy evaluation

$$V^{\pi}_{\mu,t}(s) \triangleq \mathbb{E}^{\pi}_{\mu}(U_t \mid s_t = s) \tag{3.1}$$

$$= \sum_{k=0}^{T-t} \mathbb{E}^{\pi}_{\mu}(r_{t+k} \mid s_t = s) \tag{3.2}$$

$$\tag{3.3}$$

This derivation directly gives a number of policy evaluation algorithms.

Policy evaluation

$$V^{\pi}_{\mu,t}(s) \triangleq \mathbb{E}^{\pi}_{\mu}(U_t \mid s_t = s) \tag{3.1}$$

$$= \sum_{k=0}^{T-t} \mathbb{E}^{\pi}_{\mu}(r_{t+k} \mid s_t = s), \quad U_{t+1} = \sum_{k=1}^{T-t} r_{t+k}. \tag{3.2}$$

$$= \mathbb{E}^{\pi}_{\mu}(r_t \mid s_t = s) + \mathbb{E}^{\pi}_{\mu}(U_{t+1} \mid s_t = s) \tag{3.3}$$

$$\tag{3.4}$$

This derivation directly gives a number of policy evaluation algorithms.

Policy evaluation

$$V^{\pi}_{\mu,t}(s) \triangleq \mathbb{E}^{\pi}_{\mu}(U_t \mid s_t = s) \tag{3.1}$$

$$= \sum_{k=0}^{T-t} \mathbb{E}^{\pi}_{\mu}(r_{t+k} \mid s_t = s) \tag{3.2}$$

$$= \mathbb{E}^{\pi}_{\mu}(r_t \mid s_t = s) + \mathbb{E}^{\pi}_{\mu}(U_{t+1} \mid s_t = s) \tag{3.3}$$

$$= \mathbb{E}^{\pi}_{\mu}(r_t \mid s_t = s) + \sum_{i \in \mathcal{S}} V^{\pi}_{\mu,t+1}(i)\, \mathbb{P}^{\pi}_{\mu}(s_{t+1} = i | s_t = s). \tag{3.4}$$

$$\tag{3.5}$$

This derivation directly gives a number of policy evaluation algorithms.

Algorithm 1 Direct policy evaluation

1: **for** $s \in \mathcal{S}$ **do**
2: **for** $t = 0, \ldots, T$ **do**
3:
$$\hat{V}_t(s) = \sum_{k=t}^{T} \sum_{j \in \mathcal{S}} \mathbb{P}^{\pi}_{\mu}(s_k = j \mid s_k = s)\, \mathbb{E}^{\pi}_{\mu}(r_k \mid s_k = j).$$
4: **end for**
5: **end for**

Algorithm 2 Monte-Carlo policy evaluation

for $s \in \mathcal{S}$ **do**
  for $k = 0, \ldots, K$ **do**

$$\hat{V}_k(s) = \sum_{k=0}^{T} r_{t,k}, \qquad \hat{V}(s) = \frac{1}{K} \sum_{k=1}^{K} \hat{V}_k(s).$$

  end for
end for

Remark 1

The Monte Carlo evaluation algorithm has the property:

$$\|V - \hat{V}\|_\infty \leq \sqrt{\frac{\ln(2|\mathcal{S}|/\delta)}{2K}}, \qquad \textit{with probability } 1 - \delta$$

Proof.

From Hoeffding's inequality, applied to any s, we have that

$$\mathbb{P}\left(|\hat{V}(s) - V(s)| \geq \sqrt{\frac{\ln(2|\mathcal{S}|/\delta)}{2K}}\right) \leq \delta/|\mathcal{S}|.$$

Algorithm 3 Backwards induction policy evaluation

For each state $s \in S$, for $t = 1, \ldots, T-1$:

$$\hat{V}_t(s) = r(s) + \sum_{j \in S} \mathbb{P}_{\mu,\pi}(s_{t+1} = j \mid s_t = s) \hat{V}_{t+1}(j), \tag{3.6}$$

with $\hat{V}_T(s) = r(s)$.

