CE 273
Markov Decision Processes

Lecture 8

Value Iteration

Previously on Markov Decision Processes

The objective in the discounted cost MDP problem is

$$\lim_{N\to\infty} \mathbb{E}_w \sum_{k=0}^{N-1} \left\{ \alpha^k g(x_k, u_k, w_k) \right\}$$

Under most practical situations that we encounter, this limit exists and we can also exchange the limit and expectation and write

$$\mathbb{E}_w \sum_{k=0}^{\infty} \left\{ \alpha^k g(x_k, u_k, w_k) \right\}$$

Likewise, given a particular policy $\pi = \{\mu_0, \mu_1, \ldots\}$, the value function can be written as

$$J_\pi(x_0) = \lim_{N\to\infty} \mathbb{E}_w \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}$$

We will make appropriate assumptions (such as bounded costs) that will guarantee the existence of the above limit.

Previously on Markov Decision Processes

As before, if Π denotes the set of admissible policies, the optimal cost function is given by

$$J^*(x_0) = J_{\pi^*}(x_0) = \min_{\pi \in \Pi} J_\pi(x_0)$$

Note that when writing the value functions, we can drop k and think of J as a function of x alone because no matter where we are, we have an infinite number of stages over which our objective is computed.

For most problems, it turns out that the optimal policy is also stationary! That is, $\pi = \{\mu, \mu, \ldots\}$. So we can simply write $J_\mu(x)$ as the cost of the policy instead of $J_\pi(x)$.

Thus, unlike the finite horizon case, we need not find an infinite number of functions $J_k^*(x_k)$ and $\mu_k^*(x_k)$ but just compute $J^*(x)$ and $\mu^*(x)$.

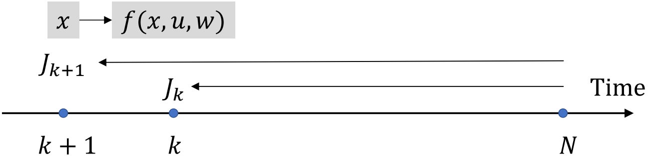
Previously on Markov Decision Processes

The new DP algorithm is

$$J_0(x) = 0 \,\forall\, x \in X$$

$$J_{k+1}(x) = \min_{u \in U(x)} \mathbb{E}\Big\{ g(x,u,w) + \alpha J_k(f(x,u,w)) \Big\}$$

Time is now measured backward from some N which tends to ∞.



Thus, after N iterations, we would have found the optimal cost for the N-stage discounted problem with terminal cost function $\alpha^N J$.

If we stop the algorithm after k iterations, we would have found the optimal cost for the k-stage discounted problem with terminal cost function $\alpha^k J$.

Previously on Markov Decision Processes

Definition

Given a function $J : X \to \mathbb{R}$, define $(TJ)(x)$ as

$$(TJ)(x) = \min_{u \in U(x)} \mathbb{E}\Big\{ g(x,u,w) + \alpha J(f(x,u,w)) \Big\}$$

Definition

Given a function $J : X \to \mathbb{R}$, define $(T_\mu J)(x)$ as

$$(T_\mu J)(x) = \mathbb{E}\Big\{ g(x,\mu(x),w) + \alpha J(f(x,\mu(x),w)) \Big\}$$

We can also define composition mappings

$$(T^0 J)(x) = J(x) \,\forall\, x \in X$$
$$(T^k J)(x) = (T(T^{k-1}J))(x) \,\forall\, x \in X$$

$(T^k J)(x)$ is equivalent to k iterations of the new DP algorithm and is hence the optimal cost of the k-stage discounted problem with terminal costs $\alpha^k J$.

Likewise, $(T_\mu^0 J)(x) = J(x)$ and $(T_\mu^k J)(x) = (T_\mu(T_\mu^{k-1}J))(x) \,\forall\, x \in X$

Previously on Markov Decision Processes

We will mostly deal with countable state, control, and disturbance spaces. In such cases, we can write the DP equations and the T operators in more compact form.

