Finite Horizon MDP and LP in Inventory Control

Prakash Gawas
153190008

IIT Bombay

October 26, 2016

Markov Decision Process (MDP)

Markov decision processes (MDPs) is a mathematical framework for modelling sequential decision making in problems. In these problems, a decision maker, at every decision epoch must see the current state of the system and take a decision/action. Depending on the state and action chosen he will receive an immediate reward or incur a cost. The state of the system than evolves to a possibly different state depending on the transition probability. This process continues over time and the decision maker receives a sequence of rewards/costs. The available actions, the rewards and the transition probabilities only depend on the current state and not on the past history. The aim of the decision maker is to maximise or minimize a function of such a sequence.

Applications

- Gambling
- Optimal Stopping
- Inventory Management
- Queuing Systems
- Routing Problems
- Communication Models

Single Product Stochastic Inventory Control

Consider you are a manager in charge of the inventory of a particular product in a firm. Your task is to decide at the start of every month, whether or not to order additional stock from supplier after seeing the current stocking hand. In doing so you will face a trade off between cost associated with keeping Inventory and the lost profit associated with not being able too fulfil demand of the customer. Your objective is find a policy to minimize some function of the costs incurred over the planning period. The demand is random and follows a known distribution.

Assumptions

- All orders are placed at the start of the period and received immediately and lag is zero.
- Demands arrive immediately after the orders arrive and are fulfilled instantly.
- Holding and Shortage costs are charged linearly.
- Unsatisfied demand is lost forever.
- The capacity of the storing facility is M.
- Demand $\xi_1, \xi_2 ..... \xi_N$ are independent and identically distributed by a common discrete distribution function $P(\xi)$.

Costs

- Order cost : $C_O(a) = \begin{cases} 0\ , & a = 0 \\ K + c_u a, & a > 0 \end{cases}$

  where K is fixed ordering cost, c_u is unit ordering cost and a is amount quantity ordered.

- Holding Cost : $C_H(x, a) = \begin{cases} 0, & \xi \geq x + a \\ c_H(x + a - \xi), & \xi < x + a \end{cases}$

  where c_k is the unit holding cost and x is current inventory level.

- Shortage Cost : $C_S(x, a) = \begin{cases} 0, & \xi \leq x + a \\ c_S(\xi - x - a), & \xi > x + a \end{cases}$

  where c_s is the unit shortage cost.

Terminology

- Decision-Making Horizon - Total planning period N.
- Decision Epoch (T) - Points in the planning horizon when decisions are made i.e months in planning period.

$$T = \{1, 2, 3, ...N\}, \qquad N \leq \infty.$$

- State Space ($\mathcal{X}$) - Denotes the set of possible Inventory level.

$$\mathcal{X} = \{0, 1, 2, ...M\}.$$

- Action A_x - The set of possible order quantity from current stock x.

$$A_x = \{0, 1, 2, ...M - x\}, \qquad \text{where } x \in \mathcal{X}.$$

Terminology

- Rewards ($R_t(x, a)$) - Cost incurred after you order quantity a when the inventory level is x.

$$R_t(x, a) = \big(C_O(a) + C_H(x, a) + C_S(x, a)\big) \tag{1}$$

- Transition Probability ($p(j|x, a)$) - Defines the probability to move to a possibly different state j after ordering some quantity a when inventory level is x. Depends on the demand distribution.

$$p(j|x, a) = \begin{cases} 0, & M \geq j > x + a \\ p_{x+a-j}, & M \geq x + a \geq j > 0 \\ \sum_j^{\infty} p_j, & M \geq x + a \text{ and } j = 0 \end{cases} \tag{2}$$

where $p_i = P(\xi = i)$.

Terminology

- Policy (π) - This represents a prescription of the amount of quantity to be ordered a at any state x and at any decision epoch t. Following a policy π means you will receive sequence of rewards. The sample policy below gives what action to be taken given any state x and decision epoch t.

Table: Example Policy

x\t	1	2	3
1	1	2	0
2	1	2	0
3	0	0	3

- Optimum Policy (π^*) - Policy which will minimize the some function of these sequence of rewards.

Aim

The aim is to find a policy π^* prior to the first decision epoch so that we minimize the expected value of a function of these rewards i.e :

$$\min_{\pi \in \Pi} E\left[f(R_1, R_2, ...)\right] \tag{3}$$

where Π is set of all possible policies π.

Risk Neutral MDP (RNMDP)

In risk neutral MDP (also called Linear Cost MDP) one minimizes the sum of the sequence of the rewards.

$$\min_{\pi \in \Pi} E\left[\sum_{t} R_t\right] \tag{4}$$

One can even consider a discount factor β to minimize sum of discounted values.

$$\min_{\pi \in \Pi} E\left[\sum_{t} \beta^t R_t\right] \tag{5}$$

One controls the first moment of sum cost in linear MDPs.