Theorem 2

Algorithm 3 results in estimates with the property:

$$\hat{V}_t(s) = V^{\pi}_{\mu,t}(s) \tag{3.7}$$

Algorithm 4 Finite-horizon backwards induction

Input μ, $\mathcal{S}_T$.
Initialise $V_T(s)$, for all $s \in \mathcal{S}_T$.
for $n = T - 1, T - 2, \ldots, 1$ **do**
  for $s \in \mathcal{S}_n$ **do**
    $\pi_n(s) = \arg\max_a \mathbb{P}_\mu(s'|s, a)[\mathbb{E}_\mu(r|s', s) + V^*_{n+1}(s')]$
    $V_n(s) = \sum_{s' \in \mathcal{S}_{n+1}} \mathbb{P}_\mu(s'|s, \pi_n(s))[\mathbb{E}_\mu(r|s', s) + V_{n+1}(s')]$
  end for
end for
Return $\pi = (\pi_n)_{n=1}^T$.

Notes

- $\mathbb{P}_{\mu,\pi}(s'|s) = \sum_a \mathbb{P}_\mu(s'|s, a)\, \mathbb{P}_\pi(a|s)$.
- Finite horizon problems only, or approximations (e.g. lookahead in game trees).
- For stochastic problems , we marginalize over states.
- As we know the optimal choice at the last step, we can find the optimal policy!
- Can be used with estimates of the value function.

Theorem 3

For a T-horizon problems, backwards induction is optimal, i.e.

$$V_n(s) = V^*_{\mu,n}(s) \tag{3.8}$$

Proof.

1. First we show that $V_t \geq V^*_t$.

□

Theorem 3

For a T-horizon problems, backwards induction is optimal, i.e.

$$V_n(s) = V^*_{\mu,n}(s) \tag{3.8}$$

Proof.

1. First we show that $V_t \geq V^*_t$.
2. For $n = T$, $V_T(s) = r(s) = V^{\pi}_{\mu,T}(s)$.

□

Theorem 3

For a T-horizon problems, backwards induction is optimal, i.e.

$$V_n(s) = V^*_{\mu,n}(s) \tag{3.8}$$

Proof.

1. First we show that $V_t \geq V^*_t$.
2. For $n = T$, $V_T(s) = r(s) = V^\pi_{\mu,T}(s)$.
3. Assume that for $n \geq t + 1$, (3.8) holds.

□

Theorem 3

For a T-horizon problems, backwards induction is optimal, i.e.

$$V_n(s) = V^*_{\mu,n}(s) \tag{3.8}$$

Proof.

1. First we show that $V_t \geq V^*_t$.
2. For $n = T$, $V_T(s) = r(s) = V^\pi_{\mu,T}(s)$.
3. Assume that for $n \geq t + 1$, (3.8) holds.
4. Then it holds for $n = t$ since:

$$V_t(s) = \max_a \left\{ r(s) + \sum_{j \in \mathcal{S}} p(j|s,a) V_{t+1}(j) \right\}$$

Theorem 3

For a T-horizon problems, backwards induction is optimal, i.e.

$$V_n(s) = V^*_{\mu,n}(s) \tag{3.8}$$

Proof.

1. First we show that $V_t \geq V^*_t$.
2. For $n = T$, $V_T(s) = r(s) = V^{\pi}_{\mu,T}(s)$.
3. Assume that for $n \geq t + 1$, (3.8) holds.
4. Then it holds for $n = t$ since:

$$V_t(s) \geq \max_a \left\{ r(s) + \sum_{j \in \mathcal{S}} p(j|s,a) V^*_{\mu,t+1}(j) \right\} \qquad \text{(by step 3)}$$

Theorem 3

For a T-horizon problems, backwards induction is optimal, i.e.

$$V_n(s) = V^*_{\mu,n}(s) \tag{3.8}$$

Proof.

1. First we show that $V_t \geq V^*_t$.
2. For $n = T$, $V_T(s) = r(s) = V^{\pi}_{\mu,T}(s)$.
3. Assume that for $n \geq t + 1$, (3.8) holds.
4. Then it holds for $n = t$ since:

$$V_t(s) \geq \max_a \left\{ r(s) + \sum_{j \in \mathcal{S}} p(j|s,a) V^{\pi'}_{\mu,t+1}(j) \right\}$$

Theorem 3

For a T-horizon problems, backwards induction is optimal, i.e.

$$V_n(s) = V^*_{\mu,n}(s) \tag{3.8}$$

Proof.