Suppose the state space is $X = \{1, \dots, n\}$. The transitions no longer are a function of k and hence we can write

$$p_{ij}(u) = \mathbb{P}[x_{k+1} = j | x_k = i, u_k = u] \, \forall \, i, j \in X, u \in U(i)$$

The two T mappings take the form

$$(TJ)(i) = \min_{u \in U(i)} \left\{ g(i, u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j) \right\} \forall \, i \in X$$

$$(T_\mu J)(i) = \left\{ g(i, \mu(i)) + \alpha \sum_{j=1}^{n} p_{ij}(\mu(i)) J(j) \right\} \forall \, i \in X$$

Note that it has been implicitly assumed that g does not depend on the disturbance. How can we relax that?

Previously on Markov Decision Processes

One can also write vector forms of these equations.

$$J = \begin{pmatrix} J(1) \\ \vdots \\ J(n) \end{pmatrix} \qquad TJ = \begin{pmatrix} (TJ)(1) \\ \vdots \\ (TJ)(n) \end{pmatrix} \qquad T_\mu J = \begin{pmatrix} (T_\mu J)(1) \\ \vdots \\ (T_\mu J)(n) \end{pmatrix}$$

For a given policy μ, we can also write the one-step transition probability matrix as

$$P_\mu = \begin{pmatrix} p_{11}(\mu(1)) & \dots & p_{1n}(\mu(1)) \\ \vdots & \ddots & \vdots \\ p_{n1}(\mu(n)) & \dots & p_{nn}(\mu(n)) \end{pmatrix}$$

and the cost vector for a fixed policy μ as

$$g_\mu = \begin{pmatrix} g(1, \mu(1)) \\ \vdots \\ g(n, \mu(n)) \end{pmatrix}$$

Thus, the T-mu operator in matrix form can be written as

$$T_\mu J = g_\mu + \alpha P_\mu J$$

Lecture Outline

1. Analysis Review
2. Value Iteration
3. Variants of Value Iteration

Analysis Review

Let X be a vector space. Define a norm, a real valued-function $\|\cdot\|$, which satisfies the following conditions for all $x \in X$,

1. $\|x\| \geq 0$, and $\|x\| = 0 \Leftrightarrow x = 0$
2. $\|ax\| = |a|\|x\|$ for any scalar a
3. $\|x + y\| \leq \|x\| + \|y\|$

Definition (Cauchy Sequence)

Let X be a normed vector space. A sequence $\{x_k\}$ is said to be a Cauchy sequence if for any $\epsilon > 0 \, \exists \, N$ such that $\|x_m - x_n\| \leq \epsilon \, \forall \, m, n \geq N$.

In other words, for a Cauchy sequence, $\|x_m - x_n\| \to 0$ as $m, n \to \infty$.

Definition (Complete Space)

The space X is said to be complete if every Cauchy sequences converges to a point in X.

A complete normed vector space is also called a **Banach space**.

Example of Banach spaces include

- $\mathbb{R}^n$ with the Euclidean norm
- $B(X)$, the set of all bounded functions $J : X \to \mathbb{R}$ with the sup-norm
$$\|J\| = \sup_{x \in X} |J(x)|$$

The sup-norm is also called the ℓ_∞-norm and is also denoted using $\|\cdot\|_\infty$

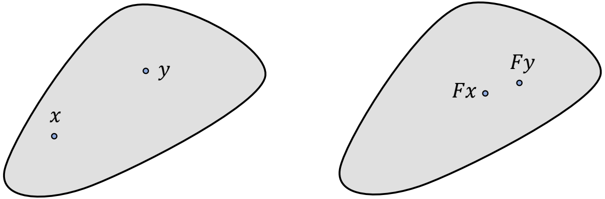
Analysis Review

Contraction Mappings

Definition (Contraction Mapping)

A function $F : X \to X$ is said to be a contraction mapping if for some $\rho \in (0, 1)$,

$$\|Fx - Fy\| \leq \rho \|x - y\| \, \forall \, x, y \in X$$



The scalar ρ is called the modulus of contraction of F.Where are we going with this? We will show that T and T_μ are contraction mappings.

Theorem (Banach Fixed Point Theorem)

Let $B(X)$ be a Banach space and suppose that $F : B(X) \to B(X)$, is a contraction mapping with modulus of contraction ρ. Then there exists a unique $J^* \in B(X)$ such that

1. $\lim_{k\to\infty} F^k J = J^* \,\forall\, J \in B(X)$
2. $J^* = FJ^*$
3. $\|F^k J - J^*\| \leq \rho^k \|J - J^*\| \,\forall\, k$

Proof.