RNMDP - Simplification

Your objective is to minimize :

$$E[\sum_t r_t(x,a)] = \sum_t E[r_t(x,a)] \tag{6}$$

$$= \sum_t E\Big[E\Big[R_t|(X_1,A_1),(X_2,A_2),...,(X_{T-1},A_{T-1}),(X_T)\Big]\Big]. \tag{7}$$

Hence we can define value function:

$$v_t(x) = \sum_t E[r_t(x,a)] \tag{8}$$

Solution Approaches - Dynamic Program (DP)

DP involves finding the optimal expected cost using Backward Induction Algorithm for a finite horizon problem. Based on the Principle of Optimality, DP divides the problem into many sub problems and than optimizes these problems. The corresponding equations given in [1] are:

$$v_t(x) = \min_{a \in A_x} \left\{ r_t(x,a) + \sum_{j \in \mathcal{X}} p(j|x,a) v_{t+1}(j) \right\} \tag{9}$$

for $t = 1, ..., N-1$ and $x \in \mathcal{X}$

$$v_N(x) = r_N(x) \tag{10}$$

for $x \in \mathcal{X}$
where v_t is the accumulated reward from time t and onwards.
Solution to these equations provide the optimal policy for period t onwards.

Solution Approaches - Linear Program (LP) - Primal

Using the optimality equations, one can derive what are called fixed point equations which are then used to derive LPs to solve the MDP. The following Primal LP is given in [2] to solve finite horizon problem.

$$\max \sum_{x \in \mathcal{X}} \alpha(x) v_0(s) \tag{11}$$

subject to:

$$v_t(x) \leq r_t(x, a) + \sum_{x' \in \mathcal{X}} p(x'|x, a) v_{t+1}(x') \tag{12}$$

for $t = 0, ..., T-2$.

$$v_{t-1}(x) \leq r_{T-1}(x, a) + \sum_{x' \in \mathcal{X}} p(x'|x, a) r_T(x') \tag{13}$$

$\forall x \in \mathcal{X}$ and $a \in A_x$ and α is the initial distribution on X.

Solution Approaches - Linear Program (LP) - Dual

The Primal LP gives the optimal sum of cost, while the dual LP gives the optimal policy to the problem.

$$\min \quad \sum_{t=1}^{T-1} \sum_{a} \sum_{x} r_t(x,a) y(t,x,a) + \sum_{a} \sum_{x,x'} r_T(x') y(T-1,x,a) p(x'|x,a) \tag{14}$$

subject to:

$$\sum_{a} y(0,x',a) = \alpha(x') \ \ \forall x' \in \mathcal{X} \tag{15}$$

$$\sum_{a} y(t,x',a) = \sum_{a} \sum_{s} p(s'|x,a) y(t-1,x,a) \ \ \forall 1 \leq t \leq T-1 \text{ and } x' \in \mathcal{X} \tag{16}$$

and $y(t,x,a) \geq 0$ for $a \in A_x$ and $x \in \mathcal{X}$.
$y(t,x,a)$ represents the total joint probability under α of being in state x and choosing action a at time t.

Inventory Control

In [3], Herbert Scarf considers a similar dynamic inventory problem. He has shown under the assumptions of linear costs (holding and shortage) that the optimal is of the type (s, S). This means that one would order only when the stock in hand is less than s than you would order enough quantity so that the stock rises to S. He has shown that the total cost function follows K-convexity and hence the optimal policy turns out to have a simple form.

$$C_n(x) = \min_a (K + c_u a)\mathbf{1}_{\{a>0\}} + L(x+a) + \sum_0^\infty C_{n-1}(x+a-\xi)P(\xi) \quad (17)$$

where

$$L(x+a) = \sum_{\xi=0}^{x+a-1} c_H(x+a-\xi)P(\xi) + \sum_{\xi=x+a}^{\infty} c_S(\xi-x-a)P(\xi) \quad (18)$$

Hence the reward

$$r_t(x, a) = (K + c_u a)\mathbf{1}_{\{a>0\}} + L(x + a). \tag{19}$$

and

$$r_T(x) = 0 \tag{20}$$

Solution - DP

Consider Uniform distribution of demand between $(0, 20)$. Let $M = 20, N = 10, K = 0.5, c_u = 0.4, c_h = 0.2, c_s = 0.8$. The solution obtained through DP is :

Table: Optimal Solution for Uniform Distribution between $(0, 20)$ using DP

	Decision Epoch										
States	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	14	14	14	14	14	14	14	14	13	8	0
1	13	13	13	13	13	13	13	13	12	7	0
2	12	12	12	12	12	12	12	12	11	6	0
3	11	11	11	11	11	11	11	11	10	5	0
4	10	10	10	10	10	10	10	10	9	0	0
5	9	9	9	9	9	9	9	9	8	0	0
6	8	8	8	8	8	8	8	8	7	0	0
7	7	7	7	7	7	7	7	7	6	0	0

	Decision Epoch										
States	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
8	6	6	6	6	6	6	6	6	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0

Optimal policy

Table: Optimal (S, s) policy for Uniform Distribution with parameters $(0, 20)$ using DP

	Decision Epoch										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
S	14	14	14	14	14	14	14	14	13	8	0
s	9	9	9	9	9	9	9	9	8	4	0

Solution - LP

The problem was solved for the same settings and solution obtained were same as DP.