1. First we show that $V_t \geq V^*_t$.
2. For $n = T$, $V_T(s) = r(s) = V^\pi_{\mu,T}(s)$.
3. Assume that for $n \geq t + 1$, (3.8) holds.
4. Then it holds for $n = t$ since:

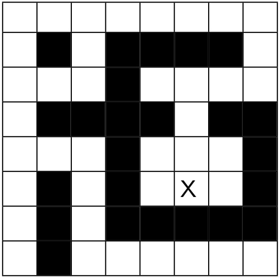
$$V_t(s) \geq V_t^{\pi'}(s)$$

5. The above holds for any policy π', including $\pi' = \pi$, the policy returned by backwards induction. Then:

$$V^*_{\mu,t}(s) \geq V^\pi_{\mu,t}(s) = V_t(s) \geq V^*_{\mu,t}(s).$$

□

Deterministic shortest-path problems



Properties

- $\gamma = 1$, $T \to \infty$.
- $r_t = -1$ unless $s_t = X$, in which case $r_t = 0$.
- $\mathbb{P}_\mu(s_{t+1} = X | s_t = X) = 1$.
- $\mathcal{A} = \{\text{North}, \text{South}, \text{East}, \text{West}\}$
- Transitions are deterministic and walls block.

What is the shortest path to the destination from any point?

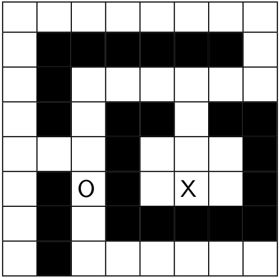
Shortest-path problem solution

14	13	12	11	10	9	8	7
15		13					6
16	15	14		4	3	4	5
17					2		
18	19	20		2	1	2	
19		21		1	0	1	
20		22					
21		23	24	25	26	27	28

Properties

- $\gamma = 1$, $T \to \infty$.
- $r_t = -1$ unless $s_t = X$, in which case $r_t = 0$.
- The length of the shortest path from s equals the negative value of the optimal policy.
- Also called *cost-to-go*.
- Remember Dijkstra's algorithm?

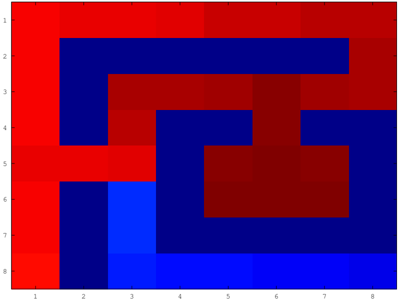
Stochastic shortest path problem, with a pit



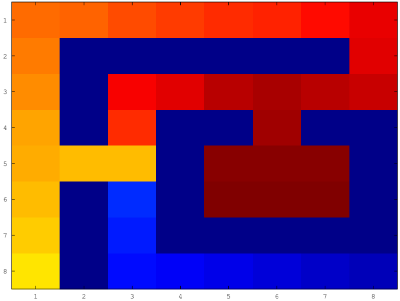
Properties

- $\gamma = 1$, $T \to \infty$.
- $r_t = -1$, but $r_t = 0$ at X and -100 at O and episode ends.
- $\mathbb{P}_\mu(s_{t+1} = X | s_t = X) = 1$.
- $\mathcal{A} = \{\mathrm{North}, \mathrm{South}, \mathrm{East}, \mathrm{West}\}$
- Moves to a random direction with probability ω. Walls block.

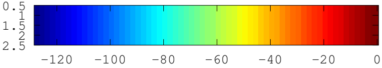
For what value of ω is it better to take the dangerous shortcut? (However, if we want to take into account risk explicitly we must modify the agent's utility function)



(a) $\omega = 0.1$



(b) $\omega = 0.5$



(c) value

Figure: Pit maze solutions for two values of ω.

Continuing stochastic MDPs

Inventory management

- There are K storage locations.
- Each place can store n_i items.
- At each time-step there is a probability ϕ_i that a client try to buy an item from location i, $\sum_i \phi_i \leq 1$. If there is an item available, you gain reward 1.
- Action 1: ordering u units of stock, for paying $c(u)$.
- Action 2: move u units of stock from one location i to another, j, for a cost $\psi_{ij}(u)$.

