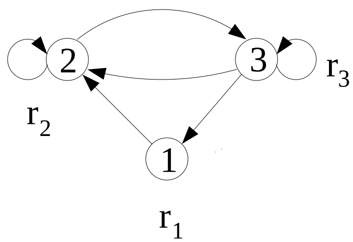
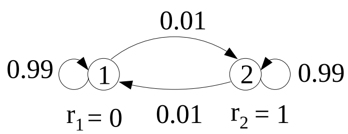
Review: Finite State Markov Chain with Rewards



- $S = \{1, ..., J\}$
- Reward r_i in state $i \in S$
- $\{X_n; n \geq 0\}$
- $\{R_n; n \geq 0\}$

- $v_i(n) = E\Big[\sum_{k=0}^{n} R_k \mid X_o = i\Big]$
- $g = \lim_{n\to\infty} \frac{1}{n} E\Big[\sum_{k=0}^{n-1} R_k \mid X_o = i\Big]$
- Single Recurrent Class: $g = \sum_{i=1}^{J} \pi_i r_i$
- Focus on $v_i(n)$
- Example



  - $\lim_{n\to\infty} \frac{1}{n} E\Big[\sum_{k=1}^{n} R_k \mid X_o = i\Big] = g = \sum_{i=1}^{J} \pi_i r_i$
  - Time-average is the same no matter where we start.
  - What about transient behavior: $v_1(n) - v_2(n)$?

Relative Gain: Results

- Assumption: Single recurrent class and perhaps some transient states
- Let π be the steady state probability vector and let

$$g = \sum_i r_i \pi_i$$

- Let the vector w be a solution to $w + ge = r + [P]w$
- Then

$$v(n) = nge + w + [P]^n \Big\{v(0) - w\Big\}$$

Markov Decision Theory

- At each state i we can choose between K_i actions where each actions is characterized by a reward $r_i^{(k)}$ and transition probability $P_{ij}^{(k)}$, $j = 1, ..., J$.
- Policy: "Decide for each state $i \in S$ which action k, $k = 1, ..., K_i$ to apply at state i"
- Stationary policy: decision does not depend on time n
- Dynamic Policy
- For a given stationary policy A, we have $g^A = \sum_i \pi_i^A r_i^A$
- Questions
  - Optimal Dynamic Policy
  - Optimal Stationary Policy

Dynamic Programming: Optimal Dynamic Policy

- Optimal decision at time 1:

$$v_i^*(1) = \max_{k=1,\ldots,K_i} \left\{ r_i^{(k)} + \sum_j P_{ij}^{(k)} v_j(0) \right\}$$

- Optimal decision at time 2:

$$v_i^*(2) = \max_{k=1,\ldots,K_i} \left\{ r_i^{(k)} + \sum_j P_{ij}^{(k)} v_j^*(1) \right\}$$

- Optimal decision at time n:

$$v_i^*(n) = \max_{k=1,\ldots,K_i} \left\{ r_i^{(k)} + \sum_j P_{ij}^{(k)} v_j^*(n-1) \right\}$$

or

$$v_i^*(n) = \max_A \left\{ r^A + [P^A] v^*(n-1) \right\}$$

- Note: finite number of policies
- Dynamic programming algorithm
- Conceptually easy
- What about asymptotic behavior as n becomes large?

Optimal Stationary Policy

- Assumption: For all policies A the Markov chain with $[P^A]$ is recurrent. Strong Assumption!
- For fixed policy A

$$v^A(n) = ng^A e + w^A + [P^A]^n \left\{ v(0) - w^A \right\}$$

where

$$w^A + g^A e = r^A + [P^A] w^A$$

- Goal: Find policy B such that $g^B \geq g^A$ for all policies A.

Optimal Stationary Policy: Guess

- Intuition: For $n \to \infty$, optimal dynamic policy becomes optimal stationary policy
- Suppose for $n \geq m$, the optimal policy is always B, then for large n we have (see Theorem 7 in Chapter 4)

$$v^*(n) \approx v^B(n) \approx ng^B e + w^B + \beta e$$

- Because B is optimal policy for $n \geq m$, for any policy A we have

$$r^B + [P^B] v^*(n) \geq r^A + [P^A] v^*(n)$$

- Using $v^*(n) \approx ng^B e + w^B + \beta e$, we have

$$r^B + [P^B] w^B \geq r^A + [P^A] w^B$$

- Is such a policy B an optimal stationary policy?

Optimal Stationary Policy: Analysis

- **Lemma**: If $v(0) = w^B$ and

$$r^B + [P^B]w^B \geq r^A + [P^A]w^B$$

for all policies A, then policy B is an optimal dynamic policy and

$$v^*(n) = w^B + ng^B e$$

- **Theorem** The policy B is an optimal stationary policy if and only if

$$r^B + [P^B]w^B \geq r^A + [P^A]w^B$$

for all policies A.

Howard's Policy Improvement Algorithm

1. Choose an arbitrary initial policy B
2. Calculate w^B
3. If $r^B + [P^B]w^B \geq r^A + [P^A]w^B$ for all policies A, then stop - B is an optimal stationary policy
4. Otherwise, choose a policy A such that

$$r^A + [P^B]w^B \geq, \neq r^B + [P^A]w^B$$

5. Update policy B to the new policy A and go to step (2)

Optimal Dynamic Policy vs. Optimal Stationary Policy

- Difference in $v(n)$ between optimal dynamic policy and optimal stationary policy?
- **Lemma** Let $v(0)$ and $v'(0)$ be such that $v(0) \leq v'(0)$. Then, for any stationary policy A we have

$$v^A(n) \leq v'^A(n).$$

Similarly, for an optimal dynamic policy we have

$$v^*(n) \leq v'^*(n).$$

- **Lemma** For an optimal stationary policy B, the function

$$f(n) = \pi^B v^*(n) - ng^B$$

is monotonic non-decreasing in n and has some limit β'

Optimal Dynamic Policy vs. Optimal Stationary Policy

- **Theorem** Assume that B is an optimal stationary policy and that the Markov chain with $[P^B]$ is ergodic. Then

$$\lim_{n\to\infty} v^*(n) - ng^B = w^B + \left(\beta' - \pi^B w^B\right)e$$

where β' is the constant from the above lemma.