Table: Optimal (S, s) policy for Uniform Distribution with parameters $(0, 20)$ using LP

	Decision Epoch										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
S	14	14	14	14	14	14	14	14	13	8	0
s	9	9	9	9	9	9	9	9	8	4	0

Optimal Policy with $K = 0$

The problem was solved again with the same settings but with $K = 0$ and the optimal policy obtaines was (s, S) with $s = S$.

Table: Optimal (S, s) policy for Uniform Distribution with parameters $(0, 20)$ with $K = 0$

	Decision Epoch										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
S	13	13	13	13	13	13	13	13	12	8	0
s	13	13	13	13	13	13	13	13	12	8	0

Rewards - Alternate Model

You can consider shift of few cost terms from current time slot to the next, which would not change the actual problem and which would pave way for risk MDP. It is easily known that if the state s at the start of next period is greater than zero than we would directly incur holding cost equivalent to $c_h s$ in the current period. If $s = 0$ than we for sure know that there was no holding cost and only shortage cost would apply. Therefore we can model the cost r_t as function of the next state of inventory which would give us information on which of the cost is applicable in the previous period.

Defining new Rewards

$$r_t(x_t, a_t) = \begin{cases} c_o a_t & , t = 1 \\ c_o a + c_h(x_t)\mathbf{1}_{\{x_t>0\}} + c_s(\tilde{\xi})\mathbf{1}_{\{x_t=0\}} & , 1 < t < T \\ c_h(x_t)\mathbf{1}_{\{x_t>0\}} + c_s(\tilde{\xi}) & , t = T \end{cases} \quad (21)$$

where $\tilde{\xi}$ is the excess variable demand and distribution of whose is to be known to compute the reward. This excess demand $\tilde{\xi}$ depends on the inventory state, the action taken in that state and demand in that period. Advantages of this formulation:

1. The holding cost does not depend on random demand.
2. Shortage cost depends only upon excess demand $\tilde{\xi}$, which would than be independent of further state evolution. Further if the demands are memoryless, than $\tilde{\xi}$ has the same distribution as ξ.

Solution - Alternate Model

The demand distribution is geometric with parameter $p = 0.3$. Let $M = 20, N = 10, K = 0.2, c_u = 0.4, c_h = 0.1, c_s = 1$.

Table: Optimal (s, S) policy for Geometric distribution

	Decision Epoch										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
S	6	6	6	6	6	6	5	5	3	2	0
s	2	2	2	2	2	2	2	2	1	0	0

If you solve the problem for the same settings using the rewards as defined in (19) and (20), the same solution will be obtained.

Risk Sensitive MDP (RSMDP)

As a manager of the inventory, you would like to control the fluctuations around the expected value. Here you want to be more sure about your total cost. Hence one can use RSMDPs to model such cases. RSMDPs can have varied applications in diverse fields. From investment of funds to manufacturing processes one can use RSMDPs to find optimal decisions. RSMDPs basically mean giving varied importance to higher moments of the total expected cost.

Mean Variance MDP (MVMDP)

One way is to optimize a function of mean and variance only. In [4], authors consider mean variance optimization.

$$E\Big[\sum_t R_t\Big] + \theta Var\Big(\sum_t R_t\Big) \tag{22}$$

where θ is some control parameter. But there are many problems associated with MVMDP

- Absence of principle of optimality which could have led to simple recursive algorithms
- Variance $(Var(W)) = E(W^2) - (E(W))^2$ is not a linear function of the probability measure of the process where W is sum of the rewards for the time horizon.
- These MDPs are typically NP hard but may admit a pseudo polynomial time solution.

Exponential form of RSMDP

In [2], authors give importance to higher moments of the sum cost by considering the following objective function:

$$E[e^{\gamma \sum_{t \leq T} R_t}] \tag{23}$$

where γ is risk parameter, which provides importance to higher moments. This form of objective function covers some shortcomings of the MVMDPs.
Note that

$$E[e^{\gamma \sum_{t \leq T} R_t}] = E[\sum_k \gamma^k (\sum_{t \leq T} R_t)^k / k!]. \tag{24}$$

Exponential form RSMDP - Simplification

Your objective is to minimize:

$$E[e^{\sum_{t=1}^{T}\gamma R_t}] = E[\Pi_{t=1}^{T} e^{\gamma R_t}] \tag{25}$$
$$= E\left[E\left[\Pi_{t=1}^{T} e^{\gamma R_t} | (X_1, X_2, ...X_T)(A_1, A_1, ...A_{T-1})\right]\right]$$

As compared to linear cost MDPs we cannot simplify this easily. In linear MDPs (7), we can simplify the total in terms of sum of expectations of the costs which is easy to compute. But this is not the case in Risk sensitive MDPs as seen above and we cannot average out the risk reward for a given state s and the corresponding action a.

Exponential form of RSMDP - Approximation

The authors in [2], consider the following transformation of the RNMDP objective function.

$$\tilde{J}_o(\alpha, \pi) = \frac{1}{\gamma} \log \left(J_o(\alpha, \pi) \right) \tag{26}$$

where

$$J_o(\alpha, \pi) = E^{(\alpha,\pi)}[e^{\gamma \sum_t r_t(X_t, A_t)}] \tag{27}$$

For smaller values of γ the objective value takes the form:

$$\tilde{J}_o(\alpha, \pi) \simeq E^{(\alpha,\pi)}\left[\sum_t r_t(X_t, A_t)\right] + \frac{\gamma}{2} Var^{(\alpha,\pi)}\left[\sum_t r_t(X_t, A_t)\right] \tag{28}$$

As one can see as $\gamma \to 0$ the objective function approaches risk neutral cost.