An easy special case

- $K = 1$.
- There is one type of item only.
- Orders are placed and received every n timesteps.

Inventory management

An easy special case

- $K = 1$.
- Deliveries happen once every m timesteps.
- Each time-step a client arrives with probability ϕ.

Properties

- The state set .
- The action set .
- The transition probabilities

Inventory management

An easy special case

- $K = 1$.
- Deliveries happen once every m timesteps.
- Each time-step a client arrives with probability ϕ.

Properties

- The state set is the number of items we have: $\mathcal{S} = \{0, 1, \ldots, n\}$.
- The action set .
- The transition probabilities

Inventory management

An easy special case

- $K = 1$.
- Deliveries happen once every m timesteps.
- Each time-step a client arrives with probability ϕ.

Properties

- The state set is the number of items we have: $\mathcal{S} = \{0, 1, \ldots, n\}$.
- The action set $\mathcal{A} = \{0, 1, \ldots, n\}$ since we can order from nothing up to n items.
- The transition probabilities

Inventory management

An easy special case

- $K = 1$.
- Deliveries happen once every m timesteps.
- Each time-step a client arrives with probability ϕ.

Properties

- The state set is the number of items we have: $\mathcal{S} = \{0, 1, \ldots, n\}$.
- The action set $\mathcal{A} = \{0, 1, \ldots, n\}$ since we can order from nothing up to n items.
- The transition probabilities $P(s'|s, a) = \binom{m}{d} \phi^d (1 - \phi)^{m-d}$, where $d = s + a - s'$, for $s + a \leq n$.

Discounted total reward

$$U_t = \lim_{T \to \infty} \sum_{k=t}^{T} \gamma^k r_k, \qquad \gamma \in (0, 1)$$

Definition 4

A policy π is stationary if $\pi(a_t \mid s_t) = \pi(a_n \mid s_n)$ for all n, t.

Remark 2

We can use the Markov chain kernel $\boldsymbol{P}$ to write the expected reward vector as

$$\boldsymbol{v}^\pi = \sum_{t=0}^{\infty} \gamma^t \boldsymbol{P}_{\mu,\pi}^t \boldsymbol{r} \tag{5.1}$$

Theorem 5

For any stationary π, $\boldsymbol{v}^\pi$ is the unique solution of

$$\boldsymbol{v} = \boldsymbol{r} + \gamma \boldsymbol{P}_{\mu,\pi} \boldsymbol{v}. \quad \leftarrow \textit{fixed point} \tag{5.2}$$

In addition, the solution is:

$$\boldsymbol{v}^\pi = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi})^{-1} \boldsymbol{r}. \tag{5.3}$$

Proof.

□

Theorem 5

For any stationary π, $\boldsymbol{v}^\pi$ is the unique solution of

$$\boldsymbol{v} = \boldsymbol{r} + \gamma \boldsymbol{P}_{\mu,\pi} \boldsymbol{v}. \quad \leftarrow \textit{fixed point} \tag{5.2}$$

In addition, the solution is:

$$\boldsymbol{v}^\pi = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi})^{-1} \boldsymbol{r}. \tag{5.3}$$

Proof.

1. $\boldsymbol{r} = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi}) \boldsymbol{v}$

Theorem 5

For any stationary π, $\boldsymbol{v}^\pi$ is the unique solution of

$$\boldsymbol{v} = \boldsymbol{r} + \gamma \boldsymbol{P}_{\mu,\pi} \boldsymbol{v}. \quad \leftarrow \textit{fixed point} \tag{5.2}$$

In addition, the solution is:

$$\boldsymbol{v}^\pi = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi})^{-1} \boldsymbol{r}. \tag{5.3}$$

Proof.

