Simple Mathematics Handbook for RL
Useop Gim 2022

# 1 Basic

**Theorem 1.** *Markov Property*
*It is the probability of event which is related with the past events.*

$$P(x_t|x_0, x_1, ..x_{t-1}) = P(x_t)$$

*The Markov model is based on the above property.*

**Theorem 2.** *Markov Process(Markove Chain)*
*In the Markov process is a process which following the Markov Property with tuple* $(S, P)$

**Theorem 3.** *State transition Probability* $P$
*The probability making transform state is called "state transition probability."*

**Theorem 4.** *Reward* $R$
*The reward in Markov Process is obtained from transition.*

**Theorem 5.** *Markov Decision Process*
*The Markov Decision Process represent the Markov Process as tuple.*

**Theorem 6.** *Markov Reward Process*
*It adds a reward and a discount factor element in process* $(S, P, R, \gamma)$

**Theorem 7.** *Markov Decision Process*
*It adds an action element in process* $(S, A, P, R, \gamma)$

**Theorem 8.** *Policy* $\pi$
*From the Markov Process, policy is the distribution of all action by current state.*

$$\pi(a|s) = P\left[A_t = a|S_t\right]$$

**Theorem 9.** *State-value*
*Let* $G$ *is total future reward*
*The state-value is the value of current state*

$$v(s) = \mathbb{E}\left[G_t|S_t = s\right]$$

**Theorem 10.** *Action-value*
*The action-value is the value of action for current state*

$$q(s, a) = \mathbb{E}\left[G_t|S_t = s, A_t = a\right]$$

**Theorem 11.** *Bellman Expected equation*

$$v_\pi(s) = \mathbb{E}\left[R_{t+1} + \gamma v(s_{t+1})|s_t\right]$$
$$q_\pi(s, a) = \mathbb{E}\left[R_{t+1} + \gamma q(s_{t+1}, a_{t+1})|s_t, a_t\right]$$

**Theorem 12.** *Bellman Optimality equation*

$$v_*(s) = max\mathbb{E}\left[R_{t+1} + \gamma v(s_{t+1})|s_t\right]$$

$$q_*(s,a) = max\mathbb{E}\left[R_{t+1} + \gamma q(s_{t+1}, a_{t+1})|s_t, a_t\right]$$

**Theorem 13.** *Transition by policy*
*The probability of transition from state $s_t$ to $s_{t+1}$ with policy*

$$P^\pi(s_{t+1}|s_t) = \sum_{a\in A_t} \pi(a|s_t)p(s_{t+1}|s_t,a)r(s,a)$$

**Theorem 14.** *Reward by policy*
*The probability of transition from state $s_t$ to $s_{t+1}$ with policy*

$$R^\pi(s_t,a) = \sum_{a\in A_t} \pi(a|s_t)r(s_t,a)$$

# 2 Policy Gradient Theorem

**Theorem 15.** *State-value by policy*
*First convert State-value function as policy form*

$$\begin{aligned}
v_\pi(s_t) &= \mathbb{E}_\pi\left[R_{t+1} + \gamma v(s_{t+1})\right] \\
&= R^\pi_{s_t} + \gamma \sum_{s_t} v(s_{t+1}) \\
&= \sum_{a\in A} \pi(a|s_t)r(s_t,a) + \gamma \sum_{s_{t+1}\in S}\sum_{a\in A_t} \pi(a|s_t)p(s_{t+1}|s_t,a)v_\pi(s_{t+1})
\end{aligned}$$

*Through the summation property*

$$v_\pi(s_t) = \sum_{a\in A_t} \pi(a|s_t)r(s_t,a) + \gamma \sum_{a\in A_t} \pi(a|s_t) \sum p(s_{t+1}|s_t,a)v_\pi(s_{t+1})$$

**Theorem 16.** *Action-value by policy*

$$\begin{aligned}
q_\pi(s_t,a) &= \mathbb{E}\left[R_{t+1} + \gamma q(s_{t+1}, a_{t+1})|s_t, a_t\right] \\
&= r(s,a) + \gamma \sum_{s_{t+1}\in S} p(s_{t+1}|s,a) \sum_{a'\in A} \pi(a'|s_{t+1})q_\pi(s_{t+1},a')
\end{aligned}$$