Solution approaches - DP

The value function is defined as the optimal value of the above risk sensitive objective with initial condition $X_0 \equiv x$:

$$v^*(x) := \min_{\pi \in \mathcal{D}} \tilde{J}_o(\alpha, \pi) \qquad \text{for any } x \in \mathcal{X} \tag{29}$$

The authors have given a Dynamic Programming equations using backward induction as below for any $x \in \mathcal{X}$.

$$v_T(x) = r_T(x), \text{ and for any } 0 \leq t \leq T-1, \tag{30}$$

$$v_t(x) = \min_{a \in \mathcal{A}} \left\{ r_t(x,a) + \frac{1}{\gamma} log \Big[\sum_{x' \in \mathcal{X}} p(x'|x,a) e^{\gamma v_{t+1}(x')} \Big] \right\}, \tag{31}$$

and their solution provides the value function $v^*(x) = v_0^*(x)$.

Solution approaches - DP - Simplification

The authors, to simplify the DP also provide the following translation of the value function:

$$u_t(x) = e^{\gamma v_t(x)} \text{ for all } 0 \leq t \leq T, \text{ and } x \in \mathcal{X} \tag{32}$$

By monotonicity and continuity $u_0^*(x) = e^{\gamma v_0^*(x)}$ is the minimum value of the risk cost J_0 in (25):

$$u_0^*(x) = \min_{\pi} J_0(x, \pi). \tag{33}$$

Hence the DP equations are now rewritten as:

$$u_T(x) = e^{\gamma v_T(x)}, \text{ for any } x \in \mathcal{X} \tag{34}$$

$$u_t(x) = \min_{a \in \mathcal{A}} \left\{ e^{\gamma v_t(x)} \sum_{x' \in \mathcal{X}} p(x'|x, a) u_{t+1}(x') \right\}, \tag{35}$$

for any $0 \leq t \leq T - 1$, and $x \in \mathcal{X}$.

Solution approaches - LP (Primal)

The authors give the following Primal LP to find the optimal value:

$$\max \sum_{x \in \mathcal{X}} \alpha(x) u_0(x) \tag{36}$$

subject to

$$u_{T-1}(x) \leq b_{x,a} \tag{37}$$

for all x, a,

$$u_t(x) \leq e^{\gamma v_t(x)} \sum_{x' \in \mathcal{X}} p(x'|x,a) u_{t+1}(x') \tag{38}$$

$\forall\ x \in \mathcal{X}$, $a \in A_x$ and $t \leq T-2$,
with $b_{(x,a)} := e^{\gamma r_{T-1}(x,a)} \sum_{x'} p(x'|x,a) e^{r_T(x')}$.

Solution approaches - LP (Dual)

The authors give the following Dual LP to find the optimal policies:

$$\min \quad \sum_{a} \sum_{x} b_{(x,a)} y(T-1, x, a) \tag{39}$$

subject to

$$\sum_{a} y(0, x', a) = \alpha(x') \text{ for all } x' \in \mathcal{X} \tag{40}$$

$$\sum_{a} y(t, x', a) = \sum_{a} \sum_{x} e^{\gamma r_{t-1}(x,a)} p(x'|x, a) y(t-1, x, a) \tag{41}$$

$\forall\ 1 \leq t \leq T-1$ and $x' \in \mathcal{X}$ with
$b_{(x,a)} := e^{\gamma r_{T-1}(x,a)} \sum_{x'} p(x'|x, a) e^{r_T(x')}$ and $y(t, x, a) \geq 0$

RSMDP and Inventory Control - Rewards

You consider the transformed risk reward as below:

$$= E\left[E\left[\Pi_{t=1}^{T} e^{\gamma R_t} | (X_1, A_1), (X_2, A_2), ..., (X_{T-1}, A_{T-1}), (X_T)\right]\right], \quad (42)$$

substitute the shifted rewards from (21) and simplifying,

$$= E\left[\Pi_{t=1}^{T-1} e^{c_o A_t} \Pi_{t=1}^{T} e^{(c_h X_t)\mathbf{1}_{\{x_t>0\}}} M\right] \quad (43)$$

where

$$M = E[\Pi_{t=1}^{T-1} e^{(c_s \tilde{\xi})\mathbf{1}_{\{X_t=0\}}} | (X_1, ...X_T)(A_1, ...A_{T-1})] \quad (44)$$

The distribution of excess demand $\tilde{\xi}$ in general depends upon the inventory state and action of previous slot, however will not influence further evolution of the system. We can assume memory less demands (geometric), in which case $\tilde{\xi}$ is again distributed geometrically and one can average this out to compute the shortage cost.

