ECE276B: Planning & Learning in Robotics
Lecture 10: Bellman Equations I

Lecturer:

Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistants:

Tianyu Wang: tiw161@eng.ucsd.edu

Yongxi Lu: yol070@eng.ucsd.edu

UC San Diego

JACOBS SCHOOL OF ENGINEERING

Electrical and Computer Engineering

Policy Evaluation Theorem

Under the termination state assumption, the cost vector $J^\pi(1), \ldots, J^\pi(n)$ for any proper policy π is the unique solution of:

$$J^\pi(i) = g(i, \pi(i)) + \sum_{j=1}^{n} P_{ij}^{\pi(i)} J^\pi(j). \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

Furthermore, given any initial conditions V_0, the sequence V_k generated by the recursion below converges to J^π:

$$V_{k+1}(i) = g(i, \pi(i)) + \sum_{j=1}^{n} P_{ij}^{\pi(i)} V_k(j), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

- **Proof**: This is a special case of the SSP Bellman Equation Theorem. Consider a modified problem, where the only allowable control at state i is $\pi(i)$. Since the proper policy π is the only policy under consideration, the proper policy assumption is satisfied and the arg min over $u \in \mathcal{U}(i)$ has to be $\pi(i)$.

Value Iteration

- **Value Iteration** (VI): applies the DP recursion with an arbitrary initialization $V_0(i)$ for all $i \in \mathcal{X} \setminus \{0\}$:

$$V_{k+1}(i) = \min_{u \in \mathcal{U}(i)} \Big[g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} V_k(j) \Big], \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

- VI requires an infinite number of iterations for $V_k(i)$ to converge to $J^*(i)$
- In practice, define a threshold for $\| V_{k+1}(i) - V_k(i) \|$ for all $i \in \mathcal{X} \setminus \{0\}$

Policy Iteration

- **Policy Iteration** (PI): iterates the following two steps over policies π instead of values/cost-to-go:
  1. **Policy Evaluation**: Given a policy π, compute J^π by solving the linear system of equations:

$$J^\pi(i) = g(i, \pi(i)) + \sum_{j=1}^{n} P_{ij}^{\pi(i)} J^\pi(j), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

  2. **Policy Improvement**: Obtain a new stationary policy π':

$$\pi'(i) = \arg\min_{u \in \mathcal{U}(i)} \Big[g(i, u) + \sum_{j=1}^{n} P_{ij}^{u} J^\pi(j) \Big], \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

- Repeat the two steps above until $J^{\pi'}(i) = J^\pi(i)$ for all $i \in \mathcal{X} \setminus \{0\}$

Theorem: Optimality of PI

Under the termination state and proper policy assumptions, the PI algorithm converges to an optimal policy after a finite number of steps.

Proof of Optimality of PI (Step 1)

- Let π be a fixed proper policy and $V_0(i) = J^\pi(i)$ for all $i \in \mathcal{X} \setminus \{0\}$. Consider the following recursion in k:

$$V_{k+1}(i) = g(i, \pi'(i)) + \sum_{j=1}^{n} P_{ij}^{\pi'(i)} V_k(j), \qquad i \in \mathcal{X} \setminus \{0\}$$

- Then, for all $i \in \mathcal{X} \setminus \{0\}$:

$$\begin{aligned} J^\pi(i) = V_0(i) &\overset{\text{Policy Evaluation}}{\underset{\text{Theorem}}{=\!=\!=}} g(i, \pi(i)) + \sum_{j=1}^{n} P_{ij}^{\pi(i)} V_0(j) \\ &\overset{\text{Policy}}{\underset{\text{Improvement}}{\geq}} g(i, \pi'(i)) + \sum_{j=1}^{n} P_{ij}^{\pi'(i)} V_0(j) =: V_1(i) \\ &\overset{\text{Since } V_0(i) \geq V_1(i)}{\underset{\text{for all } i \in \mathcal{X} \setminus \{0\}}{\geq}} g(i, \pi'(i)) + \sum_{j=1}^{n} P_{ij}^{\pi'(i)} V_1(j) =: V_2(i) \end{aligned}$$

