# Dynamic Programming for Markov Decision Process II



Notation:

- upercase (e.g. $R_t$): random variable
- lowercase (e.g. $s_t$): instance of random variable
- bold straight (e.g. $\mathbf{v}$): determineistic vector

Preliminary: Dynamic Programming for Markov Decision Process I

## Generalization

There are multiple ways to generalize the MDP:

- deterministic policy $\pi(s)$ $\longrightarrow$ stochastic policy $\pi(a \mid s)$.
- deterministic rewards $r(s,a)$ $\longrightarrow$ stochastic rewards $p(r \mid s,a)$.
- infinite time horizon and time invariance $\longrightarrow$ finite time horizon and time dependence.

In the following, we focus on generalization to stochastic policy and stochastic rewards.

### Stochastic Policy

Instead of mapping each state to a fixed action, a stochastic policy samples an action from the conditional distribution $\pi(\cdot \mid s)$.

$$A \sim \pi(a \mid s) \tag{1}$$

Remarks:

- Stochastic policy satisfies the law of total probability.

$$\sum_{a\in\mathcal{A}} \pi(a \mid s) = 1,\ \forall s \in \mathcal{S}\ (\text{for finite } \mathcal{A}) \tag{2}$$

$$\int_{a\in\mathcal{A}} \pi(a \mid s)\,\mathrm{d}a = 1,\ \forall s \in \mathcal{S}\ (\text{for continuous } \mathcal{A}) \tag{3}$$

- A determinstic policy $s \mapsto a_s$ can be seen as a special of stochastic policy

$$\pi(a \mid s) = \delta(a - a_s), \quad \forall a \in \mathcal{A} \tag{4}$$

  where $\delta(\cdot)$ is Kronecker delta (for finite $\mathcal{A}$) or Dirac delta (for continuous $\mathcal{A}$).

- Under this setting, our stochastic policy is time-invariant (or ***stationary***). i.e. The distribution $\pi(\cdot \mid s)$ does not depend on when we arrived at $s$.
- If both $\mathcal{A}$ and $\mathcal{S}$, we can represent a stochastic policy as a $|\mathcal{S}| \times |\mathcal{A}|$ matrix. Each entry $\pi(a \mid s)$ represents the proability of taking action $a$ at state $s$.

**Example**: An extremely simplifed model of stock trading. $\mathcal{S}$ denotes the market condition. $\mathcal{A}$ denotes your trading action.

$$\mathcal{S} = \{\text{bullish}, \text{bearish}\}$$
$$\mathcal{A} = \{\text{buy}, \text{hold}, \text{sell}\}$$

A trading strategy inspired by the principle *"Be fearful when others are greedy and to be greedy only when others are fearful"* can be modelled as a stochastic policy:

| Market Condition | buy | hold | sell |
|---|---|---|---|
| **bullish** | 0.2 | 0.4 | 0.4 |
| **bearish** | 0.6 | 0.3 | 0.1 |

## Why stochastic policy?

If the agent has perfect knowledge of the environment (state transiton, rewards, etc.), then any deviation from the optimal policy (which is deterministic, as shown later) will indeed yield a lower total reward. In this case, a deterministic policy suffices. The agent only needs to compute the optimal policy -- this is called ***planning problem***. e.g. shortest path problem in a graph with known weights.

In contrast, a ***learning problem*** assumes that the agent does not have the perfect environment model. e.g. shortest path problem in a graph with unknown edge weights. The optimal policy derived from an imperfect environment model is unlikely to be optimal in real environment. In this case, stochastic policy allows the agent to explore the environment by occasionally try different actions, even at a risk of receiving a lower total reward.

## Exploitation vs. Exploration

- ***Exploitation***: execute the optimal policy based on current understanding of the environment. It yields the highest total reward based on current understanding but misses the oppurtunity for a potentially higher total reward.
- ***Exploration***: deviate from the optimal policy by occasionally trying different actions. It may yield a lower or higher total reward compared to exploitation.

Real-life analogy: Suppose you're looking for good restaurants in your city. You've found that restaurant A is your favorite spot for lunch and restaurant B is your favorite for dinner. Exploitation means you always return to A for lunch and B for dinner — you're likely to be satisfied, but you miss the chance to discover new favorites. Exploration means occasionally trying a new place, say restaurant X. It might not be as good as A, or it could turn out to be even better.

In planning problems, only exploitation is necessary because the environment is fully known. In learning problems, however, it is crucial to balance exploitation and exploration.