Proof of (1): Pick an arbitrary $J \in B(X)$. Consider the sequence $\{J_k\}$, where $J_0 = J, J_1 = FJ, J_2 = F^2 J, \ldots$, i.e., $J_{k+1} = F^k J$. We will first show that $\{J_k\}$ is a Cauchy sequence.

$$\|J_{k+1} - J_k\| \leq \rho \|J_k - J_{k-1}\| \,\forall\, k = 1, 2, \ldots$$

Re-applying this inequality, we can write

$$\|J_{k+1} - J_k\| \leq \rho^k \|J_1 - J_0\| \,\forall\, k = 1, 2, \ldots$$

Proof.

For every $m \geq 1$, using triangle inequality,

$$\begin{aligned}\|J_{k+m} - J_k\| &\leq \sum_{i=1}^{m} \|J_{k+i} - J_{k+i-1}\| \\ &\leq \sum_{i=1}^{m} \rho^{k+i-1} \|J_1 - J_0\| \\ &\leq \sum_{i=1}^{\infty} \rho^{k+i-1} \|J_1 - J_0\| \\ &= \frac{\rho^k}{1-\rho} \|J_1 - J_0\|\end{aligned}$$

Thus, $\{J_k\}$ is a Cauchy sequence and must converge to some $J^* \in B(X)$ since $B(X)$ is complete. Hence, (1) is proved.

Proof.

Proof of (2): For all $k \geq 1$,

$$0 \leq \|FJ^* - J^*\| \leq \|FJ^* - J_k\| + \|J_k - J^*\| \quad \text{(Triangle Inequality)}$$
$$\leq \rho\|J^* - J_{k-1}\| + \|J_k - J^*\| \text{ (Contraction Mapping)}$$

Taking limit as $k \to 0$ and using the sandwich theorem, $J^* = FJ^*$.

Suppose J^* was not unique. Let another $\hat{J}$ be a fixed point that satisfies $\hat{J} = F\hat{J}$. $\|J^* - \hat{J}\| = \|FJ^* - F\hat{J}\| \leq \rho\|J^* - \hat{J}\|$, which implies $J^* = \hat{J}$.

Proof of (3): Rate of convergence

$$\|F^k J - J^*\| = \|F^k J - FJ^*\| \leq \rho\|F^{k-1} J - J^*\|$$

Proceeding similarly,

$$\|F^k J - J^*\| \leq \rho^k\|J - J^*\|$$

■

Value Iteration

Value Iteration

Wish List

Recall that we hypothesized that the following is true for infinite horizon MDPs:

1. $J^*(i) = \lim_{k\to\infty}(T^k J)(i) \, \forall \, i \in X$ for any bounded function J.
2. $J^* = TJ^*$, i.e., J^* is a fixed point of the mapping T.
3. If $\mu(i)$ attains the minimum in the RHS of the above equation, then it is optimal.

These conditions naturally lead to an algorithm to compute the optimal value functions.

Let us now formally prove these using the Banach fixed point theorem. We'd also like to establish how far we are from the optimal solution and optimal policy after a fixed number of iterations.

Value Iteration
Assumptions

We will make the following assumptions throughout infinite horizon models unless otherwise stated.

- Stationary costs and dynamics
- Bounded costs, i.e., $|g(i,u)| \leq M \, \forall i \in X, u \in U(i)$
- Countable state, control, and disturbance space

For computing the optimal value functions and policies, we further assume that the state, control, and disturbance spaces are finite.

Value Iteration

Useful Lemmas

For any two functions $J : X \to \mathbb{R}$ and $J' : X \to \mathbb{R}$ we write

$$J \leq J' \text{ if } J(i) \leq J'(i) \, \forall \, i \in X$$

Lemma (Monotonicity Lemma)

For any $J : X \to \mathbb{R}$ and $J' : X \to \mathbb{R}$ such that $J \leq J'$ and a stationary policy μ,

1. $T^k J \leq T^k J'$
2. $T_\mu^k J \leq T_\mu^k J'$

Proof (Sketch).

Recall that $T^k J$ is the optimal value function of the k-stage problem with terminal costs $\alpha^k J$. Thus, if the terminal costs are $\alpha^k J'$ instead, using induction, we can show that $T^k J \leq T^k J'$. ■

As a consequence, note that if $J \leq TJ$, then $T^k J \leq T^{k+1} J$, $\forall \, k \geq 1$.