1. $\boldsymbol{r} = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi}) \boldsymbol{v}$
2. Since $\|\gamma \boldsymbol{P}_{\mu,\pi}\| < 1 \cdot \|\boldsymbol{P}_{\mu_\pi}\| = 1$, the following inverse exists:

$$(\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi})^{-1} = \lim_{n \to \infty} \sum_{t=0}^{n} (\gamma \boldsymbol{P}_{\mu,\pi})^t$$

□

Theorem 5

For any stationary π, $\boldsymbol{v}^\pi$ is the unique solution of

$$\boldsymbol{v} = \boldsymbol{r} + \gamma \boldsymbol{P}_{\mu,\pi} \boldsymbol{v}. \quad \leftarrow \textit{fixed point} \tag{5.2}$$

In addition, the solution is:

$$\boldsymbol{v}^\pi = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi})^{-1} \boldsymbol{r}. \tag{5.3}$$

Proof.

1. $\boldsymbol{r} = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi}) \boldsymbol{v}$
2. Since $\|\gamma \boldsymbol{P}_{\mu,\pi}\| < 1 \cdot \|\boldsymbol{P}_{\mu_\pi}\| = 1$, the following inverse exists:

$$(\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi})^{-1} = \lim_{n \to \infty} \sum_{t=0}^{n} (\gamma \boldsymbol{P}_{\mu,\pi})^t$$

3. Using step 1 and then 2,

$$\boldsymbol{v} = (\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi})^{-1} r = \sum_{t=0}^{\infty} \gamma^t \boldsymbol{P}_{\mu,\pi}^t \boldsymbol{r} = \boldsymbol{v}^\pi,$$

where the last step is by Remark 2

Definition 6 (Bellman operator)

$$\mathscr{L}_\pi \boldsymbol{v} \triangleq \boldsymbol{r} + \gamma \boldsymbol{P}_\pi \boldsymbol{v}$$

$$\mathscr{L} \boldsymbol{v} \triangleq \sup_\pi \{\boldsymbol{r} + \gamma \boldsymbol{P}_\pi \boldsymbol{v}\}, \qquad \boldsymbol{v} \in \mathcal{V}$$

$$\boldsymbol{v} = \mathscr{L} \boldsymbol{v} \qquad \text{(Bellman optimality equation)}$$

Theorem 7

For any bounded $\boldsymbol{r}$, it holds that for $\boldsymbol{v} \in \mathcal{V}$:

- *If $\boldsymbol{v} \geq \mathscr{L}\boldsymbol{v}$, then $\boldsymbol{v} \geq \boldsymbol{v}^*$*
- *If $\boldsymbol{v} \leq \mathscr{L}\boldsymbol{v}$, then $\boldsymbol{v} \leq \boldsymbol{v}^*$*
- *If $\boldsymbol{v} = \mathscr{L}\boldsymbol{v}$, then $\boldsymbol{v}$ is unique and $\boldsymbol{v} = \boldsymbol{v}^*$,*

where $\boldsymbol{v}^* = \sup_\pi \boldsymbol{v}^\pi$.

Theorem 8 (Banach Fixed-Point theorem)

Suppose $\mathcal{S}$ is a Banach space (i.e. a complete normed linear space) and $T : \mathcal{S} \to \mathcal{S}$ is a contraction mapping (i.e. $\exists \gamma \in [0, 1)$ s.t. $\|Tu - Tv\| \leq \gamma \|u - v\|$ for all $u, v \in \mathcal{S}$). Then

- *There is a unique $u^* \in U$ s.t. $Tu^* = u^*$ and*
- *For any $u^0 \in \mathcal{S}$ the sequence $\{u^n\}$:*

$$u^{n+1} = Tu^n = T^{n+1} u^0$$

converges to u^*.

Proof.

For any $m \geq 1$

$$\|u^{n+m} - u^n\| \leq \sum_{k=0}^{m-1} \|u^{n+k+1} - u^{n+k}\| = \sum_{k=0}^{m-1} \|T^{n+k} u^1 - T^{n+k} u^0\|$$

□

Theorem 8 (Banach Fixed-Point theorem)

Suppose $\mathcal{S}$ is a Banach space (i.e. a complete normed linear space) and $T : \mathcal{S} \to \mathcal{S}$ is a contraction mapping (i.e. $\exists \gamma \in [0, 1)$ s.t. $\|Tu - Tv\| \leq \gamma \|u - v\|$ for all $u, v \in \mathcal{S}$). Then

- *There is a unique $u^* \in U$ s.t. $Tu^* = u^*$ and*
- *For any $u^0 \in \mathcal{S}$ the sequence $\{u^n\}$:*

$$u^{n+1} = Tu^n = T^{n+1} u^0$$

converges to u^*.