## Stochastic Rewards

Given a state action pair $(s, a)$, instead of receving a deterministic reward, the agent receives a stochastic reward sampled from $p(\cdot \mid s, a)$.

$$R \sim p(r \mid s, a)$$

Moreover, we let $\mathcal{R} \subseteq \mathbb{R}$ denote the reward sets. Then, $p(r \mid s, a)$ is either a PMF (if $\mathcal{R}$ is finite) or a PDF (if $\mathcal{R}$ is continuous).

Remarks:

- Stochastic reward satisfies the law of total probability

$$\sum_{r \in \mathcal{R}} p(r \mid s, a) = 1, \ \forall (s, a) \in \mathcal{S} \times \mathcal{A} \tag{5}$$

$$\text{or} \int_{r \in \mathcal{R}} p(r \mid s, a) \, \mathrm{d}r = 1, \ \forall (s, a) \in \mathcal{S} \times \mathcal{A} \tag{6}$$

- A determinstic reward $(s, a) \mapsto r_{sa}$ can be seen as a special of stochastic reward

$$p(r \mid s, a) = \delta(r - r_{sa}), \quad \forall r \in \mathcal{R} \tag{7}$$

  where $\delta(\cdot)$ is Kronecker delta (for finite $\mathcal{R}$) or Dirac delta (for continuous $\mathcal{R}$).
- The stochastic reward is again stationary. i.e. The distribution of the reward does not depend on when we arrive at $s$ or when we take action $a$.

# Markov Decision Process

A generalized Markov decision process (MDP) consists of

- $\mathcal{S}$: set of states.
- $\mathcal{A}$: set of actions.
- $\mathcal{R}$: set of rewards.
- $p(s' \mid s, a)$: state transition probability.
- $p(r \mid s, a)$: reward probability.
- $\gamma \in [0, 1)$: discount factor.

The agent executes a stochastic policy $\pi$. Then, we are interested in

1. policy evaluation
2. computing optimal policy

Consider the execution of $\pi$ from a fixed initial state $s$.

$$s \xrightarrow[R]{A} S' \xrightarrow[R']{A'} \cdots$$

Key observation:

- The immediate reward $R$ is now stochastic. The stochasticity of immediate reward comes from
  1. stoachsticity in action $A$ due to stochastic policy
  2. reward probability

- Starting from the next state, the stochasticity in $R'$ comes from
  1. stoachsticity in state $S'$ due to state transition proability
  2. stoachsticity in action $A'$ due to stochastic policy
  3. reward probability

The total (discounted) reward is defined exactly the same as in MDP I:

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots \tag{8}$$
$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k} \tag{9}$$

Remarks:

- For each $\tau = t, t+1, \cdots$, $R_\tau \sim p(r \mid S_\tau, A_\tau)$.
- Recursive structure:

$$G_t = R_t + \gamma G_{t+1} \tag{10}$$

## Value functions

State value:

$$v_\pi(s) = \mathbb{E}\left[G_t \mid S_t = s\right], \quad \forall s \in \mathcal{S} \tag{11}$$

Q-function:

$$q_\pi(s, a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right], \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{12}$$

Remarks:

- $v_\pi(s) \triangleq$ the expected total reward of following a stochastic policy $\pi$ fron now onwards.
- $q_\pi(s, a) \triangleq$ the expected total reward of taking a **deterministic** action $a$ now and then following a stochastic policy $\pi$.
- Relation between $q_\pi$ and $v_\pi$:

$$v_\pi(s) = \mathbb{E}_{a \sim \pi(a|s)}[q_\pi(s, a)]$$
$$q_\pi(s, a) = \mathbb{E}_{r \sim p(r|s,a)}[r] + \gamma \mathbb{E}_{s' \sim p(s'|s,a)}[v_\pi(s')]$$

  - Proof: c.f. section *Bellman Equations for Q-Function*

Suppose the agent execute some policy $\pi$ on initial state $s$ and then follow another policy $\pi'$ onwards. Then, the expected total reward is

$$\mathbb{E}\left[G_t \mid S_t = s\right] = \mathbb{E}_{a \sim \pi(a|s)}\left[q_{\pi'}(s, a)\right] \tag{13}$$

If $\pi$ happens to be deterministic (i.e. $a = \pi(s)$), the total reward becomes

$$\mathbb{E}\left[G_t \mid S_t = s\right] = q_{\pi'}(s, \pi(s)) \tag{14}$$

Illustration:

$$s \xrightarrow[R]{A \sim \pi(a|s)} S' \xrightarrow[R']{A' \sim \pi'(a|s)} \cdots$$

# Bellman Equations

## Bellman Equations for State Value

Bellman equation for $v_\pi$: For each $s \in \mathcal{S}$, it holds that

$$v_\pi(s) = \mathbb{E}_{a\sim\pi(a|s)}\Big[\mathbb{E}_{r\sim p(r|s,a)}[r]\Big] + \gamma \mathbb{E}_{a\sim\pi(a|s)}\Big[\mathbb{E}_{s'\sim p(s'|s,a)}[v_\pi(s')]\Big] \tag{BE-V}$$

