# MDP Cheatsheet Reference

Author: John Schulman
(★) = facts that are a bit more technical

## 1 Markov Decision Process

Infinite-horizon, discounted setting:

- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $P(s,a,s')$: transition kernel
- $R(s,a,s')$: reward function
- $\gamma \in [0,1]$: discount
- $\mu$: initial state distribution (optional)

## 2 Backup Operators

At the core of policy and value iteration are the "Bellman backup operators" $T, T^\pi$, which are mappings $\mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ that update the value function.

$$TV(s) := \max_a \sum_{s'} P(s,a,s')[R(s,a,s') + \gamma V(s')]$$

$$T^\pi V(s) := \sum_{s'} P(s,\pi(s),s')[R(s,\pi(s),s') + \gamma V(s')]$$

Note that $TV(s)$ means that we are evaluating $TV$ (a vector, in the finite case) at state $s$, i.e., it would more properly be written $(TV)(s)$. The same convention is used when considering $T^n V(s)$ and so forth.

### Properties of $T$

- Unique fixed point is $V^*$, defined by $V^*(s) = \mathbb{E}[R_0 + \gamma R_1 + \ldots | s_0 = s]$, where actions are chosen according to an optimal policy: $a_t = \pi^*(s_t)$.
- $n$th iterate can be interpreted as the optimal expected return in $n$-step finite-horizon problem: $T^n V(s) = \max_{\pi_0, \pi_1, \ldots, \pi_{n-1}} \mathbb{E}\ [R_0 + \gamma R_1 + \cdots + \gamma^{n-1} R_{n-1} + \gamma^n V(s_n) | s_0 = s]$, where $a_t = \pi(s_t)\ \forall t$ and we are using the shorthand $R_t := R(s_t, a_t, s_{t+1})$, and the expectation is taken with respect to all states $s_t$ for $t > 0$.
- (★) $T$ is a contraction under the max norm $|\cdot|_\infty$
- $T$ is monotonic, so $V \le TV \Rightarrow V \le TV \le T^2 V \le \cdots \le V^*$, and $V \ge TV \Rightarrow V \ge TV \ge T^2 V \ge \cdots \ge V^*$

### Properties of $T^\pi$

- Unique fixed point is $V^\pi$, defined by $V^\pi(s) = \mathbb{E}[R_0 + \gamma R_1 + \ldots | s_0 = s]$, where actions are chosen according to the policy $a_t = \pi(s_t)$.
- $n$th iterate can be interpreted as the expected return of a $n$-step rollout under $\pi$, with terminal cost $V$: $(T^\pi)^n V(s) = \mathbb{E}[R_0 + \gamma R_1 + \cdots + \gamma^{n-1} R_n + \gamma^n V(s_n) | s_0 = s]$ where $a_t = \pi(s_t)\ \forall t$.
- (★) $T^\pi$ is a contraction under the weighted $\ell_2$ norm $\|\cdot\|_\rho$ where $\rho$ is the steady-state distribution of the Markov chain induced by executing policy $\pi$. $T^\pi$ is also a contraction under the max norm $|\cdot|_\infty$.
- $T^\pi$ is monotonic

## 3 Algorithms

**Algorithm 1** Value Iteration

```
Initialize V^(0).
for n = 1,2,... do
    for s ∈ S do
        V^(n)(s) = max_a Σ_{s'} P(s,a,s')(R(s,a,s') + γ V^(n-1)(s'))
    end for
    ▷ The above loop over s could be written as V^(n) = T V^(n-1)
end for
```

### Properties of value iteration

- If initialized with $V^{(0)} = 0$ and $R(s,a,s') \ge 0$, values monotonically increase, i.e., $V^{(0)}(s) \le V^{(1)}(s) \le \ldots \forall s$.
- Error $V^{(n)} - V^*$ and maximum suboptimality of resulting policy are bounded by $\gamma^n |R|_\infty / (1-\gamma)$.

The policy update step could be written in "operator form" as $\pi^{(n)} = GV^{\pi^{(n-1)}}$ where $GV$ denotes the greedy policy for value function $V$, i.e., $GV(s) = \text{argmax}_a \sum_{s'} P(s,a,s')[R(s,a,s') + \gamma V(s')]$, $\forall s \in S$.

