Approximating Sequence Method

MDP with Infinite State Space

Cf. Linn Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley Series in Probability & Statistics, 1999.

D.L.Bricker, 2001
Dept of Industrial Engineering
The University of Iowa

Assume that the state space of a Markov Decision Problem (MDP) is countable but *infinite*.

Four different optimization criteria are considered:

Cases	Expected discounted costs	Average cost/stage
Finite horizon	1	2
Infinite horizon	3	4

1. Expected discounted cost over finite horizon
2. Expected cost/stage over finite horizon
3. Expected discounted cost over infinite horizon
4. Expected cost/stage over infinite horizon

Denote the original MDP by **Δ**, with infinite (but countable) state space S.

It is common, for computational purposes, to approximate **Δ** by a MDP with *finite* state space of size N.

As N is increased, the approximating MDP is "improved".

We are interested in the limit as N→∞.

Approximating Sequence

Definition

Consider the sequence $\{\Delta_N\}_{N\geq N_0}$ of MDPs, where

- the state space of Δ_N is the nonempty *finite* set $S_N \subset S$,
- the action set for state $i\in S_N$ is A_i, and
- the cost for action $a\in A_i$ is C_i^a.

Let $\{S_N\}_{N\geq N_0}$ be an increasing sequence of subsets of S such that

- $\underset{N}{\cup} S_N = S$, and
- for each $i\in S_N$ and $a\in A_i$, $P_{i\cdot}^a(N)$ is a probability distribution on S_N such that $\lim_{N\to\infty} P_{ij}^a(N) = P_{ij}^a$

Then $\{\Delta_N\}_{N\geq N_0}$ is an ***approximating sequence*** (AS) for the MDP Δ, and N is the *approximation level.*

Augmentation Procedure

The usual way to define an approximating distribution is by means of an ***augmentation procedure***:

Suppose that in state $i\in S_N$, action $a\in A_i$ is chosen.

For $j\in S_N$ the probability P_{ij}^a is unchanged.

Suppose, however, that $P_{ir}^a>0$ for some $r\notin S_N$,

i.e., there is a positive probability that the system makes a transition to a state outside of S_N.

This is said to be ***excess probability*** associated with (i,a,r,N).

Augmentation Procedure

In order to define a valid MDP, this excess probability must be distributed among the states of S_N according to some specified ***augmentation distribution*** q_j(i,a,r,N), where

$$\sum_j q_j(i,a,r,N) = 1 \text{ for each (i,a,r,N).}$$

The quantity q_j(i,a,r,N) specifies what portion of the excess probability P_{ir}^a is redistributed to state $j\in S_N$.

Augmentation Procedure

Definition: The approximating sequence $\{\Delta_N\}$ is an *augmentation-type approximating sequence* (**ATAS**) if the approximating distributions are defined as follows:

$$P_{ij}^a(N) = P_{ij}^a + \sum_{r\notin S_N} P_{ij}^a q(i,a,r,N)$$

Notes:

- The original probabilities on S_N are *never* decreased, but may be augmented by addition of portions of excess probability.
- Often it is the case that there is some *distinguished state* z such that for each (i,a,r,N), $q_z(i,a,r,N)=1$

(That is, all excess probability is sent to the distinguished state.)

Infinite Horizon Case

For the discounted-cost MDP Δ with infinite horizon, and infinite state space *S*, let

$$V_b(i) = \min_{a\in A_i}\left\{C_i^a + b\sum_j P_{ij}^a V_b(j)\right\}, \quad \forall j\in S$$

Suppose we have an *approximating sequence* $\{\Delta_N\}$, with corresponding optimal values V_b^N

Major questions of interest:

- When does $\lim_{N\to\infty} V_b^N(i) = V_b(i) < +\infty$?
- If p^N is the optimal policy for Δ_N, when does p^N converge to an optimal policy for Δ?

Infinite Horizon Discounted Cost Assumption **DC(β)**:

For i∈S we have

$$\limsup_{N\to\infty} V_b^N(i) \equiv W_b(i) < +\infty$$

and

$$W_b(i) \le V_b(i)$$

Theorem (Sennott, page 76):

The following are equivalent:

- $\lim_{N\to\infty} V_b^N = V_b < +\infty$
- Assumption **DC(β)** holds.

If one (& therefore both) of these conditions are valid, and $\{\ {}_b^N\}$is an optimal stationary policy for $_N$. Then any limit point of the sequence is optimal for Δ.