Value Iteration

Useful Lemmas

Suppose $e : X \to \mathbb{R}$ denotes the unit function that takes a value 1 for all i and let r be a scalar.

$$\begin{aligned}
\big(T(J + re)\big)(i) &= \min_{u \in U(i)} \mathbb{E}\left\{ g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u)(J+re)(j) \right\} \\
&= \min_{u \in U(x)} \mathbb{E}\left\{ g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u)J(j) + \alpha r \right\} \\
&= \big(TJ\big)(i) + \alpha r
\end{aligned}$$

Similarly, we can show $\big(T_\mu(J + re)\big)(i) = (T_\mu J)(i) + \alpha r$. These results can be extended using induction as

Lemma (Constant Shift Lemma)

For every k, and $J : X \to \mathbb{R}$ and stationary policy μ

1. $\big(T^k(J + re)\big)(i) = \big(T^k J)\big)(i) + \alpha^k r$
2. $\big(T^k_\mu(J + re)\big)(i) = \big(T^k_\mu J)\big)(i) + \alpha^k r$

Theorem (Banach Fixed Point Theorem)

Let $B(X)$ be a Banach space and suppose that $F : B(X) \to B(X)$, is a contraction mapping with modulus of contraction ρ. Then there exists a unique $J^* \in B(X)$ such that

1. $\lim_{k\to\infty} F^k J = J^* \, \forall \, J \in B(X)$
2. $J^* = FJ^*$
3. $\|F^k J - J^*\| \leq \rho^k \|J - J^*\| \, \forall \, k$

Let us now use the Banach Fixed Point Theorem. Suppose X is the state space and F is replaced with T.

Let $B(X)$ denote the set of all bounded functions $J : X \to \mathbb{R}$ with the sup-norm, which is a Banach space.

What else do we need to apply the above theorem?

Value Iteration

Proposition

$T : B(X) \to B(X)$ is a contraction mapping with $\rho = \alpha$

Proof.

Let $J, J' \in B(X)$ and $r = \|J - J'\| = \sup_{i \in X} |J(i) - J'(i)|$. $r < \infty$ since J, J' are bounded. Hence, we may write

$$J(i) - r \leq J'(i) \leq J(i) + r \, \forall \, i \in X$$

Using Monotonicity Lemma,

$$(T(J - re))(i) \leq (TJ')(i) \leq (T(J + re))(i) \, \forall \, i \in X$$

Using the Constant Shift Lemma,

$$(TJ)(i) - \alpha r \leq (TJ')(i) \leq (TJ)(i) + \alpha r \, \forall \, i \in X$$

which implies

$$|(TJ)(i) - (TJ')(i)| \leq \alpha r \, \forall \, i \in X$$
$$\Rightarrow \|TJ - TJ'\| \leq \alpha \|J - J'\|$$

Thus, T is a contraction mapping. ∎

Hence, from the Banach Fixed Point Theorem, $\lim_{k\to\infty} T^k J = J^*$, where J^* is the fixed point of T. Are we done?

Technically, we've just shown that $J^* \in B(X)$ but haven't formally proved that it is the same J^* which minimizes the objective (we have informally interpreted this J^* as the limit of $T^k J$ using a finite horizon model)

$$J^*(x_0) = \min_{\mu} \lim_{N\to\infty} \mathbb{E}\left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k), w_k) \right\}$$

We will skip this part, but establishing it is not very difficult.

Value Iteration

Summary of Results

Proposition

For any bounded function $J : X \to \mathbb{R}$,

$$J^* = \lim_{k \to \infty} T^k J$$

Proposition (Bellman Equations)

The optimal value functions satisfy

$$J^* = TJ^*$$

and J^* *is the unique solution of this equation.*

Value Iteration

Summary of Results

In a similar fashion, we can show that T_μ is also a contraction mapping and invoke the Banach fixed point theorem to derive the following results.

Proposition

For any bounded function $J : X \to \mathbb{R}$*,*

$$J_\mu = \lim_{k \to \infty} T_\mu^k J$$

Proposition

The value functions associated with a stationary policy μ *satisfy*

$$J_\mu = T_\mu J_\mu$$

and J_μ *is the unique solution of this equation.*

Combining the results from the last two slides, we can also say something about the optimal policies

Proposition

A stationary policy μ is optimal $\Leftrightarrow$ it attains the minimum in the Bellman equations, i.e.,

$$TJ^* = T_\mu J^*$$

The proof of this proposition is trivial.

