Outline

1. Introduction to Infinite Horizon Problems
2. Transition Probability Notation - Main Results
3. SSP Problems: Elaboration
4. Algorithms - Approximate Value Iteration
Infinite number of stages, and stationary system and cost

- System \( x_{k+1} = f(x_k, u_k, w_k) \) with state, control, and random disturbance.
- Policies \( \pi = \{\mu_0, \mu_1, \ldots\} \) with \( \mu_k(x) \in U(x) \) for all \( x \) and \( k \).
- Special scalar \( \alpha \) with \( 0 < \alpha \leq 1 \). If \( \alpha < 1 \) the problem is called discounted.
- Cost of stage \( k \): \( \alpha^k g(x_k, \mu_k(x_k), w_k) \).
- Cost of a policy \( \pi = \{\mu_0, \mu_1, \ldots\} \)
  \[
  J_\pi(x_0) = \lim_{N \to \infty} E_{w_k} \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}
  \]
- Optimal cost function \( J^*(x_0) = \min_\pi J_\pi(x_0) \).
- If \( \alpha = 1 \) we assume a special cost-free termination state \( t \). The objective is to reach \( t \) at minimum expected cost. The problem is called stochastic shortest path (SSP) problem.
Main Results: Intuitive Justification (Math Proof Required)

Value iteration (VI) convergence: Fix horizon $N$, let terminal cost be 0

- Let $V_{N-k}(x)$ be the optimal cost starting at $x$ with $k$ stages to go, so
  \[ V_{N-k}(x) = \min_{u \in U(x)} E_w \left\{ \alpha^{N-k} g(x, u, w) + V_{N-k+1}(f(x, u, w)) \right\} \]

- Reverse the time index: Define $J_k(x) = V_{N-k}(x)/\alpha^{N-k}$ and divide with $\alpha^{N-k}$:
  \[ J_k(x) = \min_{u \in U(x)} E_w \left\{ g(x, u, w) + \alpha J_{k-1}(f(x, u, w)) \right\} \quad (VI) \]

- $J_N(x)$ is equal to $V_0(x)$, which is the $N$-stages optimal cost starting from $x$

- Hence, intuitively, VI converges to $J^*$:
  \[ J^*(x) = \lim_{N \to \infty} J_N(x), \quad \text{for all states } x \quad (??) \]

The following Bellman equation holds: Take the limit in Eq. (VI)

\[ J^*(x) = \min_{u \in U(x)} E_w \left\{ g(x, u, w) + \alpha J^*(f(x, u, w)) \right\}, \quad \text{for all states } x \quad (??) \]

Optimality condition: Let $\mu(x)$ attain the min in the Bellman equation for all $x$

The policy $\{\mu, \mu, \ldots\}$ is optimal (??). (This type of policy is called stationary.)
Transition Probability Notation for Finite-State Problems

- States: $i = 1, \ldots, n$. Successor states: $j$. (For SSP there is also the extra termination state $t$.)
- Probability of $i \rightarrow j$ transition under control $u$: $p_{ij}(u)$
- Cost of $i \rightarrow j$ transition under control $u$: $g(i, u, j)$

VI (translated to the new notation - note that $J_k(t) = 0$ for SSP)

$$J_{k+1}(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) (g(i, u, j) + \alpha J_k(j))$$

Bellman equation (translated to the new notation - note that $J^*(t) = 0$ for SSP)

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

$$J^*(i) = \min_{u \in U(i)} \left[ p_{it}(u) g(i, u, t) + \sum_{j=1}^{n} p_{ij}(u) (g(i, u, j) + J^*(j)) \right] \quad \text{(for SSP)}$$
Convergence of VI

Given any initial conditions $J_0(1), \ldots, J_0(n)$, the sequence $\{J_k(i)\}$ generated by VI

$$J_{k+1}(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) (g(i, u, j) + \alpha J_k(j)),$$

converges to $J^*(i)$ for each $i$.

Bellman’s equation

The optimal cost function $J^* = (J^*(1), \ldots, J^*(n))$ satisfies the equation

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

and is the unique solution of this equation.

Optimality condition

A stationary policy $\mu$ is optimal if and only if for every state $i$, $\mu(i)$ attains the minimum in the Bellman equation.
Assumption (Termination Inevitable Under all Policies)

There exists $m > 0$ such that regardless of the policy used and the initial state, there is positive probability that $t$ will be reached within $m$ stages; i.e., for all $\pi$

$$\max_{i=1, \ldots, n} P\{x_m \neq t \mid x_0 = i, \pi\} < 1.$$ 

VI Convergence: $J_k \to J^*$ for all initial conditions $J_0$, where

$$J_{k+1}(i) = \min_{u \in U(i)} \left[ p_{it}(u)g(i, u, t) + \sum_{j=1}^{n} p_{ij}(u)(g(i, u, j) + J_k(j)) \right], \quad i = 1, \ldots, n$$

Bellman’s equation: $J^*$ satisfies

$$J^*(i) = \min_{u \in U(i)} \left[ p_{it}(u)g(i, u, t) + \sum_{j=1}^{n} p_{ij}(u)(g(i, u, j) + J^*(j)) \right], \quad i = 1, \ldots, n,$$

and is the unique solution of this equation.