The following theorem of Sennot (p. 77) gives a sufficient condition for **DC(β)** to hold (and hence for the convergence of the approximating sequence method):

Theorem:

Assume that there exists a finite constant B such that $C_i^a \le B$ for every $i \in S$ and $a \in A_i$. Then **DC(β)** is valid for $b \in (0,1)$

Example: Inventory Replenishment

Consider again our earlier application to inventory replenishment:

♦ The daily demand is random, with Poisson distribution having mean of 3 units.

♦ The inventory on the shelf (the *state*) is counted at the end of each business day, and a *decision* is then made to raise the inventory level to *S* at the beginning of the next business day.

♦ There is a fixed cost *A=10* of placing an order, a holding cost *h=1* for each item in inventory at the end of the day, and a penalty *p=5* for each unit backordered.

We imposed limits of 7 units of stock-on-hand and 3 backorders, and found that the policy which minimizes the expected cost/day is of type **(s,S)** = **(2, 6)**, i.e., if the inventory position is 2 or less, order enough to bring the inventory level up to 6.

Consider the problem with ***infinitely-many states***, i.e.,

$$S = \{-\infty, \ldots -2,\ -1, 0, 1, 2, 3, 4, \ldots +\infty\}$$

and the objective of minimizing the ***discounted cost***, with discount factor

$$b = \frac{1}{1+0.20} = 0.833333.$$

What is the optimal replenishment policy?

Approximating Sequence Method

N = 1

To define the first MDP in the sequence, Δ_1, use state space

$$S_1 = \{-2, -1, 0, 1, 2, \ldots 6\},$$

i.e., assume a limit of 2 backorders and 6 units in stock. The optimal policy is **(s, S) = (2, 6)**:

```
   State       |  Action  |       V
  BO= two      |  SOH= 6  | 72.3583
  BO= one      |  SOH= 6  | 57.3583
  SOH= zero    |  SOH= 6  | 52.3583
  SOH= one     |  SOH= 6  | 53.3583
  SOH= two     |  SOH= 2  | 52.4908
  SOH= three   |  SOH= 3  | 50.4510
  SOH= four    |  SOH= 4  | 49.2100
  SOH= five    |  SOH= 5  | 48.5763
  SOH= six     |  SOH= 6  | 48.3583
```

N = 2

We now increase the state space to

$$S_2 = \{-3, -2, -1, 0, 1, 2, \ldots 6, 7\},$$

i.e., assume a limit of 3 backorders and 7 units in stock, and find that the optimal policy is **(s, S) = (2, 7):**

```
  State       |  Action  |       V
 BO= three    |  SOH= 7  | 98.2503
 BO= two      |  SOH= 7  | 73.2503
 BO= one      |  SOH= 7  | 58.2503
 SOH= zero    |  SOH= 7  | 53.2503
 SOH= one     |  SOH= 7  | 54.2503
 SOH= two     |  SOH= 7  | 55.2503
 SOH= three   |  SOH= 3  | 53.2667
 SOH= four    |  SOH= 4  | 51.3011
 SOH= five    |  SOH= 5  | 50.4785
 SOH= six     |  SOH= 6  | 50.2025
 SOH= seven   |  SOH= 7  | 50.2503
```

N = 3

We now increase the state space to $S_3 = \{-4, -3, -2, -1, 0, 1, 2, \ldots 7, 8\}$,

i.e., assume a limit of 4 backorders and 8 units in stock, and find that the optimal policy is **(s, S) = (2, 8):**

```
  State       |  Action  |       V
 BO= four     |  SOH= 8  | 130.6728
 BO= three    |  SOH= 8  |  95.6728
 BO= two      |  SOH= 8  |  70.6728
 BO= one      |  SOH= 8  |  55.6728
 SOH= zero    |  SOH= 8  |  50.6728
 SOH= one     |  SOH= 8  |  51.6728
 SOH= two     |  SOH= 8  |  52.6728
 SOH= three   |  SOH= 3  |  51.8500
 SOH= four    |  SOH= 4  |  49.3778
 SOH= five    |  SOH= 5  |  48.4689
 SOH= six     |  SOH= 6  |  48.2269
 SOH= seven   |  SOH= 7  |  48.3086
 SOH= eight   |  SOH= 8  |  48.6728
```

N = 4

We now increase the state space to $S_4 = \{-5, \ldots, -1, 0, 1, 2, \ldots, 9, 10\}$,

and find that the optimal policy is **(s, S) = (2, 10):**

```
  State       |  Action   |       V
 BO= five     |  SOH= 10 | 176.7718
 BO= four     |  SOH= 10 | 131.7718
 BO= three    |  SOH= 10 |  96.7718
 BO= two      |  SOH= 10 |  71.7718
 BO= one      |  SOH= 10 |  56.7718
 SOH= zero    |  SOH= 10 |  51.7718
 SOH= one     |  SOH= 10 |  52.7718
 SOH= two     |  SOH= 10 |  53.7718
 SOH= three   |  SOH= 3  |  53.5004
 SOH= four    |  SOH= 4  |  50.7828
 SOH= five    |  SOH= 5  |  49.8438
 SOH= six     |  SOH= 6  |  49.6259
 SOH= seven   |  SOH= 7  |  49.7289
 SOH= eight   |  SOH= 8  |  50.1051
 SOH= nine    |  SOH= 9  |  50.7841
 SOH= ten     |  SOH= 10 |  51.7718
```

N = 5 Increase the state space to $S_5 = \{-6, \ldots , -1, 0, 1, 2, \ldots 11, 12\}$.