Equivalent formulation:

$$v_\pi(s) = \mathbb{E}_{a\sim\pi(a|s)}\Big[\mathbb{E}_{r\sim p(r|s,a)}[r] + \gamma\mathbb{E}_{s'\sim p(s'|s,a)}[v_\pi(s')]\Big] \tag{BE-V1}$$

$$v_\pi(s) = \mathbb{E}_{r\sim p(r|s)}[r] + \gamma\mathbb{E}_{s'\sim p(s'|s)}[v_\pi(s')] \tag{BE-V2}$$

For finite $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$, Bellman equation becomes

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_r p(r \mid s,a)\cdot r + \gamma \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s,a)\cdot v_\pi(s') \tag{15}$$

$$= \sum_a \pi(a \mid s) \left[\sum_r p(r \mid s,a)\cdot r + \gamma \sum_{s'} p(s' \mid s,a)\cdot v_\pi(s')\right] \tag{16}$$

Remarks:

- Bellman equation holds for all $s \in \mathcal{S}$. For finite state space, there are $|\mathcal{S}|$ equations in total. There are further equivalent formulation of Bellman equations in the Appendix.
- In $(\text{BE-V})$: The 1st term represents the expected immediate reward. The 2nd term in $(\text{BE-V})$ represents the discounted expected future reward.
- In $(\text{BE-V2})$: The distributions $p(r \mid s)$ and $p(s' \mid s)$ are often denoted by $p_\pi(r \mid s)$ and $p_\pi(s' \mid s)$ since they both implicitly depend on $\pi$. Specifically, they can be obtained from system parameters by applying the law of total probability (c.f. Appendix)

  $$p_\pi(r \mid s) = \mathbb{E}_{a\sim\pi(a|s)}\big[p(r \mid s,a)\big] \tag{17}$$

  $$p_\pi(s' \mid s) = \mathbb{E}_{a\sim\pi(a|s)}\big[p(s' \mid s,a)\big] \tag{18}$$

- Illustration:

$$s \xrightarrow[R_t]{A_t} S_{t+1} \xrightarrow[R_{t+1}]{A_{t+1}} \cdots$$

*Proof*: We first show the validity of $(\text{BE-V2})$. Then, we show $(\text{BE-V})$ and $(\text{BE-V1})$ are equivalent to $(\text{BE-V2})$.

Plugging $G_t = R_t + \gamma G_{t+1}$ into the value function, we get

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[R_t + \gamma G_{t+1} \mid S_t = s] \\ &= \underbrace{\mathbb{E}[R_t \mid S_t = s]}_{\text{(I)}} + \gamma \underbrace{\mathbb{E}[G_{t+1} \mid S_t = s]}_{\text{(II)}} \end{aligned}$$

$(\text{I}) \triangleq$ expected immediate reward:

$$\underbrace{\mathbb{E}[R_t \mid S_t = s]}_{\text{(I)}} = \mathbb{E}_{r\sim p(r|s)}[r]$$

$(\mathrm{II}) \triangleq$ expected future reward:

$$\begin{aligned}\underbrace{\mathbb{E}[G_{t+1} \mid S_t = s]}_{(\mathrm{II})} &= \mathbb{E}_{s' \sim p(s'|s)}\Big[\mathbb{E}[G_{t+1} \mid S_t = s, S_{t+1} = s']\Big] \\ &= \mathbb{E}_{s' \sim p(s'|s)}\Big[\mathbb{E}[G_{t+1} \mid S_{t+1} = s']\Big] \\ &= \mathbb{E}_{s' \sim p(s'|s)}[v_\pi(s')]\end{aligned}$$

Combining $(\mathrm{I})$ and $(\mathrm{II})$, we obtain $(\mathrm{BE\text{-}V2})$:

$$v_\pi(s) = \underbrace{\mathbb{E}_{r \sim p(r|s)}[r]}_{(\mathrm{I})} + \gamma \underbrace{\mathbb{E}_{s' \sim p(s'|s)}[v_\pi(s')]}_{(\mathrm{II})}$$

$(\mathrm{BE\text{-}V})$ and $(\mathrm{BE\text{-}V1})$ follow from $(\mathrm{BE\text{-}V2})$ and the law of total expecation (c.f. Appendix) as

$$\begin{aligned}\underbrace{\mathbb{E}_{r \sim p(r|s)}[r]}_{(\mathrm{I})} &= \mathbb{E}_{a \sim \pi(a|s)}\Big[\mathbb{E}_{r \sim p(r|s,a)}[r]\Big] \\ \underbrace{\mathbb{E}_{s' \sim p(s'|s)}[v_\pi(s')]}_{(\mathrm{II})} &= \mathbb{E}_{a \sim \pi(a|s)}\Big[\mathbb{E}_{s' \sim p(s'|s,a)}[v_\pi(s')]\Big]\end{aligned}$$