- Therefore: $V_0(i) \geq V_1(i) \geq V_2(i) \geq \ldots \geq V_k(i)$, for all $i \in \mathcal{X} \setminus \{0\}$

Proof of Optimality of PI (Step 2)

- **Claim**: If π is proper, then π' is proper
- **Proof** (by contradiction): Suppose π' is improper so that $J^{\pi'}(i) = \infty$ for at least one state i as $T \to \infty$. The definition of V_k is the DP recursion after an index substitution $k := T - t$, initialized from $V_0(i) = J^{\pi}(i)$, and with constrained control space $\mathcal{U}(i) = \{\pi'(i)\}$ so that:

$$V_T(i) = \mathbb{E}\left[\sum_{t=0}^{T-1} g(x_t, \pi'(x_t)) + J^{\pi}(x_T) \middle| x_0 = i\right]$$

As $T \to \infty$, the first term above corresponds to $J^{\pi'}(i)$ and we have that $V_T(i) \to \infty$. This contradicts: $V_0(i) \geq V_1(i) \geq V_2(i) \geq \ldots$. Therefore, π' is proper.

Proof of Optimality of PI (Step 3)

- Since π' is proper, by the Policy Evaluation Theorem, the Policy Evaluation step always has a unique solution $J^{\pi'}$. Furthermore, as $k \to \infty$, $V_k \to J^{\pi'}$ and therefore $J^{\pi}(i) \geq J^{\pi'}(i)$ for all $i \in \mathcal{X} \setminus \{0\}$.
- Since the number of stationary policies is finite, eventually we have $J^{\pi} = J^{\pi'}$ after a finite number of steps.
- Once J^{π} has converged, it follows from the Policy Improvement step:

$$J^{\pi'}(i) = J^{\pi}(i) = \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} J^{\pi}(j) \right), \qquad i \in \mathcal{X} \setminus \{0\}$$

- Since this is the Bellman Equation for the SSP problem, we have converged to an optimal policy $\pi^* = \pi$ and the optimal cost $J^* = J^{\pi}$.

Comparison between VI and PI

- PI and VI actually have a lot in common, if we re-write VI as follows:
  2. **Policy Improvement**: Given $V_k(i)$ obtain a stationary policy:

$$\pi(i) = \arg\min_{u \in \mathcal{U}(i)} \Big[g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} V_k(j) \Big], \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

  1. **Value Update**: Given $\pi(i)$ and $V_k(i)$, compute

$$V_{k+1}(i) = g(i, \pi(i)) + \sum_{j=1}^{n} P_{ij}^{\pi(i)} V_k(j), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

- PI performs Policy Evaluation, which solves a system of linear equations and is equivalent to running the Value Update step of VI an infinite number of times!

Comparison between VI and PI

- **Complexity of VI per Iteration**: $O(|\mathcal{X}|^2|\mathcal{U}|)$: evaluating the expectation (i.e., sum over j) requires $|\mathcal{X}|$ operations and there are $|\mathcal{X}|$ minimizations over $|\mathcal{U}|$ possible control inputs.
- **Complexity of PI per Iteration**: $O(|\mathcal{X}|^2\,(|\mathcal{X}| + |\mathcal{U}|))$: the Policy Evaluation step requires solving a system of $|\mathcal{X}|$ equations in $|\mathcal{X}|$ unknowns ($O(|\mathcal{X}|^3)$), while the Policy Improvement step has the same complexity as one iteration of VI.
- PI is more computationally expensive than VI
- Theoretically it takes an infinite number of iterations for VI to converge
- PI converges in $|\mathcal{U}|^{|\mathcal{X}|}$ iterations (all possible policies) in the worst case

Variants: Gauss-Seidel Value Iteration

- A regular VI implementation stores the values from a previous iteration and updates them for all states simultaneously:

$$\bar{V}(i) \leftarrow \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} V(j) \right), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

$$V(i) \leftarrow \bar{V}(i), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