The optimal policy is again **(s, S) = (2, 10):**

```
     State        |  Action  |        V
     BO= six      |  SOH= 10 | 231.8900
     BO= five     |  SOH= 10 | 176.8900
     BO= four     |  SOH= 10 | 131.8900
     BO= three    |  SOH= 10 |  96.8900
     BO= two      |  SOH= 10 |  71.8900
     BO= one      |  SOH= 10 |  56.8900
     SOH= zero    |  SOH= 10 |  51.8900
     SOH= one     |  SOH= 10 |  52.8900
     SOH= two     |  SOH= 10 |  53.8900
     SOH= three   |  SOH= 3  |  53.7796
     SOH= four    |  SOH= 4  |  50.9538
     SOH= five    |  SOH= 5  |  49.9933
     SOH= six     |  SOH= 6  |  49.7723
     SOH= seven   |  SOH= 7  |  49.8706
     SOH= eight   |  SOH= 8  |  50.2390
     SOH= nine    |  SOH= 9  |  50.9098
     SOH= ten     |  SOH= 10 |  51.8900
     SOH= eleven  |  SOH= 11 |  53.1630
     SOH= twelve  |  SOH= 12 |  54.7082
```

N = 6 Increase the state space to $S_5 = \{-7, \ldots , -1, 0, 1, 2, \ldots 11, 15\}$.

The optimal policy is again **(s, S) = (2, 10):**

```
     State        |  Action  |     V
 BO= seven        |  SOH= 10 | 296.9292
 BO= six          |  SOH= 10 | 231.9292
 BO= five         |  SOH= 10 | 176.9292
 BO= four         |  SOH= 10 | 131.9292
 BO= three        |  SOH= 10 |  96.9292
 BO= two          |  SOH= 10 |  71.9292
 BO= one          |  SOH= 10 |  56.9292
 SOH= zero        |  SOH= 10 |  51.9292
 SOH= one         |  SOH= 10 |  52.9292
 SOH= two         |  SOH= 10 |  53.9292
 SOH= three       |  SOH= 3  |  53.8742
 SOH= four        |  SOH= 4  |  51.0097
 SOH= five        |  SOH= 5  |  50.0426
       :                :           :
 SOH= fourteen    |  SOH= 14 |  58.5790
 SOH= fifteen     |  SOH= 15 |  60.8442
```

The optimal policies have converged to **(s, S) = (2, 10)**

Finite Horizon Case

For the MDP Δ with finite horizon *n* and infinite state space *S*, let

$$v_{\beta,n}(i) = \min_{a\in A_i}\left\{C_i^a + \beta\sum_j P_{ij}^a v_{\beta,n-1}(j)\right\}, \quad \forall j\in S, n\geq 1$$

Suppose we have an *approximating sequence* $\{\Delta_N\}$, with corresponding optimal values $v_{\beta,n}^N$

Major questions of interest:

- When does $\lim_{N\to\infty} v_{\beta,n}^N(i) = v_{\beta,n}(i)$?
- If π^N is the optimal policy for Δ_N, when does π^N converge to an optimal policy for Δ?

Finite Horizon Assumption **FH(β,n)**:

For $i\in S$ we have

$$\limsup_{N\to\infty} v_{\beta,n}^N \equiv w_{\beta,n} < +\infty$$

and

$$w_{\beta,n}(i) \leq v_{\beta,n}(i)$$

Theorem (Sennott, page 43):

Let n≥1 be fixed. The following are equivalent:

- $\lim_{N\to\infty} v_{\beta,n}^N = v_{\beta,n} < +\infty$
- Assumption **FH(β,n)** holds.

The following theorem of Sennot (p. 45) gives a sufficient condition for **FH(β,n)** to hold (and hence for the convergence of the approximating sequence method):

Theorem:

Suppose that there exists a finite constant **B** such that

$$C_i^a \leq B$$

$$F_i \leq B$$

where F_i is the terminal cost of state $i\in S$. Then **FH(β,n)** holds for all β and n≥1.