RSMDP - Transformed Reward

$$M = E[\Pi_{t=2}^{T} e^{(c_s \tilde{\xi}) \mathbf{1}_{\{x_t=0\}}}] \tag{45}$$

$$= \Pi_{t=2}^{T} E[e^{(c_s \tilde{\xi}) \mathbf{1}_{\{x_t=0\}}}] \tag{46}$$

The new transformed rewards:

$$r_t(x_t, a_t) = \begin{cases} c_o a_t & , t = 1 \\ c_o a + c_h(x_t)\mathbf{1}_{\{x_t>0\}} + \tilde{c}\mathbf{1}_{\{x_t=0\}} & , 1 < t < T \\ c_h(x_t)\mathbf{1}_{\{x_t>0\}} + \tilde{c}\mathbf{1}_{\{x_t=0\}} & , t = T \end{cases} \tag{47}$$

where $\tilde{c} = log\left(E[e^{(\gamma c_s \tilde{\xi})}]\right)$.

Convergence of RSMDP to RNMDP as $\gamma \to 0$

In [2] , authors have shown the convergence of Primal RSMDP LP to Primal RNMDP LP as $\gamma \to 0$. However there was discontinuity observed in the Dual convergence of RSMDP.
The convergence in case of Primal RSMDP LP has been proved to be true when we consider the transformed reward as in (44). This was an important step because of the presence of the constant in (44).

RSMDP - Solution $\gamma \to 0$

The demand distribution is geometric with parameter $p = 0.3$. Let $M = 20, N = 10, K = 0.2, c_u = 0.4, c_h = 0.1, c_s = 1$. Also consider $\gamma = 0.001$. The optimal policy obtaines is:

Table: Optimal (s, S) policy for Geometric distribution

	Decision Epoch										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
S	6	6	6	6	6	6	5	5	3	2	0
s	2	2	2	2	2	2	2	2	1	0	0

As can be seen from the table the optimal policy matches with the RNMPD optimal policy from Table 6.

RSMDP - Solution $\gamma = 0.2$

The demand distribution is geometric with parameter $p = 0.3$. Let $M = 20, N = 10, K = 0.2, c_u = 0.4, c_h = 0.1, c_s = 1$. The optimal policy obtained is:

Table: Optimal (s, S) Policy with $\gamma = 0.2$

	Decision Epoch										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
S	9	9	9	9	9	9	8	7	6	3	0
s	5	5	5	5	5	5	5	4	3	1	0

As can be seen from the table the optimal policies remain (s, S). This change from the linear cost solution is expected as the shortage cost is more than the holding cost. Since you are averse towards taking risk, you will try to protect against under stocking.

Comparison of RSMDP and RNMDP optimal policies

Comparison of the optimal policies were made between RSMDP and RNMDP by sequentially varying one of the cost and graphs were plotted. The value of γ used for RSMDP is 0.2.

Holding Cost

Let $K = 0.2, c_u = 0.2, c_s = 1$.