- **Gauss-Seidel Value Iteration** updates the values in place:

$$V(i) \leftarrow \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} V(j) \right), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

- Gauss-Seidel VI often leads to faster convergence and requires less memory than VI

Variants: Asynchronous/Generalized Policy Iteration

- Assuming that the Value Update and Policy Improvement steps are executed an infinite number of times for all states, all combinations of the following converge:
  - Any number of Value Update steps in between Policy Improvement steps
  - Any number of states updated at each Value Update step
  - Any number of states updated at each Policy Improvement step

Connections to Linear Algebra (SSP)

- In the Policy Evaluation Theorem and in PI's Policy Evaluation step, we are essentially solving a linear system of equations:

$$\mathbf{v} = \mathbf{g} + P\mathbf{v} \qquad \Rightarrow \qquad (I - P)\mathbf{v} = \mathbf{g}$$

where for $i, j = 1, \dots, n$, $\mathbf{v}_i := J^\pi(i)$, $\mathbf{g}_i := g(i, \pi(i))$, $P_{ij} := P_{ij}^{\pi(i)}$.

- There exists a unique solution for $\mathbf{v}$, iff $(I - P)$ is invertible. This is guaranteed as long as π is a proper policy.

- **Proof**: $(I - P)$ is invertible iff P does not have eigenvalues at 1. By the Chapman-Kolmogorov equation, $[P^T]_{ij} = \mathbb{P}(x_T = j \mid x_0 = i)$ and since π is proper, $[P^T]_{ij} \to 0$ as $T \to \infty$ for all $i, j \in \mathcal{X} \setminus \{0\}$. Since P^T vanishes as $T \to \infty$ all eigenvalues of P must have modulus less than 1 and therefore $(I - P)$ exists.

Connections to Linear Algebra (SSP)

- The Policy Evaluation Thm is an iterative solution to $(I - P)\mathbf{v} = \mathbf{g}$:

$$
\begin{aligned}
\mathbf{v}_1 &= \mathbf{g} + P\mathbf{v}_0 \\
\mathbf{v}_2 &= \mathbf{g} + P\mathbf{v}_1 = \mathbf{g} + P\mathbf{g} + P^2\mathbf{v}_0 \\
&\vdots \\
\mathbf{v}_T &= (I + P + P^2 + P^3 + \ldots + P^{T-1})\mathbf{g} + P^T\mathbf{v}_0 \\
&\vdots \\
\mathbf{v}_\infty &\to (I - P)^{-1}\mathbf{g}
\end{aligned}
$$

Connections to Linear Algebra (Discounted Problem)

- We can obtain a Policy Evaluation Theorem for the Discounted problem through the SSP equivalence
- As before, define an auxiliary SSP by introducing a virtual terminal state 0 and transitions $\tilde{P}^u_{ij} = \gamma P^u_{ij}$, $\tilde{P}^u_{i,0} = 1 - \gamma$, $\tilde{P}^u_{0,0} = 1$, $\tilde{P}^u_{0,j} = 0$.
- The Policy Evaluation Theorem for the auxiliary SSP is: $\mathbf{v} = \mathbf{g} + \tilde{P}\mathbf{v}$
- This leads to a Policy Evaluation Theorem for the Discounted problem:

$$\mathbf{v} = \mathbf{g} + \gamma P \mathbf{v} \qquad \Rightarrow \qquad (I - \gamma P)\mathbf{v} = \mathbf{g}$$

  where P is the transition kernel of the Discounted problem under the policy π, equivalent with the SSP policy $\tilde{\pi}$.

- The matrix P has eigenvalues with modulus ≤ 1. Hence, all eigenvalues of γP must have modulus < 1, so that $(\gamma P)^T \to 0$ as $T \to \infty$ and $(I - \gamma P)^{-1}$ exists.