■

## Vector-Form Bellman Equation

Recall the Bellman equation for finite $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_r p(r \mid s, a) \cdot r + \gamma \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \cdot v_\pi(s')$$

Switching the orders of sums in both terms, we get

$$\begin{aligned}v_\pi(s) &= \sum_r \underbrace{\sum_a \pi(a \mid s) p(r \mid s, a)}_{p_\pi(r|s)} \cdot r + \gamma \sum_{s'} \underbrace{\sum_a \pi(a \mid s) p(s' \mid s, a)}_{p_\pi(s'|s)} \cdot v_\pi(s') \\ &= \underbrace{\sum_r p_\pi(r \mid s) \cdot r}_{r_\pi(s)} + \gamma \sum_{s'} p_\pi(s' \mid s) \cdot v_\pi(s')\end{aligned}$$

Assume $\mathcal{S} = \{\varsigma_1, \ldots, \varsigma_{|\mathcal{S}|}\}$. Consider $v_\pi(\varsigma_i), i = 1, \ldots, |\mathcal{S}|$ as unknowns. For a certain $\pi$, the 1st term $r_\pi(s)$ is just a known number. The 2nd term is a linear combination of $\varsigma_1, \ldots, \varsigma_{|\mathcal{S}|}$. Therefore, we have

$$\underbrace{\begin{bmatrix} v_\pi(\varsigma_1) \\ v_\pi(\varsigma_2) \\ \vdots \\ v_\pi(\varsigma_n) \end{bmatrix}}_{\mathbf{v}_\pi} = \underbrace{\begin{bmatrix} r_\pi(\varsigma_1) \\ r_\pi(\varsigma_2) \\ \vdots \\ r_\pi(\varsigma_n) \end{bmatrix}}_{\mathbf{r}_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(\varsigma_1 \mid \varsigma_1) & \ldots & p_\pi(\varsigma_n \mid \varsigma_1) \\ p_\pi(\varsigma_1 \mid \varsigma_2) & \ldots & p_\pi(\varsigma_n \mid \varsigma_2) \\ \vdots & \cdots & \vdots \\ p_\pi(\varsigma_1 \mid \varsigma_n) & \ldots & p_\pi(\varsigma_n \mid \varsigma_n) \end{bmatrix}}_{\mathbf{P}_\pi} \cdot \underbrace{\begin{bmatrix} v_\pi(\varsigma_1) \\ v_\pi(\varsigma_2) \\ \vdots \\ v_\pi(\varsigma_n) \end{bmatrix}}_{\mathbf{v}_\pi}$$

> Bellman equation (vector form)
>
> $$\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_\pi \tag{19}$$

Remarks:

- In finite settings, the Bellman equation has again the same vector form, even though policy and rewards are stochastic. Here, the $\mathbf{r}_\pi$ is a vector of expected immediate rewards w.r.t. $r \sim p_\pi(r \mid s)$, which is the statistical average of $p_\pi(r \mid s, a)$ over $a \sim \pi(a \mid s)$. The row-stochastic matrix $\mathbf{P}_\pi$ consists of state transition probabilities $p_\pi(s' \mid s)$ under $\pi$, which is the statistical average of $p_\pi(s' \mid s, a)$ over $a \sim \pi(a \mid s)$

.

- The vector-form can again be solve analytically. However, analytical solution is rarely used unless the state space is small. In practice, iterative method is more frequently used (detailed later).

## Bellman Equations for Q-Function

Relation between $q_\pi$ and $v_\pi$:

$$v_\pi(s) = \mathbb{E}_{a\sim\pi(a|s)}[q_\pi(s,a)] \tag{Q2V}$$
$$q_\pi(s,a) = \mathbb{E}_{r\sim p(r|s,a)}[r] + \gamma \mathbb{E}_{s'\sim p(s'|s,a)}[v_\pi(s')] \tag{V2Q}$$

Bellman equation for $q_\pi$:

$$q_\pi(s,a) = \mathbb{E}_{r\sim p(r|s,a)}[r] + \gamma \mathbb{E}_{s'\sim p(s'|s,a)}\Big[\mathbb{E}_{a'\sim\pi(a'|s')}[q_\pi(s',a')]\Big] \tag{BE-Q}$$