**Theorem 17.** *State-value and Action-value relationship*

$$v_\pi(s_t) = \sum_{a_t\in A_t} \pi(a_t|s_t)q_\pi(s_t,a_t)$$

$$q_\pi(s_t,a) = r(s_t,a) + v_\pi(s_t)$$

**Theorem 18.** *Policy Gradient Theorem*
*For the stochastic policy* $\pi$

$$
\begin{aligned}
\nabla v_{\pi} &= \nabla\left(\sum_a \pi q\right)\\
&= \sum_a \left(\nabla \pi(a|s) q_\pi + \sum \pi(a|s)\nabla q_\pi\right)\\
&= \sum_a \left(\nabla \pi(a|s) q_\pi + \pi(a|s)\sum_{s'} \nabla\left(\sum_{s'} P^a_{s's}\cdot q_\pi(s',a')\right)\right)\\
&= \sum_a \left(\nabla \pi(a|s) q_\pi + \pi(a|s)\sum_{s'} \nabla\left(\sum_{s'} P^a_{s's}\cdot (r+v_\pi(s'))\right)\right)\\
&= \sum_a \left(\nabla \pi(a|s) q_\pi + \pi(a|s)\sum_{s'} P^a_{s's}\cdot \nabla v_\pi(s')\right)
\end{aligned}
$$

*Therefore*

$$
v_\pi(s) = \sum_a \left(\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s)\sum_{s'} P^a_{s's}\cdot \nabla v_\pi(s')\right)
$$

*Then the* $v_\pi(s,a)$ *is repeated Thus let* $\phi(s) = \sum_a \nabla\pi(a|s)q_\pi(s,a)$

$$
\begin{aligned}
\nabla v_\pi(s) &= \phi(s) + \sum_a \left(\pi(a|s)\sum_{s'} P^a_{s's}\cdot \nabla v_\pi(s')\right)\\
&= \phi(s) + \sum_a \sum_{s'} \pi(a|s) P^a_{s's}\cdot \nabla v_\pi(s')\\
&= \phi(s) + \sum_{s'} \sum_a \pi(a|s) P^a_{s's}\cdot \nabla v_\pi(s')\\
&= \phi(s) + \sum_{s'} \sum_k p_\pi(s\to s',k)\cdot \nabla v_\pi(s')\\
&= \phi(s) + \sum_{s'} \sum_k p_\pi(s\to s',k)\cdot (\phi(s')+...)\\
&= \phi(s) + \sum_x \sum_{k=0} p_\pi(s\to x,k)\cdot \phi(x)
\end{aligned}
$$

*Andso let* $\sum_{k=0} p_\pi(s \to x, k)$ *as* $\eta(s)$

$$\begin{aligned}
\nabla v_\pi &= \sum_s \eta(s)\phi(s) \\
&= \left(\sum_s \eta(s)\right) \sum \frac{\eta(s)}{\sum \eta(s)} \phi(s) \\
&= \sum_s d_\pi(s)\phi(s) && \textit{Since } \textstyle\sum \eta \textit{ is constant} \quad d \textit{ is the stationary distribution} \\
&= \sum_s d_\pi(s) \sum_a \nabla\pi(a|s) q_\pi(s,a)
\end{aligned}$$

*Hence the derivative of the expected value function is obateind from the gradient of policy without taking derivative for the reward function*

## 3 Base

**Definition 1.** *DP, MC, TD*
*DP is dynamic programming, it uses the model base*
*MC is Monte Carlo, it uses bunch of samples then estimates the probability (ex calculate circle of pi)*
*TD is Temporal Difference, it uses the difference of the transitional value from one step behind state*

**Definition 2.** *Exploitation vs Exploration*
*Exploitation is deciding the best action through the using the given samples*
*Exploration is collecting samples*

**Definition 3.** *TP, TN, FP, FN, SE, SP, FPR, ROC, AUC*
*TP is true positive*
*TN is true negative*
*FP is false postive (type 1 error)*
*FN is false negative (type 2 error)*
*SE is Sensitivity* $\frac{TP}{TP+FN}$ *which is the rate of correct positive*
*SP is Specificity* $\frac{TN}{TN+FP}$ *which is the rate of correct negative*
*FPR is False Positive Rate which is the rate of wrong positive*
*ROC is the the curve by vertical SE horizontal FPR or (1-SP)*
*AUC is the area of under ROC*