IE 3186: Approximate Dynamic Programming **Fall 2018**

Lecture 3: Variants of Value Iteration Algorithm

Lecturer: Daniel Jiang *Scribes: Ziyue Sun, Tarik Bilgic*

References:

D.P. Bertsekas. *Dynamic Programming and Optimal Control: Approximate Dynamic Programming*, Vol. 2, 4th ed, Athena Scientific, Belmont MA, 2012. (§2.2)

3.1 Bounds on Value Iteration

Recall the value iteration algorithm:

1. Set $J_0 \in \mathbb{R}^{|\mathcal{X}|}$ arbitrarily.
2. On iteration $k+1$, set $J_{k+1} = (TJ_k)$ for all x. In other words, for each k and each $x \in \mathcal{X}$,
$$J_{k+1}(x) = \min_{u \in \mathcal{U}(x)} \mathbf{E}\left[g(x,u,w) + \gamma J_k(x,u,w)\right]$$

Also recall that $\lim_{k\to\infty} T^k J_0 = J^*$, which follows by to the contraction property:

$$\left\|J_{k+1} - J^*\right\|_\infty \leqslant \gamma \left\|J_k - J^*\right\|_\infty = \max_x \left|J_{k+1}(x) - J^*(x)\right|.$$

Can we say something about the progress of VI at the progress of VI at time k? Let the state space be $\mathcal{X} = \{1, 2, \ldots, n\}$. Denote

$$\epsilon = \min_{i \in \mathcal{X}} \left[(TJ)(i) - J(i)\right].$$

Then, applying T on both sides of $J + \epsilon e \leqslant TJ$ gives

$$\begin{aligned} T(J+\epsilon e) &\leqslant T^2 J \quad \text{(monotonicity)} \\ TJ + \gamma\epsilon e &\leqslant T^2 J \quad \text{(constant shift)}. \end{aligned}$$

It follows from the previous two equations that

$$J + \epsilon e + \gamma \epsilon e \leqslant TJ + \gamma\epsilon e \leqslant T^2 J. \tag{3.1}$$

Repeat the steps again to get $TJ + (\gamma + \gamma^2)\epsilon e \leqslant T^2 J + \gamma^2 \epsilon e \leqslant T^3 J$, which results in

$$J + \left(\sum_{i=0}^{k} \gamma^i \right) \epsilon e \leqslant TJ + \left(\sum_{i=1}^{k} \gamma^i \right) \epsilon e \leqslant T^2 J + \left(\sum_{i=2}^{k} \gamma^i \right) \epsilon e.$$

Let $k \to \infty$:

$$J + \frac{\epsilon e}{1-\gamma} \leqslant TJ + \frac{\gamma \epsilon e}{1-\gamma} \leqslant T^2 J + \frac{\gamma^2 \epsilon e}{1-\gamma} \leqslant J^*. \tag{3.2}$$

Let $T^k J$ replace J (and using the $T^{k+1}J - T^k J$ version of ϵ):

$$T^{k+1}J + \underbrace{\frac{\gamma}{1-\gamma} \min_i \left(T^{k+1}J(i) - T^k J(i) \right) e}_{\underline{c}_{k+1}} \leqslant J^*. \tag{3.3}$$

Note: this relates $T^{k+1}J$ (J_{k+1} in VI) to J^* in the terms of a quantity related to the Bellman error, which is observable at any iteration k. On the other hand, the distance to optimal is not observable.

From (3.1): $TJ + \gamma \epsilon e \leqslant T^2 J$, letting $T^{k-1}J$ replace J, we have

$$T^k J + \gamma e \left(\frac{1-\gamma}{\gamma} \underline{c}_k \right) \leqslant T^{k+1} J,$$

from which we conclude

$$\gamma \left(\frac{1-\gamma}{\gamma} \right) \underline{c}_k \leqslant \min_i \left(T^{k+1}J(i) - T^k J(i) \right) \leqslant \frac{1-\gamma}{\gamma} \underline{c}_{k+1}$$

and $\gamma \underline{c}_k \leqslant \underline{c}_{k+1}$. Based on (3.2) and (3.3),

$$T^k J + \frac{\underline{c}_{k+1}}{\gamma} \leqslant T^{k+1} J + \underline{c}_{k+1} \leqslant J^*,$$
$$T^k J + \underline{c}_k \leqslant T^{k+1} J + \underline{c}_{k+1} \leqslant J^*.$$