Value Iteration

Fix a tolerance level $\epsilon > 0$
Select $J_0 \in B(X)$ and $k \leftarrow 0$
$J_1 \leftarrow TJ_0$
while $\|J_{k+1} - J_k\| > \frac{\epsilon(1-\alpha)}{2\alpha}$ **do**
    $k \leftarrow k + 1$
    $J_{k+1} \leftarrow TJ_k$
end while

Select μ_ϵ that satisfies $T_{\mu_\epsilon} J_{k+1} = TJ_{k+1}$

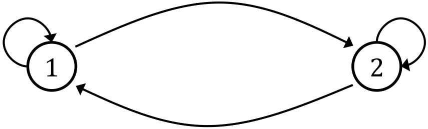
In other words, the policy constructed at termination can be written as

$$\mu_\epsilon(i) \in \arg\min_{u \in U(i)} \mathbb{E}\left\{ g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J_{k+1}(j) \right\}$$

Value Iteration

Example

Perform five iterations of the VI algorithm for the following example with two states 1 and 2. Assume that the discount factor is 0.9.



- $U(1) = \{u_1, u_2\}$
- $g(1, u_1) = 2, g(1, u_2) = 0.5$
- $p_{1j}(u_1) = [3/4 \ \ 1/4]$
- $p_{1j}(u_2) = [1/4 \ \ 3/4]$

- $U(2) = \{u_1, u_2\}$
- $g(2, u_1) = 1, g(2, u_2) = 3$
- $p_{2j}(u_1) = [3/4 \ \ 1/4]$
- $p_{2j}(u_2) = [1/4 \ \ 3/4]$

Value Iteration

Example

Table: Value Iteration Results

	1			2		
k	u_1	u_2	$J_k(1)$	u_1	u_2	$J_k(2)$
0	-	-	0.000	-	-	0.000
1	2.000	**0.500**	0.500	**1.000**	3.000	1.000
2	2.563	**1.288**	1.288	**1.563**	3.788	1.563
3	3.221	**1.844**	1.844	**2.221**	4.344	2.221
4	3.745	**2.414**	2.414	**2.745**	4.914	2.745
5	4.247	**2.896**	2.896	**3.247**	5.396	3.247

Proposition

μ_ϵ is ϵ-optimal, i.e., $\|J_{\mu_\epsilon} - J^*\| \leq \epsilon$

Proof.

Recall that J_{μ_ϵ} is the value function that is a fixed point of T_{μ_ϵ}, i.e., $J_{\mu_\epsilon} = T_{\mu_\epsilon} J_{\mu_\epsilon}$. Also, by construction, $T_{\mu_\epsilon} J_{k+1} = T J_{k+1}$.

$$\|J_{\mu_\epsilon} - J^*\| \leq \|J_{\mu_\epsilon} - J_{k+1}\| + \|J_{k+1} - J^*\|$$

Consider the first term $\|J_{\mu_\epsilon} - J_{k+1}\|$:

$$\begin{aligned} \|J_{\mu_\epsilon} - J_{k+1}\| &\leq \|J_{\mu_\epsilon} - TJ_{k+1}\| + \|TJ_{k+1} - J_{k+1}\| \\ &= \|T_{\mu_\epsilon} J_{\mu_\epsilon} - T_{\mu_\epsilon} J_{k+1}\| + \|TJ_{k+1} - TJ_k\| \\ &\leq \alpha \|J_{\mu_\epsilon} - J_{k+1}\| + \alpha \|J_{k+1} - J_k\| \end{aligned}$$

Thus, $\|J_{\mu_\epsilon} - J_{k+1}\| \leq \frac{\alpha}{1-\alpha} \|J_{k+1} - J_k\|$. In a similar manner, we can show that, the second term, $\|J_{k+1} - J^*\| \leq \frac{\alpha}{1-\alpha} \|J_{k+1} - J_k\|$.