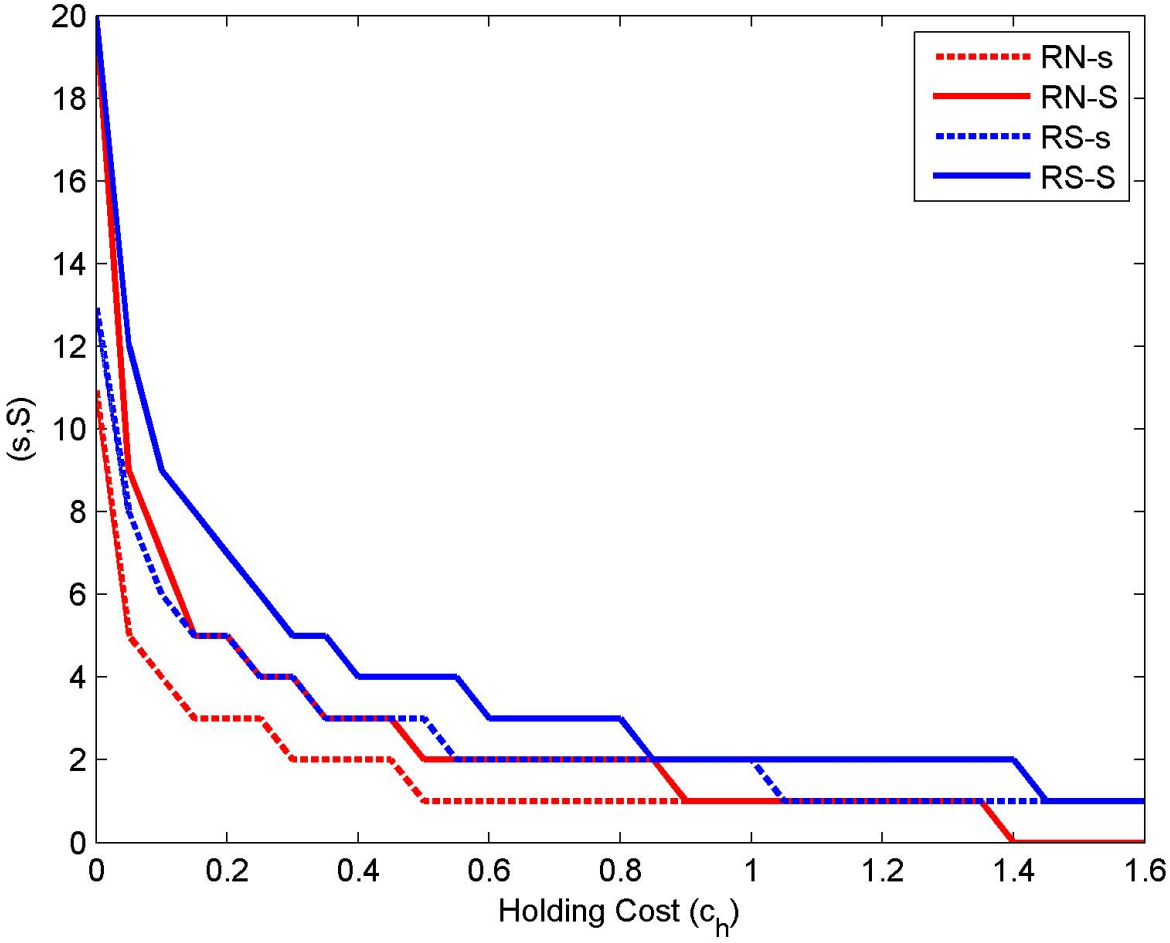

Figure: Risk Sensitive (RS, $\gamma = 0.2$) versus Risk Neutral (RN) for varying holding costs

Shortage Cost

Let $K = 0.1, c_u = 0.05, c_h = 0.5$.

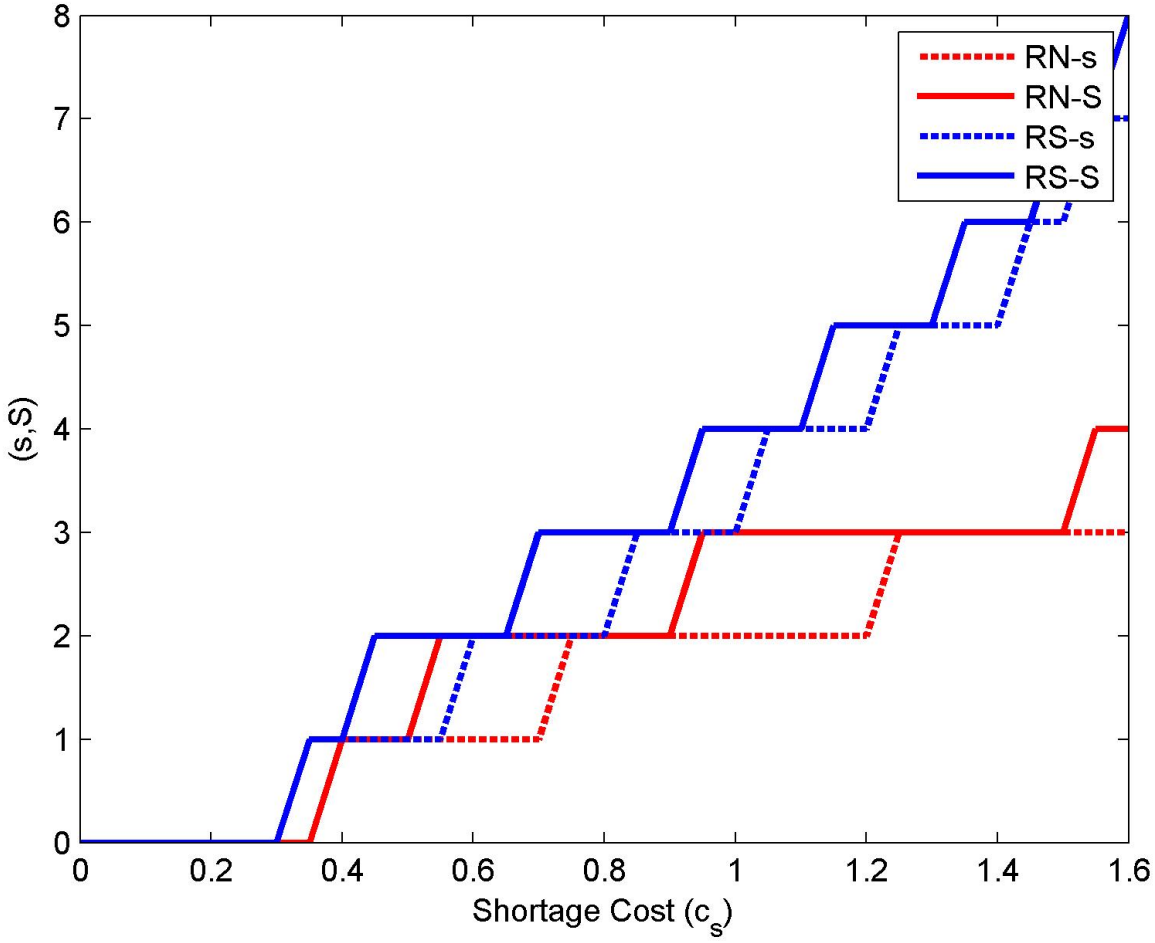

Figure: Risk Sensitive (RS, $\gamma = 0.2$) versus Risk Neutral (RN) for varying shortage costs

Unit Order Cost

Let $K = 0.05, c_h = 0.05, c_s = 1$.