Proposition 3.1 (Monotonic error bound for VI)**.** *For any value function J, state i, and iteration k:*

$$\begin{aligned} \left(T^k J\right)(i) + \underline{c}_k &\leqslant \left(T^{k+1} J\right)(i) + \underline{c}_{k+1} \leqslant J^*(i) \\ &\leqslant \left(T^{k+1} J\right)(i) + \bar{c}_{k+1} \\ &\leqslant \left(T^k J\right)(i) + \bar{c}_k \end{aligned}$$

where $\bar{c}_k = \gamma/(1-\gamma) \max_i \left[\left(T^k J\right)(i) - \left(T^{k-1} J\right)(i) \right]$ and $\underline{c}_k$ is the same as above.

Note that both $\underline{c}_k$ and $\bar{c}_k$ converge to J^* by VI. This proposition allows the process of VI to be evaluated by the Bellman error.

Example 3.2. *Here is an example of how this process might look for a simple two-state, two-action MDP.*

k	$\left(T^kT\right)(1)+\underline{c}_k$	$\left(T^kT\right)(1)+\bar{c}_k$	$\left(T^kT\right)(2)+\underline{c}_k$	$\left(T^kT\right)(2)+\bar{c}_k$
0	5	4.5	5.5	10.5
3	6.8	7.8	7.2	8.1
6	7.3	7.4	7.5	7.6
		⋮		
15	7.328	7.328	7.572	7.572

3.1.1 Performance of Greedy Policy

Suppose we stop VI at some J. Just because J is close to J^* does not make it immediately clear that $\mu = \text{greedy}(J)$ is a good policy. A simple analysis: by the Proposition 1, we have:

$$\underline{c}_1 \leqslant J^*(i) - (TJ)(i) \leqslant \bar{c}_1 \tag{3.4}$$

Let $J_\mu(i)$ be the value of the greedy policy with respect to J. Applying the proposition with $k=1$ and T_μ replacing T, we have

$$\underline{c}_1 \leqslant J_\mu(i) - (T_\mu J)(i) \leqslant \bar{c}_1 \tag{3.5}$$

Rearranging (3.4) and (3.5),

$$\begin{aligned} J_\mu(i) &\leqslant \bar{c}_1 + (T_\mu J)(i) \\ -J^*(i) &\leqslant -\underline{c}_1 - (TJ)(i). \end{aligned}$$

Adding the two inequalities and then maximizing over states yields

$$\max_i \left(J_\mu(i) - J^*(i)\right) \leqslant \frac{\gamma}{1-\gamma}\left\{\max_i \left(J_\mu(i) - J^*(i)\right) - \min_i \left(J_\mu(i) - J^*(i)\right)\right\}.$$

3.1.2 Removing Suboptimal Actions

Can we speed up VI by removing suboptimal actions? Note that $\tilde{\mu}$ is suboptimal if

$$\mathbf{E}\left[g\left(i,\tilde{\mu}\right) + \gamma \underline{J}^*\left(f\left(i,\tilde{\mu},w\right)\right)\right] > J^*(i).$$

Let's say $\underline{J} \leqslant J^* \leqslant \bar{J}$. Then if $\mathbf{E}\left[g\left(i,\tilde{\mu}\right) + \gamma J\left(f\left(i,\tilde{\mu},w\right)\right)\right] > \overline{J}(i)$, $\tilde{\mu}$ is suboptimal. Remove $\tilde{\mu}$ from $\mathcal{U}(i)$.

3.2 Gauss-Seidel Version of Value Iteration

The update step $J_{k+1} = TJ_k$ means the Bellman operator T is applied simultaneously to all states. In reality, we use looping through the states one by one. Why not use the newest information (i.e., update J as soon as you complete the Bellman optimization step)? In the Gauss-Seidel version of VI, iterations are made one-state at a time.