*Proof*: Equation $(\mathrm{Q2V})$ follows from the law of total expecation:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t \mid S_t = s] \\ &= \mathbb{E}_{a\sim\pi(a|s)}\Big[\underbrace{\mathbb{E}[G_t \mid S_t = s, A_t = a]}_{q_\pi(s,a)}\Big] \end{aligned}$$

■

To show $(\mathrm{V2Q})$, we use $G_t = R_t + \gamma G_{t+1}$:

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}[R_t + \gamma G_{t+1} \mid S_t = s, A_t = a] \\ &= \underbrace{\mathbb{E}[R_t \mid S_t = s, A_t = a]}_{(\mathrm{I})} + \gamma \underbrace{\mathbb{E}[G_{t+1} \mid S_t = s, A_t = a]}_{(\mathrm{II})} \end{aligned}$$

The expected immediate reward $(\mathrm{I})$ is

$$\underbrace{\mathbb{E}[R_t \mid S_t = s, A_t = a]}_{(\mathrm{I})} = \mathbb{E}_{r\sim p(r|s,a)}[r]$$

The expected future reward $(\mathrm{II})$ is

$$\begin{aligned} \underbrace{\mathbb{E}[G_{t+1} \mid S_t = s, A_t = a]}_{(\mathrm{II})} &= \mathbb{E}_{s'\sim p(s'|s,a)}\Big[\mathbb{E}[G_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s']\Big] \\ &= \mathbb{E}_{s'\sim p(s'|s,a)}\Big[\mathbb{E}[G_{t+1} \mid S_{t+1} = s']\Big] \\ &= \mathbb{E}_{s'\sim p(s'|s,a)}[v_\pi(s')] \end{aligned}$$

Combining $(\mathrm{I})$ and $(\mathrm{II})$, we obtain $(\mathrm{Q2V})$:

$$v_\pi(s) = \underbrace{\mathbb{E}_{r\sim p(r|s,a)}[r]}_{(\mathrm{I})} + \gamma \underbrace{\mathbb{E}_{s'\sim p(s'|s,a)}[v_\pi(s')]}_{(\mathrm{II})}$$

■

Plugging $(\mathrm{Q2V})$ into $(\mathrm{V2Q})$ yields $(\mathrm{BE\text{-}Q})$. ■

# Policy Evaluation

In this section, we assume that $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$ are all finite. Under this setting, Bellman equation can be solved (either analytically or numerically) entirely based on the system model.

In contrast, when $\mathcal{S}$ is continuous, the Bellman equation becomes an integral equation of $v_\pi$. It is challanging to derive model-based algorithms to solve Bellman equation without additional assumptions (e.g. all random variables are Gaussian). This difficulty arises because the expectations in the Bellman equation correspond to high-dimensional integrals, which are generally intractable to compute exactly. To address this difficulty, other techniques like function approximation are used (not detailed here).

Algorithm to compute state values:

> **BELLMAN UPDATE (vector form)**
> Initialize $\mathbf{v}_0$ arbitrarily.
> For $n = 0, 1, \ldots$, run until $\mathbf{v}_n$ converges
>
> $$\mathbf{v}_{n+1} = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_n \tag{20}$$

Equivalent element-wise form:

> **BELLMAN UPDATE (element-wise form)**
> Init $v_0(s)$ for all $s \in \mathcal{S}$
> For $n = 0, 1, \ldots$, run until $v_n(s)$ converges
> For each $s \in \mathcal{S}$, do
>
> $$v_{n+1}(s) = \sum_r p_\pi(r \mid s) \cdot r + \gamma \sum_{s'} p_\pi(s' \mid s) \cdot v_\pi(s')$$

Remarks:

- The correctness of Bellman update here follows the same principle as in MDP I. The stochasitcity of policy and rewards does not alter the contractiveness of the affine mapping $f(\mathbf{v}) = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}$, even though the definition of $\mathbf{r}_\pi$ and $\mathbf{P}_\pi$ are slightly different from those in MDP I. A detailed proof is omitted.
- In practice, $p_\pi(r \mid s)$ and $p_\pi(s' \mid s)$ can be pre-computed by

$$p_\pi(r \mid s) = \sum_a \pi(a \mid s) \cdot p(r \mid s, a)$$
$$p_\pi(s' \mid s) = \sum_a \pi(a \mid s) \cdot p(s' \mid s, a)$$

# Bellman Optimality Equation

## Optimal Policies and Optimal Value Functions

Optimal policy $\pi^*$:

$$\pi^* \triangleq \arg\max_\pi v_\pi(s) \tag{21}$$

Optimal value function $v^*$:

$$v^*(s) \triangleq v_{\pi^*}(s) = \max_\pi v_\pi(s) \tag{22}$$

Optimal Q-function $q^*$:

$$q^*(s, a) \triangleq q_{\pi^*}(s, a) = \max_\pi q_\pi(s, a) \tag{23}$$

Remark:

- The optimization variable $\pi$ is a probability distribution over $\mathcal{A}$. Optimization problems with PDF as variables are nasty to deal with. However, we will see later that $\pi^*$ is actually deterministic and thus can be written as delta function.