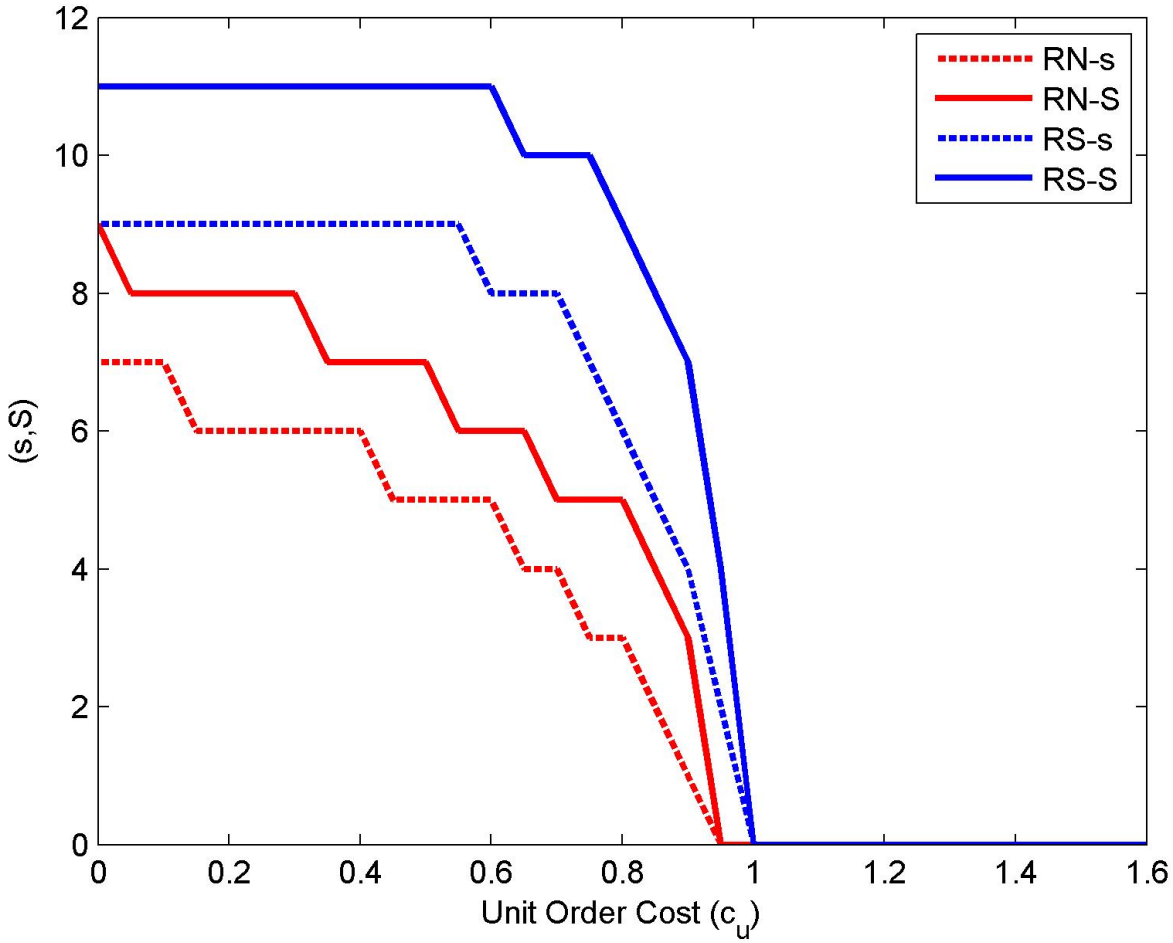

Figure: Risk Sensitive (RS, $\gamma = 0.2$) versus Risk Neutral (RN) for varying unit ordering costs

Fixed Ordering Cost

Let $c_u = 0.05, c_h = 0.05, c_s = 1$.