Proof.

For any $m \geq 1$

$$\|u^{n+m} - u^n\| \leq \sum_{k=0}^{m-1} \|u^{n+k+1} - u^{n+k}\| = \sum_{k=0}^{m-1} \|T^{n+k} u^1 - T^{n+k} u^0\|$$

$$\leq \sum_{k=0}^{m-1} \gamma^{n+k} \|u^1 - u^0\| = \frac{\gamma^n (1 - \gamma^m)}{1 - \gamma} \|u^1 - u^0\|.$$

□

Theorem 9

If $\gamma \in [0, 1)$ then the Bellman operator $\mathscr{L}$ is a contraction mapping in $\mathcal{V}$.

Proof.

Let $\boldsymbol{v}, \boldsymbol{v}' \in \mathcal{V}$. Consider $s \in \mathcal{S}$ s. t. $\mathscr{L}\boldsymbol{v}(s) \geq \mathscr{L}\boldsymbol{v}'(s)$, and let

$$a_s^* \in \arg\max_{a \in \mathcal{A}} \left\{ r(s) + \sum_{j \in \mathcal{S}} \gamma p_\mu(j \mid s, a) \boldsymbol{v}(j) \right\}.$$

Then

$$0 \leq \mathscr{L}\boldsymbol{v}(s) - \mathscr{L}\boldsymbol{v}'(s) \leq \gamma \|\boldsymbol{v} - \boldsymbol{v}'\|.$$

Repeating the argument for s such that $\mathscr{L}\boldsymbol{v}(s) \leq \mathscr{L}\boldsymbol{v}'(s)$, we obtain

$$|\mathscr{L}\boldsymbol{r}(s) - \mathscr{L}\boldsymbol{r}'(s)| \leq \gamma \|\boldsymbol{r} - \boldsymbol{r}'\|.$$

Taking the supremum, we obtain the required result. □

Theorem 10

If $\gamma \in [0, 1)$, $\mathcal{S}$ is discrete and $\boldsymbol{r}$ is bounded:

- *There is a unique $\boldsymbol{v}^* \in \mathcal{V}$ s.t. $\mathscr{L}\boldsymbol{v}^* = \boldsymbol{v}^*$ and such that $\boldsymbol{v}^* = V^*_\mu$.*
- *For a stationary π, there is a unique $\boldsymbol{v} \in \mathcal{V}$ such that $\mathscr{L}_\pi \boldsymbol{v} = \boldsymbol{v}$ and $\boldsymbol{v} = V^\pi_\mu$.*

Proof.

- From the previous theorem, $\mathscr{L}$ is a contraction. So, we can apply the Fixed-Point theorem. Thus there is a unique solution. This is the optimal value function due to Theorem9
- Use part 1 with $\Pi = \{\pi\}$.

□

Algorithm 5 Value iteration

Input μ, $\mathcal{S}$.
Initialise $\boldsymbol{v}_0 \in \mathcal{V}$.
for $n = 1, 2, \ldots$ **do**
  for $s \in \mathcal{S}_n$ **do**
    $\pi_n(s) = \arg\max_a r(s) \sum_{s' \in \mathcal{S}} \mathbb{P}_\mu(s' | s, a) \boldsymbol{v}_{n-1}(s')$
    $\boldsymbol{v}_n(s) = r(s) + \sum_{s' \in \mathcal{S}} \mathbb{P}_\mu(s' | s, \pi_n(s)) \boldsymbol{v}_{n-1}(s')$
  end for
  break if `termination-condition` is met
end for
Return π_n, V_n.

Theorem 11

The value iteration algorithm satisfies

- $\lim_{n\to\infty} \|\boldsymbol{v}_n - V^*\| = 0$.
- *There exists* $N < \infty$ *such that*

$$\|\boldsymbol{v}_{n+1} - \boldsymbol{v}_n\| \leq \epsilon(1-\gamma)/2\gamma, \qquad \forall n \geq N. \tag{5.4}$$

- *The policy* π_ϵ *that takes action* $\arg\max_a r(s) + \gamma \sum_j p(j|s,a)\boldsymbol{v}_n(s')$ *is* ϵ*-optimal.*
- $\|\boldsymbol{v}_{n+1} - V^*_\mu\| < \epsilon/2$ *for* $n > N$.