The termination criteria $\Rightarrow \|J_{k+1} - J_k\| \leq \frac{\epsilon(1-\alpha)}{2\alpha}$. Hence, $\|J_{\mu_\epsilon} - J^*\| \leq \epsilon$. ■

Variants of Value Iteration

Variants of Value Iteration

Error Bounds

Proposition (Error Bounds for VI)

For every J, state i, and k,

$$(T^k J)(i) + \underline{c}_k \leq (T^{k+1} J)(i) + \underline{c}_{k+1} \leq J^*(i) \leq (T^{k+1} J)(i) + \bar{c}_{k+1} \leq (T^k J)(i) + \bar{c}_k$$

where

$$\underline{c}_k = \frac{\alpha}{1-\alpha} \min_{i=1,\ldots,n} \left\{ (T^k J)(i) - (T^{k-1} J)(i) \right\}$$

$$\bar{c}_k = \frac{\alpha}{1-\alpha} \max_{i=1,\ldots,n} \left\{ (T^k J)(i) - (T^{k-1} J)(i) \right\}$$

Thus, at any iteration k, one can find an interval for each state within which the optimal value must lie.

The regular VI algorithm can be terminated when the difference between $\bar{c}_k$ and $\underline{c}_k$ becomes small and calculate a final estimate of the value functions using the average of the bounds

$$\hat{J}_k = T^k J + \left(\frac{\bar{c}_k + \underline{c}_k}{2} \right) e$$

Variants of Value Iteration

Gauss-Seidel Algorithm

In the VI algorithm, J_{k+1} for each state is calculated from old J_k values. This is similar to the Jacobi method for solving a system of equations.

The convergence rate can be improved by updating the J values using other J values that were updated in the same iteration. The T operator can be replaced with the F mapping defined below:

$$(FJ)(1) = \min_{u \in U(1)} \left\{ g(1,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j) \right\}$$

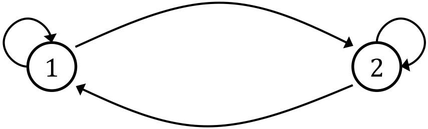
$$(FJ)(i) = \min_{u \in U(i)} \left\{ g(i,u) + \alpha \sum_{j=1}^{i-1} p_{ij}(u)(FJ)(i) + \alpha \sum_{j=i}^{n} p_{ij}(u) J(j) \right\} \forall\, i = 2, \ldots, n$$

This method is also called **Asynchronous Value Iteration**.

Variants of Value Iteration

Example

Perform five iterations of the VI with error bounds and the Gauss-Seidel algorithm for the following example with two states 1 and 2. Assume that the discount factor is 0.9.



- $U(1) = \{u_1, u_2\}$
- $g(1, u_1) = 2, g(1, u_2) = 0.5$
- $p_{1j}(u_1) = [3/4 \;\; 1/4]$
- $p_{1j}(u_2) = [1/4 \;\; 3/4]$
- $U(2) = \{u_1, u_2\}$
- $g(2, u_1) = 1, g(2, u_2) = 3$
- $p_{2j}(u_1) = [3/4 \;\; 1/4]$
- $p_{2j}(u_2) = [1/4 \;\; 3/4]$

Variants of Value Iteration

Example

Table: Value Iteration with Error Bounds

	1			2		
k	$J_k(1)$	$J_k(1) + \underline{c}_k$	$J_k(1) + \bar{c}_k$	$J_k(2)$	$J_k(2) + \underline{c}_k$	$J_k(2) + \bar{c}_k$
0	0	-	-	0	-	-
1	0.500	5.000	9.500	1.000	5.500	10.000
2	1.288	6.350	8.375	1.563	6.625	8.650
3	1.844	6.856	7.767	2.221	7.232	8.144
4	2.414	7.129	7.540	2.745	7.460	7.870
5	2.896	7.287	7.417	3.247	7.583	7.768

Variants of Value Iteration

Example

Table: Gauss-Seidel Value Iteration

	1			2		
k	u_1	u_2	$J_k(1)$	u_1	u_2	$J_k(2)$
0	-	-	0.000	-	-	0.000
1	2.000	**0.500**	0.500	**1.338**	3.113	1.338
2	2.638	**1.515**	1.515	**2.324**	4.244	2.324
3	3.546	**2.409**	2.409	**3.149**	5.111	3.149
4	4.335	**3.168**	3.168	**3.847**	5.839	3.847
5	5.004	**3.809**	3.809	**4.437**	6.454	4.437

Your Moment of Zen