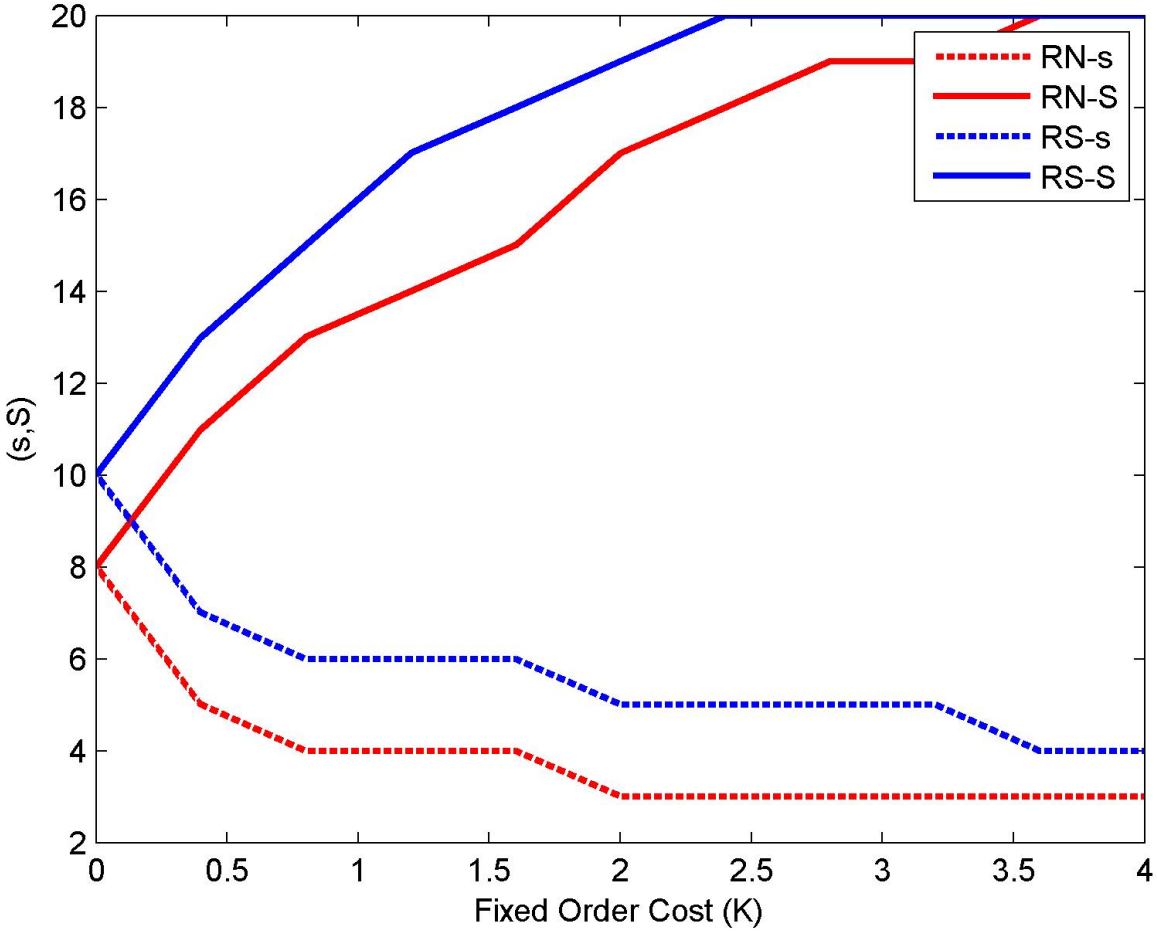

Figure: Risk Sensitive (RS, $\gamma = 0.2$) versus Risk Neutral (RN) for varying fixed ordering costs

Observations

- In figure 1, when the holding cost increases the values of s, S decrease, which mean one would store less for higher holding costs. Comparing risk neutral and risk sensitive cases, results are against what one expects, as $c_h > c_s$ the RSMDP model should protect against holding of products as the decision maker would be taking risk storing extra inventory.
- In figure 2, the RSMDP model starts ordering even when the shortage cost is less than the ordering cost. While for larger shortage cost the s, S values shoot up greatly as expected. Again one can see from the graph that risk sensitive model protects against under stocking more as compared to linear cost model even when the shortage cost is lesser than holding costs.

Observations

- In figure 3, curves are as expected i.e. when $c_u < c_s$ one has to order to reduce the costs but when $c_u > c_s$ there is no longer any incentive to order for both models.
- In figure 4, for models you see that at $K = 0$, $s = S$. As the value increases (s, S) values move in the opposite direction as expected.

Future Work

Alternate Model: You can consider an alternate model where instead of modelling shortages one can use profits from sales. Hence the new reward in this case is:

$$R_t(x,a) = \big(C_O(a) + C_H(x,a) - C_p(x,a)\big) \tag{48}$$

where

$$\text{Selling Reward : } C_P(x,a) = \begin{cases} c_p(x+a-\xi), & \xi \leq x+a \\ c_p(x+a), & \xi > x+a \end{cases}$$

where c_p is the unit selling cost. You can compare the shortage cost and selling cost models and find the difference in the policies.

Future Work

Other ideas to make this models consistent with real life situations are:

- Demand distribution other than geometric.
- Continuous distribution to model demand.
- Introduction of delivery lag.
- Perishability of products - [5] has extended work by [3] to introduce perishability cost in the model.
- Addition of constraints - [2] has also given LPs to solve constrained RSMDPs.

This model can be extended to infinite horizon problems. Algorithms like Value and Policy Iterations are available for RNMDPs. Also LPs are well studied for such cases. However LPs for Infinite Horizon RSMDPs are not formulated.

References I

[1] M. L. Puterman, "Market decision process," 1990.

[2] A. Kumar, V. Kavitha, and N. Hemachandra, "Finite horizon risk sensitive mdp and linear programming," 2015.

[3] H. Scarf, "The optimality of (s, s) policies in the dynamic inventory problem," 1959.

[4] A. Gosavi, "A risk-sensitive approach to total productive maintenance," 2006.

[5] S. Nahmias and W. P. Pierskalla, "Optimal ordering policies for a product that perishes in two periods subject to stochastic demand," 1973.

[6] R. A. Howard and J. E. Matheson, "Risk-sensitive markov decision processes," 1972.

References II

[7] Y. Liu and S. Koenig, "Risk-sensitive planning with one-switch utility functions: Value iteration," 2005.

[8] J. A. Filar, L. C. Kallenberg, and H.-M. Lee, "Variance-penalized markov decision processes," 1989.

[9] S. Mannor and J. Tsitsiklis, "Mean-variance optimization in markov decision processes," 2011.