Proof.

The first two statements follow from the fixed point theorem. Now note that

$$\|V^{\pi_\epsilon}_\mu - V^*_\mu\| \leq \|V^{\pi_\epsilon}_\mu - \boldsymbol{v}_n\| + \|\boldsymbol{v}_n - V^*_\mu\|$$

We can bound these two terms easily:

$$\|V^{\pi_\epsilon} - \boldsymbol{v}_{n+1}\| \leq \frac{\gamma}{1-\gamma}\|\boldsymbol{v}_{n+1} - \boldsymbol{v}_n\|, \qquad \|\boldsymbol{v}_{n+1} - V^*_\mu\| \leq \frac{\gamma}{1-\gamma}\|\boldsymbol{v}_{n+1} - \boldsymbol{v}_n\|$$

□

Theorem 12

Value iteration converges linearly at rate γ *and* $O(\gamma^n)$*. In addition, for* $r \in [0, 1]$ *and* $\boldsymbol{r}^0 = \mathbf{0}$

$$\|\boldsymbol{v}_n - V^*_\mu\| \leq \frac{\gamma^n}{1-\gamma}$$

$$\|V^{\pi_n}_\mu - V^*_\mu\| \leq \frac{2\gamma^n}{1-\gamma},$$

Algorithm 6 Policy iteration

Input μ, $\mathcal{S}$.
Initialise $\boldsymbol{v}_0$.
for $n = 1, 2, \ldots$ **do**
  $\pi_{n+1} = \arg\max_\pi \boldsymbol{r} + \gamma \boldsymbol{P}_\pi \boldsymbol{v}_n$ (policy improvement)
  $\boldsymbol{v}_{n+1} = V_\mu^{\pi_{n+1}}$ (policy evaluation)
  break if $\pi_{n+1} = \pi_n$.
end for
Return $\pi_n, \boldsymbol{v}_n$.

Theorem 13

If $\boldsymbol{v}_n, \boldsymbol{v}_{n+1}$ are produced by policy iteration, then $\boldsymbol{v}_n \leq \boldsymbol{v}_{n+1}$.

Proof.

From the policy improvement step

$$\boldsymbol{r} + \gamma \boldsymbol{P}_{\pi_{n+1}} \boldsymbol{v}_n \geq \boldsymbol{r} + \gamma \boldsymbol{P}_{\pi_n} \boldsymbol{v}_n = \boldsymbol{v}_n$$

where the equality is due to the fact that $(\boldsymbol{I} - \gamma \boldsymbol{P}_{\mu,\pi_n})\boldsymbol{v}_n = \boldsymbol{r}$ from the policy evaluation step. Rearranging, we get that

$$\begin{aligned} \boldsymbol{r} &\geq (\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi_{n+1}})\boldsymbol{v}_n \\ (\boldsymbol{I} - \gamma \boldsymbol{P}_{\pi_{n+1}})^{-1} \boldsymbol{r} &\geq \boldsymbol{v}_n, \end{aligned}$$

noting that the inverse is positive. Since the left side equals $\boldsymbol{v}_{n+1}$, we have proved the theorem. □

Corollary 14

If $\mathcal{S}, \mathcal{A}$ are finite then policy iteration terminates in a finite number of iterations.

Modified policy iteration

Algorithm 7 Modified policy iteration

Input μ, $\mathcal{S}$.
Initialise $\boldsymbol{v}_0$.
for $n = 1, 2, \ldots$ **do**
  $\pi_n = \arg\max_\pi \boldsymbol{r} + \gamma \boldsymbol{P}_\pi \boldsymbol{v}_{n-1}$ `// policy improvement`
  $\boldsymbol{v}_n = \mathscr{L}^k_{\pi_n} \boldsymbol{v}_{n-1}$ `// partial policy evaluation`
  break if $\pi_n = \pi_{n=1}$.
end for
Return $\pi_n, \boldsymbol{v}_n$.