Equivalent formulation of $v^*$:

$$\begin{aligned} v^*(s) &= \max_{\pi} \mathbb{E}_{a\sim\pi(a|s)}[q_\pi(s,a)] \\ &= \mathbb{E}_{a\sim\pi^*(a|s)}[q^*(s,a)] \\ &= \max_{\pi} \mathbb{E}_{a\sim\pi(a|s)}[q^*(s,a)] \end{aligned} \tag{24}$$

Remarks:

- The 1st reformulation is not useful to solve $v^*$ and $\pi^*$ since the optmization variable $\pi$ apears in both expectation and in Q-function.
- The 2nd reformulation is again not useful for solving $v^*$ and $\pi^*$ since it explicitly refers to $\pi^*$.
- The 3rd reformulation serves as the starting point to solve $v^*$ and $\pi^*$. Here, the optimization variable $\pi$ only appears in the expectation. The optimal Q-function inside the expectation does not require direct referencing to $\pi^*$ as $q^*(s,a) = \mathbb{E}_{r\sim p(r|s,a)}[r] + \gamma \mathbb{E}_{s'\sim p(s'|s,a)}[v^*(s')]$

*Proof*: The 1st reformulation follows from the definition of $v^*$ and the relation btw. $v_\pi$ and $q_\pi$. The 2nd reformulation follows from the relation btw. $v^*$ and $q^*$. It remains to show the 3rd reformulation.

Note that $\forall \pi, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$,

$$q_\pi(s,a) \leq q^*(s,a) \leq v^*(s)$$

where the 1st inequality follows from the optimality of $q^*$ while the 2nd inequality follows from the optimaliy of $v^*$. Then:

On one hand,

$$v^*(s) = \max_{\pi} \mathbb{E}_{a\sim\pi(a|s)}[q_\pi(s,a)] \leq \max_{\pi} \mathbb{E}_{a\sim\pi(a|s)}[q^*(s,a)]$$

On the other hand,

$$\mathbb{E}_{a\sim\pi(a|s)}[q^*(s,a)] \leq v^*(s), \forall \pi \implies \max_{\pi} \mathbb{E}_{a\sim\pi(a|s)}[q^*(s,a)] \leq v^*(s)$$ ■

## Bellman Optimality Equation for State Value

**Bellman Optimality Equation** (element-wise form)

$$v^*(s) = \max_a q^*(s,a) \tag{25}$$

$$= \max_a \left\{ \mathbb{E}_{r\sim p(r|s,a)}[r] + \gamma \mathbb{E}_{s'\sim p(s'|s,a)}[v^*(s')] \right\} \tag{26}$$

Optimal Policy:

$$\pi^*(a \mid s) = \delta(a - a^*) \quad \text{where } a^* = \arg\max_{a\in\mathcal{A}} q^*(s,a) \tag{27}$$

Remarks:

- BOE has again the recursive structure.

- An optimal policy is **deterministic**. It chooses the action maximizing the optimal Q-function. If there are multiple actions maximizing $q^*$, we may choose any of those maximizers. Hence, the optimal policy is **not unique**.
- The optimal value function $v^*$ is however unique due to contractive mapping theorem. (Just the same as MDP I)

*Proof*:
Consider a non-stationary policy $\pi'$ constructed as follows

1. $\pi'$ **deterministically** picks $\hat{a} \triangleq \arg\max_{a \in \mathcal{A}} q^*(s, a)$ at initial state $s$.
2. $\pi'$ thereafter follows the optimal policy $\pi^*$

The value function of $\pi'$ is thus

$$v_{\pi'}(s) = q^*(s, \hat{a}) = \max_{a \in \mathcal{A}} q^*(s, a)$$

To show: $v^*(s) = \max_a q^*(s, a) \iff$ To show: $v^*(s) = v_{\pi'}(s)$.

