# Chapter 02 – Value Iteration

*Author: Gianmarco Scarano*

*gianmarcoscarano@gmail.com*

# 1. Policy Evaluation

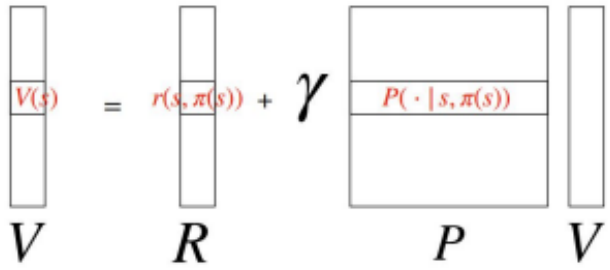
Given an MDP $(S, A, T, R, \gamma)$ and a policy $\pi$, we want to evaluate $\pi$ (for example $V^{\pi}$). We do this due to the fact that there could be $A^S$ possible policies and we want to find the optimal one.

Introducing this concept, we know that for ALL the states, the Bellman equation holds:

$$V^{\pi}(s) = r + \gamma \mathbb{E}_{s' - p(\,.\,|\,s,\ \pi(s))}[V^{\pi}(s')]$$

This let us deal with S possible equations which translate in S possible constraints.

If we combine all the constraints together, we of course could have some sort of Matrix visualization as follows:




This can be read as $V = R + \gamma * p * V$ and we can solve it by solving it for V, as:

$$V = (I - \gamma * P)^{-1} * R$$

Very computationally expensive = $O(S^3)$

## 1.1 Fixed-Point Iteration & Contractions

With this being said, we introduce a fixed point, which is nothing less than $x = f(x)$.

In order to find such points, we initialize $x_0$ and we loop $x_{i+1} = f(x_i)$.

We stop convergence where $x$ is found and does not change anymore. Namely, we stop if the difference between two points is smaller than a certain threshold.

If I keep applying the same function, I will always get the same value.

The fact that we have converge after a certain threshold is met, is given thanks to the existence of contraction mappings.

### 1.1.1 Contraction mappings

$f: M \to M$ (M is a metric space) is a contraction mapping if:

$$\textbf{(1)}\ |\, f(x) - f(x') \,| \leq k\, |\, x - x' \,| \qquad for\ k\ in\ [0, 1)$$

If we consider:

- $x = 0$
- A function $f$ as the square function: $x^2\ \epsilon$ [0, 1)
- k = 0.99
- then: f(x) = $x^2$.

Example:

$$f(x0) = f(0.99) = 0.98$$

$$f(x1) = f(0.98) = 0.96$$

$$f(x2) = f(0.96) = 0.92$$

There would be a point where $x^2 = 0.01^2$.

The value will get smaller and smaller to the point it will be almost 0, such that $x^2 = 0 = 0^2 = 0$.

In fact, if we apply the contraction mapping:

$$
\begin{aligned}
&|f(0.96) - f(0.92)| = |0.92 - 0.85| = |0.07| \\
\Rightarrow\ &|0.07| \leq k * |0.96 - 0.92| \\
\Rightarrow\ &|0.07| \leq 0.99 * |0.04| \\
\Rightarrow\ &|0.07| \leq 0.0396
\end{aligned}
$$

This doesn't hold for now, but in the end, using the square function we'll find a point where the contraction mapping holds.

In the simplest case of a contraction mapping, we could face a simple operator such as a matrix (e.g. O) and we could also replace $k$ (used in the previous formula) with $\gamma$ as they have the same range:

$$|OV - OV'| \leq \gamma |V - V'|$$

## 2. Iterative Policy Evaluation

Here, we do initialize $V_0$ in $[0, \frac{1}{(1-\gamma)}]$ (typically 0). Remember that $\frac{1}{(1-\gamma)}$ is due to the discount factor being a geometric series.

We loop using this formula $V_{i+1} = R + \gamma P V_i$ until we reach convergence.

Since we are looping throughout all the states, applying $V$ (like for the fixed point, we use $V = F(V)$), we use the Matrix form to englobe all the states. Note that in this case, our function $F(\cdot)$ is equal to using the Bellman Equation (a bit less expensive).

It's nice but still for each iteration we have a cost of $O(S^2)$.

The theorem for the Iterative Policy Evaluation is the following:

At the end we have, for all s in S

$$\|V^t(s) - V^\pi(s)\| \leq \gamma^t \|V^0 - V^\pi\|$$

$$
\begin{aligned}
\forall s,\ & \left| V^{t+1}(s) - V^\pi(s) \right| \\
&= \left| r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \left( r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right) \right| \\
&= \gamma \left| \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right| \\
&\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} \left| V^t(s') - V^\pi(s') \right| \\
&\leq \gamma \left\| V^t - V^\pi \right\|_\infty
\end{aligned}
$$

Commenting this out in general, we say that the distance of the value we computed at iteration $t$ is close to the actual truth value of the policy $\pi$ as much as the initial distance of the first one (scaled by $\gamma$ elevated to the number of iteration). We simply expand each $V$, cancel out things and bring out the Expectation $\mathbb{E}$.