- $p_{ij}(u)$: Probability of going to state j, starting from state i, by taking action u *(Transition probability notation)*
- $g(i,u) = \mathbf{E}[g(i,u,w)]$,
- Fixed order of state updates: states $1, 2, 3, \ldots, n, 1, 2, \ldots$,
- Operator W (similar to T in that $W : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}^{|\mathcal{X}|}$):

$$\begin{aligned}(WJ)(1) &= \min_{u \in \mathcal{U}(1)} g(1,u) + \gamma \sum_{j=1}^{n} p_{ij}(u)J(j) \\ &= (TJ)(1)\end{aligned}$$

For $i = 2, 3, \ldots, n$:

$$(WJ)(i) = \min_{u \in \mathcal{U}(i)} [g(i,u)] + \gamma \sum_{j<i} p_{ij}(u)WJ(j) + \gamma \sum_{j \geq i} p_{ij}(u)J(j)$$

The Gauss-Seidel V.I. proceeds via the iterations $J, WJ, W^2J, \ldots$

Proposition 3.3 (Convergence of Gauss-Seidel algorithm)**.** *For any value functions J, J' and all iterations k:*

$$||W^kJ - W^kJ'||_\infty \leq \gamma^k ||J - J'||_\infty.$$

Furthermore;

$$\begin{aligned}WJ^* &= J^* \\ \lim_{k\to\infty} W^kJ &= J^*.\end{aligned}$$

Proof. Consider $k = 1$. By definition,

$$||(WJ)(1) - (WJ')(1)||_\infty \leq \gamma ||J - J^*||_\infty \text{ by contraction property of } T.$$

Assume the equation above is true for $i = 1, \ldots, m-1$, and we will try to show the result for m:

$$\begin{aligned}|(WJ)(m) - (WJ')(m)| &\leq \gamma \max\{|(WJ)(1) - (WJ')|, \ldots, |(WJ)(m) - (WJ')(m))|, \\ &\qquad |J(m+1) - J'(m+1)|, \ldots, |J(n) - J'(n)|\} \\ &\leq \gamma \max_i \{\gamma ||J - J'||, ||J - J'||\} \\ &\leq \gamma\, ||J - J'||_\infty.\end{aligned}$$

The fixed point property $WJ^* = J^*$ follows by $TJ^* = J^*$ and the convergence to J^* follows by Banach's fixed point theorem. □

Proposition 3.4 (Comparison of G.S. and V.I.)**.** *Suppose that $J \leq TJ$. Then*

$$T^k W \leq W^k J \leq J^*,$$

which means that G.S. is at least as fast as V.I.

Proof. $T^0 J \leq W^0 J$ and assume $T^{k-1} J \leq W^{k-1} J$. Prove for k:

$$\begin{aligned}
(T^k J)(1) &= \min_u \Big[g(1,u) + \sum_j p_{1j}(u)(T^{k-1}J)(j) \Big] \\
&\leq \min_u \Big[g(1,u) + \sum_j p_{1j}(u)(W^{k-1}J)(j) \Big] \\
&\leq (W^k J(1))
\end{aligned}$$

Suppose true for states $i = 1, 2, \dots, m-1$

$$\begin{aligned}
(T^k J)(m) &= \min_u \Big[g(m,u) + \sum_{j<m} p_{mj}(u)(T^{k-1}J)(j) + \sum_{j>m} p_{mj}(u)(T^{k-1}J)(j \Big] \\
&\leq \min_u \Big[g(m,u) + \sum_{j<m} p_{mj}(u)(T^k J)(j) + \sum_{j>m} p_{mj}(u)(T^k J)(j) \Big] \\
&\leq \min_u \Big[g(m,u) + \sum_{j<m} p_{mj}(u)(W^k J)(j) + \sum_{j>m} p_{mj}(u)(W^{k-1}J)(j) \Big] \\
&= (W^k J)(m)
\end{aligned}$$

So, we conclude that $T^k J \leq W^k J$ for all k. In addition, since $J \leq TJ \leq WJ$, repeatedly applying W gives $J \leq WJ \leq W^2 J \leq \ldots \leq J^*$, which implies $T^k J \leq W^k J \leq J^*$. □