Optimality condition: $\mu$ is optimal if and only if for every $i$, $\mu(i)$ attains the minimum in the Bellman equation.
A discounted problem can be converted to an SSP problem, since the stage $k$ cost is identical in both problems, under the same policy.

Proof line of text: Start with SSP analysis, get discounted analysis as special case.

Key proof argument: The tail portion ($k$ to $\infty$) of the infinite horizon cost diminishes to 0, as $k \to \infty$, at a geometric progression rate (so the finite horizon costs converge to the infinite horizon cost).

A more general assumption for our results: Nonterminating policies are “bad”

- Every stationary policy under which termination is not inevitable from some initial states is “bad," in the sense that it has $\infty$ cost for some initial states.
- There exists at least one stationary policy under which termination is inevitable.
SSP Problems can be Tricky

Without the assumption on nonterminating policies

- Bellman equation may have any number of solutions: one, infinitely many, or none.
- Bellman equation may have one or more solutions, but $J^*$ is not a solution.
- VI may converge to $J^*$ from some initial conditions but not from others.

Challenge questions: Consider the cases $a > 0$, $a = 0$, and $a < 0$

- What is $J^*(1)$?
- What is the solution set of Bellman’s equation $J(1) = \min [b, a + J(1)]$?
- What is the limit of the VI algorithm $J_{k+1}(1) = \min [b, a + J_k(1)]$?
Answers to the Challenge Questions

Bellman Eq: $J(1) = \min \left[ b, a + J(1) \right]$; VI: $J_{k+1}(1) = \min \left[ b, a + J_k(1) \right]$

- If $a > 0$ (positive cycle): $J^*(1) = b$ is the unique solution, and VI converges to $J^*(1)$. Here the “nonterminating policies are bad" assumption is satisfied.
- If $a = 0$ (zero cycle):
  - $J^*(1) = \min[0, b]$.
  - The solution set of the Bellman equation is $=(-\infty, b]$.
  - The VI algorithm, $J_{k+1}(1) = \min \left[ b, J_k(1) \right]$, converges to $b$ starting from $J_0(1) \geq b$, and does not move from a starting value $J_0(1) \leq b$.
- If $a < 0$ (negative cycle): B-Eq has no solution, and VI diverges to $J^*(1) = -\infty$. 

Two possible controls at state 1 (costs $a$ and $b$)
Results Involving Q-Factors - Discounted Problems

**VI for Q-factors**

\[
Q_{k+1}(i, u) = \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \alpha \min_{v \in U(j)} Q_k(j, v) \right)
\]

converges to \(Q^*(i, u)\) for each \((i, u)\).

**Bellman’s equation for Q-factors**

\[
Q^*(i, u) = \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \alpha \min_{v \in U(j)} Q^*(j, v) \right)
\]

\(Q^*\) is the unique solution of this equation, and we have

\[
J^*(i) = \min_{u \in U(i)} Q^*(i, u)
\]  
(1)

**Optimality condition**

A stationary policy \(\mu\) is optimal if and only if \(\mu(i)\) attains the minimum in Eq. (1) for every state \(i\).
Consider VI with sequential approximation (fitted VI - a neural net may be used). Assume that for some $\delta > 0$

\[
\max_{i=1,\ldots,n} \left| \tilde{J}_{k+1}(i) - \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)(g(i, u, j) + \alpha \tilde{J}_k(j)) \right| \leq \delta
\]  

(1)

The cost function error is:

\[
\max_{i=1,\ldots,n} \left| \tilde{J}_k(i) - J^*(i) \right|
\]

Can be shown to be $\leq \delta/(1 - \alpha)$ (asymptotically, as $k \to \infty$).

... but this result may not be meaningful; it may be difficult to maintain Eq. (1) over an infinite horizon.

In particular, suppose $\tilde{J}_{k+1}$ is obtained using a parametric architecture:

- Start with $\tilde{J}_0$.
- Given parametric approximation $\tilde{J}_k$, obtain a parametric approximation $\tilde{J}_{k+1}$ using a least squares fit.
- We will give an example where the cost function error accumulates to $\infty$.  

Bertsekas Reinforcement Learning
Bad Example for Fitted VI

Bellman Eq: \( J(1) = \alpha J(2), \ J(2) = \alpha J(2) \)
\[ J^*(1) = J^*(2) = 0 \]

Exact VI: \( J_{k+1}(1) = \alpha J_k(2), \ J_{k+1}(2) = \alpha J_k(2) \)

Approximate VI iterate
\( \tilde{J}_{k+1} \)
\( \tilde{J}_k = (r_k, 2r_k) \)
\( J^* = (0, 0) \)

Approximation Subspace
Orthogonal Projection

By using a weighted projection we may correct the problem.
We will cover:

- Infinite horizon policy iteration without approximations
- Infinite horizon policy iteration with approximations
- Rollout and parametric approximation methods
- We will likely need more than one lecture

PLEASE READ AS MUCH OF SECTIONS 4.5-4.7 AS YOU CAN

PLEASE DOWNLOAD THE LATEST VERSIONS FROM MY WEBSITE