Since $\pi^*$ is an optimal policy, it is obvious that

$$v^*(s) \geq v_{\pi'}(s)$$

On the other side, we have by the trivial bound of expectation (c.f. Appendix)

$$v^*(s) = \max_{\pi} \mathbb{E}_{a \sim \pi(a|s)}[q^*(s, a)] \leq \max_a q^*(s, a) = v_{\pi'}(s)$$

with equality if

$$\pi(a \mid s) = \delta\left(a - \hat{a}\right) \quad \text{where } \hat{a} = \arg\max_{\tilde{a} \in \mathcal{A}} q^*(s, \tilde{a})$$ ■

# Appendix

## Law of total probability

$$p(x) = \mathbb{E}_{\theta \sim p(\theta)}\left[p(x \mid \theta)\right] \tag{28}$$
$$p(x \mid z) = \mathbb{E}_{\theta \sim p(\theta|z)}\left[p(x \mid z, \theta)\right] \tag{29}$$

Alternative notation:

$$p(x) = \mathbb{E}[p(x \mid \Theta)] \tag{30}$$
$$p(x \mid z) = \mathbb{E}[p(x \mid z, \Theta)] \tag{31}$$

**Core Inituition**: computing a probability can be seen as

1. introducing an intermediate variable $\theta$,
2. computing the conditional probability,
3. and then averaging the conditional probability over $\theta$.

*Proof*: For the sake of simplicity, we assume all random variables are continuous. Otherwise, replace the integral with sum. By the law of marginal distribution,

$$p(x) = \int_\theta p(x, \theta)\, \mathrm{d}\theta = \int_\theta p(\theta) \cdot p(x \mid \theta)\, \mathrm{d}\theta = \mathbb{E}_{\theta \sim p(\theta)}\left[p(x \mid \theta)\right]$$ ■

For $p(x \mid z)$, we have

$$
\begin{aligned}
p(x \mid z) &= \frac{p(x,z)}{p(z)} && \text{def. of } p(x \mid z) \\
&= \frac{1}{p(z)} \int_\theta p(x,z,\theta)\,\mathrm{d}\theta && \text{marginalization} \\
&= \frac{1}{p(z)} \int_\theta p(z)\cdot p(\theta \mid z)\cdot p(x \mid z,\theta)\,\mathrm{d}\theta && \text{chain rule} \\
&= \int_\theta p(\theta \mid z)\cdot p(x \mid z,\theta)\,\mathrm{d}\theta && \text{cancel out } p(z) \\
&= \mathbb{E}_{\theta\sim p(\theta|z)}\big[p(x \mid z,\theta)\big] && \text{def. of } \mathbb{E}_{\theta\sim p(\theta|z)}[\,\cdot\,] \qquad \blacksquare
\end{aligned}
$$

## Law of total expectation

$$
\mathbb{E}_{x\sim p(x)}[g(x)] = \mathbb{E}_{\theta\sim p(\theta)}\Big[\mathbb{E}_{x\sim p(x|\theta)}[g(x)]\Big] \tag{32}
$$

$$
\mathbb{E}_{x\sim p(x|z)}[g(x)] = \mathbb{E}_{\theta\sim p(\theta|z)}\Big[\mathbb{E}_{x\sim p(x|z,\theta)}[g(x)]\Big] \tag{33}
$$

Alternative notation:

$$
\mathbb{E}[g(X)] = \mathbb{E}\Big[\mathbb{E}[g(X) \mid \Theta]\Big] \tag{34}
$$

$$
\mathbb{E}[g(X) \mid z] = \mathbb{E}\Big[\mathbb{E}[g(X) \mid z, \Theta]\Big] \tag{35}
$$

**Core Inituition**: computing an expectation can be seen as

1. introducing an intermediate variable $\theta$,
2. computing the conditional expectation,
3. and then averaging the conditional expecation over $\theta$.

*Proof*: For the sake of simplicity, we assume all random variables are continuous. Otherwise, replace the integral with sum. For $\mathbb{E}_{x\sim p(x)}[g(x)]$, we have

$$
\begin{aligned}
\mathbb{E}_{x\sim p(x)}[g(x)] &= \int_x g(x)\cdot p(x)\,\mathrm{d}x && \text{def. of } \mathbb{E}_{x\sim p(x)}[\,\cdot\,] \\
&= \int_x g(x)\left(\int_\theta p(\theta)\cdot p(x \mid \theta)\,\mathrm{d}\theta\right)\mathrm{d}x && \text{by } p(x) = \mathbb{E}_{\theta\sim p(\theta)}\big[p(x \mid \theta)\big] \\
&= \int_\theta p(\theta)\int_x g(x)\cdot p(x \mid \theta)\,\mathrm{d}x\,\mathrm{d}\theta && \text{switch the order} \\
&= \int_\theta p(\theta)\,\mathbb{E}_{x\sim p(x|\theta)}[g(x)]\,\mathrm{d}\theta && \text{def. of } \mathbb{E}_{x\sim p(x|\theta)}[\,\cdot\,] \\
&= \mathbb{E}_{\theta\sim p(\theta)}\big[\mathbb{E}_{x\sim p(x|\theta)}[g(x)]\big] && \text{def. of } \mathbb{E}_{\theta\sim p(\theta)}[\,\cdot\,] \qquad \blacksquare
\end{aligned}
$$