Connections to Linear Algebra (Summary)

- Let $\mathbf{v}_i := J^\pi(i)$, $\mathbf{g}_i := g(i, \pi(i))$, $P_{ij} := P_{ij}^{\pi(i)}$ for $i, j = 1, \ldots, n$
- **Finite Horizon**: $\mathbf{v}_t = \mathbf{g}_t + P_t \mathbf{v}_{t+1}$ starting from $\mathbf{v}_T = \mathbf{g}_T$
- **SSP (First Exit)**: Let $\mathcal{T} \subseteq \mathcal{X}$ be the set of terminal states and $\mathcal{N} \subseteq \mathcal{X}$ be the set of nonterminal states. The cost-to-go/value of policy π is:

$$(I - P_{\mathcal{N}\mathcal{N}})\,\mathbf{v}_{\mathcal{N}} = \mathbf{g}_{\mathcal{N}} + P_{\mathcal{N}\mathcal{T}}\mathbf{g}_{\mathcal{T}}$$

- **Discounted**: $(I - \gamma P)\mathbf{v} = \mathbf{g}$

Connections to Linear Programming

- Suppose we initialize VI with a vector V_0 that satisfies a relaxed Bellman Equation:

$$V_0(i) \leq \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}(u) V_0(j) \right), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

- Applying VI to V_0 leads to:

$$V_1(i) = \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}(u) V_0(j) \right) \geq V_0(i), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

$$V_2(i) = \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}(u) V_1(j) \right)$$

$$\geq \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}(u) V_0(j) \right) = V_1(i), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

Connections to Linear Programming

- The above shows that $V_{k+1}(i) \geq V_k(i)$ for all k and $i \in \mathcal{X} \setminus \{0\}$
- Since VI guarantees that $V_k(i) \to J^*(i)$ as $k \to \infty$ we also have:

$$J^*(i) \geq V_0(i), \quad \forall i \in \mathcal{X} \setminus \{0\} \quad \Rightarrow \quad \sum_{i \in \mathcal{X} \setminus \{0\}} w_i J^*(i) \geq \sum_{i \in \mathcal{X} \setminus \{0\}} w_i V_0(i)$$

  for any $w_i > 0$ for all $i \in \mathcal{X} \setminus \{0\}$.

- The above holds for **any** V_0 that satisfies:

$$V_0(i) \leq \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}(u) V_0(j) \right), \qquad \forall i \in \mathcal{X} \setminus \{0\}$$

- Note that J^* also satisfies this condition with equality (Bellman Equation) and hence is the maximal V_0 (at each state) that satisfies the condition.

Linear Programming Solution to the Bellman Equation

The solution V^* to the linear program (with $w_i > 0$):

$$\max_V \sum_{i \in \mathcal{X} \setminus \{0\}} w_i V(i)$$

$$\text{s.t.} \quad V(i) \leq \left(g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} V(j) \right), \qquad \forall u \in \mathcal{U}(i), \forall i \in \mathcal{X} \setminus \{0\}$$

also solves the Bellman Equation to yield the optimal cost J^* for SSP.

Proof: LP Solution to the BE

- Let V^* be the solution to the linear program so that:

$$V^*(i) \leq \left(g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} V^*(j) \right), \qquad \forall u \in \mathcal{U}(i), \forall i \in \mathcal{X} \setminus \{0\}$$

- This implies that $V^*(i) \leq J^*(i)$ for all $i \in \mathcal{X} \setminus \{0\}$. By contradiction, suppose that $V^* \neq J^*$. Then, there exists a state $l \in \mathcal{X} \setminus \{0\}$ such that:

$$V^*(l) < J^*(l) \quad \Rightarrow \quad \sum_{i \in \mathcal{X} \setminus \{0\}} w_i V^*(i) < \sum_{i \in \mathcal{X} \setminus \{0\}} w_i J^*(i)$$

  for any positive w_i but since J^* solves the Bellman Equation:

$$J^*(i) \leq \left(g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} J^*(j) \right), \qquad \forall u \in \mathcal{U}(i), \forall i \in \mathcal{X} \setminus \{0\}$$

- Thus, V^* is not the optimal solution, which is a contradiction.