Geometric view

Definition 15

Difference operator

$$\mathscr{B}\boldsymbol{v} \triangleq \max_{\pi} \{\boldsymbol{r} + (\gamma \boldsymbol{P}_{\pi} - \boldsymbol{I})\boldsymbol{v}\} = \mathscr{L}\boldsymbol{v} - \boldsymbol{v}. \tag{5.5}$$

Hence the optimality equation becomes

$$\mathscr{B}\boldsymbol{v} = 0. \tag{5.6}$$

.

Algorithm 8 Temporal-Difference Policy Iteration

Input μ, $\mathcal{S}, \lambda$.
Initialise $\boldsymbol{v}_0$.
for $n = 1, 2, \ldots$ **do**
  $\pi_n = \arg\max_\pi \boldsymbol{r} + \gamma \boldsymbol{P}_\pi \boldsymbol{v}_{n-1}$ `// policy improvement`
  $\boldsymbol{v}_n = \boldsymbol{v}_{n-1} + \boldsymbol{\tau}_k$ `// temporal difference update.`
  break if $\pi_n = \pi_{n=1}$.
end for
Return $\pi_n, \boldsymbol{v}_n$.

$$\mathscr{L}_{\pi_{n+1}} \boldsymbol{v}_n = \mathscr{L} \boldsymbol{v}_n. \tag{5.7}$$

$$d_n(i, j) = \boldsymbol{v}_n(i) - [\boldsymbol{r}(i) + \gamma \boldsymbol{v}_n(j)]. \qquad \text{(temporal difference error)}$$

$$\boldsymbol{\tau}_n(i) = \sum_{t=0}^{\infty} \mathbb{E}_{\pi_n, \mu} \left[(\gamma\lambda)^m d_n(s_t, s_{t+1}) \mid s_0 = i \right] \tag{5.8}$$

$$\boldsymbol{v}_{n+1} = \boldsymbol{v}_n + \boldsymbol{\tau}_n \tag{5.9}$$

$$\mathscr{D}_n \boldsymbol{v} \triangleq (1 - \lambda) \mathscr{L}_{\pi_{n+1}} \boldsymbol{v}_n + \lambda \mathscr{L}_{\pi_{n+1}} \boldsymbol{v}. \qquad \text{(fixed point)}$$

Select $\boldsymbol{y} \in \mathbb{S}^{|\mathcal{S}|}$ (i.e. a state distribution). Then:

Primar linear program

$$\min_{\boldsymbol{v}} \boldsymbol{y}^\top \boldsymbol{v}$$

such that

$$\boldsymbol{v}(s) - \gamma \boldsymbol{p}_{s,a}^\top \boldsymbol{v} \geq r(s, a), \qquad \forall a \in \mathcal{A}, s \in \mathcal{S}.$$

Dual linear program

$$\max_x \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) x(s, a)$$

such that $x \in \mathbb{R}_+^{|\mathcal{S} \times \mathcal{A}|}$ and

$$\sum_{a \in \mathcal{A}} x(j, a) - \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \gamma p(j|s, a) x(s, a) = y(j).$$

with $\boldsymbol{y} \in \mathbb{S}^{|\mathcal{S}|}$.

Summary

Markov decision processes

Can represent : Shortest path problems, Stopping problems, Experiment design problems, Multi-armed bandit problems, Reinforcement learning problems.

Backwards induction (aka value iteration)

- In the class of dynamic programming algorithms.
- Tractable when either the state space $\mathcal{S}$ or the horizon T are small (finite).

Optimal decisions and Bayesian reinforcement learning

- A known environment is represented as an MDP.
- Bandit problems can be solved by representing them as infinite-state MDPs.
- In general, an unknown environment can be represented as a distribution over MDPs.
- The decision problem can again be formulated as an infinite-state MDP.

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002.

[2] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2001.

[3] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[4] Herman Chernoff. Sequential design of experiments. *Annals of Mathematical Statistics*, 30(3):755–770, 1959.

[5] Herman Chernoff. Sequential Models for Clinical Trials. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.4*, pages 805–812. Univ. of Calif Press, 1966.

[6] Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 1970.

[7] Marting L. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New Jersey, US, 1994.

[8] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML 2010*, 2010.

[9] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.