For $\mathbb{E}_{x\sim p(x|z)}[g(x)]$, we have

$$
\begin{aligned}
\mathbb{E}_{x\sim p(x|z)}[g(x)] &= \int_x g(x)\cdot p(x \mid z)\,\mathrm{d}x && \text{def. of } \mathbb{E}_{x\sim p(x|z)}[\,\cdot\,] \\
&= \int_x g(x)\left(\int_\theta p(\theta \mid z)\cdot p(x \mid z,\theta)\,\mathrm{d}\theta\right)\mathrm{d}x && \text{by } p(x \mid z) = \mathbb{E}_{\theta\sim p(\theta|z)}\big[p(x \mid z,\theta)\big] \\
&= \int_\theta p(\theta \mid z)\int_x g(x)\cdot p(x \mid z,\theta)\,\mathrm{d}x\,\mathrm{d}\theta && \text{switch the order} \\
&= \int_\theta p(\theta \mid z)\,\mathbb{E}_{x\sim p(x|z,\theta)}[g(x)]\,\mathrm{d}\theta && \text{def. of } \mathbb{E}_{x\sim p(x|z,\theta)}[\,\cdot\,] \\
&= \mathbb{E}_{\theta\sim p(\theta|z)}\big[\mathbb{E}_{x\sim p(x|z,\theta)}[g(x)]\big] && \text{def. of } \mathbb{E}_{\theta\sim p(\theta|z)}[\,\cdot\,] \qquad \blacksquare
\end{aligned}
$$

## Trivial Bounds of Expectation

Let $X : \Omega \to \mathbb{R}^n$ be a random vector with PDF $p_X$ and $g : \mathbb{R}^n \to \mathbb{R}$ be bounded. Then

$$\mathbb{E}[g(X)] \leq \max_x g(x) \tag{36}$$

where the inequality holds iff

$$p_X(x) = \delta(x - x^*) \quad \text{where } x^* \in \arg\max_x g(x)$$

## Equivalent Formulation of Bellman equation

In some literatures, the system model is given by $p(r, s' \mid s, a)$ instead of $p(r, s' \mid s, a)$ and $p(r, s' \mid s, a)$. In this setting, Bellman equation becomes

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_{a\sim\pi(a|s)}\Big[\mathbb{E}_{r\sim p(r|s,a)}[r] + \gamma \mathbb{E}_{s'\sim p(s'|s,a)}[v_\pi(s')]\Big] && \text{(BE-V1)} \\ &= \mathbb{E}_{a\sim\pi(a|s)}\Big[\mathbb{E}_{r,s'\sim p(r,s'|s,a)}[r + \gamma v_\pi(s')]\Big] && \text{(BE-V1*)} \\ v_\pi(s) &= \mathbb{E}_{r\sim p(r|s)}[r] + \gamma \mathbb{E}_{s'\sim p(s'|s)}[v_\pi(s')] && \text{(BE-V2)} \\ &= \mathbb{E}_{r,s'\sim p(r,s'|s)}[r + \gamma v_\pi(s')] && \text{(BE-V2*)} \end{aligned}$$

Remark:

- The probability $p(r, s' \mid s)$ in $(\text{BE-V2*})$ implicitly depends on $\pi$ since it is obtained by marginalizing $p(r, s' \mid s, a)$ over $a \sim \pi(a \mid s)$.

$$p(r, s' \mid s) = \mathbb{E}_{a\sim\pi(a|s)}\big[p(r, s' \mid s, a)\big]$$

*Proof*: This is direct result of the linearity of expectation.

$$\begin{aligned} \mathbb{E}_{x\sim p(x)}[g(x)] + \mathbb{E}_{y\sim p(y)}[h(y)] &= \mathbb{E}_{x,y\sim p(x,y)}[g(x) + h(y)] \\ \mathbb{E}_{x\sim p(x|z)}[g(x)] + \mathbb{E}_{y\sim p(y|z)}[h(y)] &= \mathbb{E}_{x,y\sim p(x,y|z)}[g(x) + h(y)] \end{aligned}$$

or equivalently in simplified notation

$$\begin{aligned} \mathbb{E}[g(X)] + \mathbb{E}[h(Y)] &= \mathbb{E}[g(X) + h(Y)] \\ \mathbb{E}[g(X) \mid Z = z] + \mathbb{E}[h(Y) \mid Z = z] &= \mathbb{E}[g(X) + h(Y) \mid Z = z] \end{aligned}$$

■