### Properties of policy iteration

- Computes optimal policy and value function in a finite number of iterations

**Algorithm 2** Policy Iteration

Initialize $\pi^{(0)}$.
**for** $n=1,2,...$ **do**
  $V^{(n-1)}=\text{Solve}[V=T^{\pi^{(n-1)}}V]$
  **for** $s\in S$ **do**
    $\pi^{(n)}(s)=\text{argmax}_a\sum_{s'}P(s,a,s')[R(s,a,s')+\gamma V^{(n-1)}(s')]$
    $=\text{argmax}_a Q^{\pi^{(n-1)}}(s,a)$
  **end for**
**end for**

- (★) Performance of policy monotonically increases. In fact, at the $n$th iteration, the policy improves by $(1-\gamma P^{\pi^{(n)}})^{-1}(TV^{(n-1)}-V^{(n-1)})$, where $P^\pi$ is the matrix defined by $P^\pi(s,s')=P(s,\pi(s),s')$,

**Algorithm 3** Modified Policy Iteration

Initialize $V^{(0)}$.
**for** $n=1,2,...$ **do**
  $\pi(s)=GV^{(n-1)}$
  $V^{(n)}=(T^\pi)^kV^{(n-1)}$, for integer $k\geq 1$
**end for**

**Properties of modified policy iteration**

- Computes optimal policy in a finite number of iterations, and value function converges to optimal one: $V^{(n)}\to V^*$.
- $k=1$ gives value iteration, $k=\infty$ limit gives policy iteration (except at the first iteration.)

## 4 Value Functions and Bellman Equations

The term "value function" in general refers to a function that returns the expected sum of future rewards. However, there are several different types of value function. A "state-value function" function $V(s)$ is a function of state, whereas a "state-action-value function" $Q(s,a)$ is a function of a state-action pair.

Below, we list the most common value functions with a pair of equations: the first one involving an infinite sum of rewards, the second one providing a self-consistency equation (a "Bellman equation") with a unique solution. All of the expectations are taken with respect to all states $s_t$ for $t>0$

$$V^\pi(s)=\mathbb{E}[R_0+\gamma R_1+...|s_0=s], \text{ where } a_t=\pi(s_t)\ \forall t$$
$$V^\pi(s)=\sum_{s'}P(s,\pi(s),s')[R(s,\pi(s),s')+\gamma V^\pi(s')]$$
$$Q^\pi(s,a)=\mathbb{E}[R_0+\gamma R_1+...|s_0=s,a_0=a], \text{ where } a_t=\pi(s_t)\ \forall t$$
$$Q^\pi(s,a)=\sum_{s'}P(s,a,s')[R(s,a,s')+\gamma Q^\pi(s',\pi(s'))]$$
$$V^*(s)=\mathbb{E}[R_0+\gamma R_1+...|s_0=s] \text{ where } a_t=\pi^*(s_t)\ \forall t$$
$$V^*(s)=\max_a\sum_{s'}P(s,a,s')[R(s,a,s')+\gamma V^*(s')]$$
$$Q^*(s,a)=\mathbb{E}[R_0+\gamma R_1+...|s_0=s,a_0=a], \text{ where } a_t=\pi(s_t)\ \forall t$$
$$Q^*(s,a)=\sum_{s'}P(s,a,s')[R(s,a,s')+\gamma\max_{a'}Q^*(s',a')]$$

## 5 Some Definitions

**Contraction**: a function $f$ is a contraction under norm $|\cdot|$ with modulus $\gamma$ iff $|f(x)-f(y)|\leq\gamma|x-y|$. By the Banach fixed point theorem, a contraction mapping on $\mathbb{R}^d$ has a unique fixed point.

**Stationary Distribution**: Given a transition matrix $P_{ss'}$, the stationary distribution $\rho$ is the left eigenvector, satisfying $\rho_{s'}=\rho_s P_{ss'}$. If the transition matrix satisfies appropriate conditions (see the Markov chain theory [3]), then $\rho=\lim_{n\to\infty}\nu P^k$ for any initial distribution $\nu$. In the context of MDPs, we speak speak of the *transition matrix induced by policy* $\pi$, defined by $P_{ss'}=P(s,\pi(s),s')$, and similarly, there is a stationary distribution induced by the policy $\rho_\pi$.

**Monotonic**: a function $f$ is monotonic if $x\leq y\implies f(x)\leq f(y)$. This definition can be extended to the case that $f:\mathbb{R}^d\to\mathbb{R}^d$, in which case the inequalities hold for each component on the LHS and RHS.

## References


[1] D. P. Bertsekas, D. P. Bertsekas, et al. *Dynamic programming and optimal control*, vol. 1. Athena Scientific Belmont, MA, 1995.

[2] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2005.

[3] Wikipedia. Markov chain — Wikipedia, the free encyclopedia, 2015.