What we are trying to prove is a way to compute the precision (or quality) of my policy if I know the Ground Truth (Which is the actual value of the policy, in this case $V^\pi$ and showing that the solution at $t+1$ (first row) is better than the solution at time $t$ (last row).

Our $V^{t+1}(s) = F(x)$ and our $V^\pi(s) = F(x)$. In fact, this is equal to **(1)**. Cost = $O(S^2 * \ln\left(\frac{1}{\epsilon}\right))$

# 3. How to find the Optimal Policy?

## 3.1 Bellman backup is a contraction point

But now, we are really interested in finding what is my optimal policy $\pi^*$. For this reason, we'll use the Bellman optimality and the Bellman operator.

By defining the infinity norm $||V|| = max_s|V(S)|$ and by setting $\gamma < 1$, we can easily define the (non-linear) BV operator as a Bellman equation applied to V:

$$BV\ or\ TV = max_a(r(s,a) + \gamma \mathbb{E}_{s' - p(.\,|s,\ a)}[V(s')])$$

$$\begin{aligned}
\|BV_k - BV_j\| &= \left\| \max_a \left( R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V_k(s') \right) - \max_{a'} \left( R(s,a') + \gamma \sum_{s' \in S} P(s'|s,a') V_j(s') \right) \right\| \\
&\leq \max_a \left\| \left( R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V_k(s') - R(s,a) - \gamma \sum_{s' \in S} P(s'|s,a) V_j(s') \right) \right\| \\
&= \max_a \left\| \gamma \sum_{s' \in S} P(s'|s,a)(V_k(s') - V_j(s')) \right\| \\
&\leq \max_a \left\| \gamma \sum_{s' \in S} P(s'|s,a) \|V_k - V_j\| \right\| \\
&= \max_a \left\| \gamma \|V_k - V_j\| \sum_{s' \in S} P(s'|s,a)) \right\| \\
&= \gamma \|V_k - V_j\|
\end{aligned}$$

So, basically here we want to show that the new value that we get by applying the Bellman Optimality Equation minus (-) the same thing applied at iteration J is $\leq$ than the value that we would get without the Bellman Optimality Equation.

In the 3rd last row, $V_k - V_j$ does not depend on the states, so we take them out of the sum. That sum then is equal to 1 because we are summing the whole probabilities of ALL the states (which is indeed equal to 1), that's why then we are left with the final row formula.

## 3.2 Bellman Operator for Q

$$TQ(s,a) = r(s,a) + \gamma \mathbb{E}_{s' - p(.\,|s,\ a)} max_{a'}[Q(s',a')]$$

Since $Q: S\ x\ A \rightarrow \mathbb{R}$, then also $TQ: S\ x\ A \rightarrow \mathbb{R}$

## 3.3 Value Iteration

All what we said, of course holds also for $V^*$ and we can obtain $Q^* = TQ^*$, since $Q^*$ is a fixed-point solution to $Q = TQ$.

- We initialize $||Q_0||$ in $[0, \frac{1}{(1-y)}]$ (typically 0)
- Until convergence, for all states and for all the actions: $Q_{i+1} = TQ_i$ which is substantially the Bellman Operation ($T$) applied on $Q_i$.
- $||Q_{i+1} - Q^*|| = ||TQ_i - TQ^*|| \leq \gamma||Q_i - Q^*|| \leq \gamma^{i+1}||Q_0 - Q^*||$
- The more we iterate, the more we get towards the optimal policy. This is needed when the Ground Truth Policy is not given. If the Ground Truth Policy is given, we can simply use the "Iterative Policy Evaluation".

Finally, we know that $\pi^*(s) = argmax_a Q^*(s,a)$ and since $Q_i(s,a) \cong Q^*(s,a)$, we could choose:

$$\pi_i(s) = argmax_a Q_i(s,a)$$

Without explaining and dig into the theorem, we simply state that the quality of such policy for all the states is the following:

$$V^{\pi i}(s) \geq V^*(s) - \frac{2 \cdot \gamma^t}{(1-y)} \cdot ||Q_0 - Q^*||$$

$\pi_i(s)$, though, is not the *OPTIMAL* policy since it's an approximation, but we know that it is greater than the (OPTIMAL policy – 2*$\gamma$ etc.) we calculated just above.