A Report

on

”Finite Horizon Markov Decision Programs and Linear Programs”

Submitted in partial fulfillment of the requirements

of the degree of

Master of Technology

by

(Gawas Prakash Arjun)
(Roll no. 153190008)

Supervised by

(Prof. Veeraruna Kavitha)
(Prof. Ashutosh Mahajan)

Inter-disciplinary program
in
Industrial Engineering and Operations Reserach (IEOR)
IIT BOMBAY
(2016-17)

Declaration

I declare that this written submission represents my idea in my own words and where others's idea or words have been included ,I have adequately cited and referenced the original source.I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any/data/fact/source in my submission.I understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: Prakash Arjun Gawas

(153190002)

Acknowledment

I would like to extend thanks to the many people who so generously contributed to the work presented in the thesis.Special mention goes to my enthusiastic supervisor, **Prof. Veeraruna Kavitha** for giving me opportunity of working under his guidance.His Direction, motivation,affectionate guidance and support have been the source of inspiration to bring the report this shape.I thank all the other faculty member of Industrial Engineering And Operations Research, who made me realize the virtue of learning through sustained hard work.

I would like to thanks all those whose name i missed but have contributed in any form for building up of the thesis upto now.

Prakash Arjun Gawas
(153190002)

Abstract

The project study is based on the

Contents

List of Figures

List of Tables

Chapter 1

Preamble

1.1 Introduction

Markov Decision process (MDP) is a mathematical framework that solves the problem of sequential decision making under uncertainty, to obtain an optimal policy. At each decision epoch the system state provides the decision maker with all necessary information for choosing an action from available actions in that state. As result of an action the decision maker receives an immediate reward and the system state evolves to a possibly different state at the next epoch depending on the controlled (action based) transition probability. Each action, the decision maker chooses in any particular state and at any time slot t, will have an associated reward/running cost (R_t). The reward depends also on the current system state. In case of finite horizon problems, it also considers a terminal cost that only depends upon the state at termination. Hence over time the decision maker will receive a sequence of reward/costs.

A policy is a sequence of such actions/rules, which the decision maker chooses at any time epoch depending on the state of the system. The aim of MDP is to find a policy prior to the first decision epoch that maximizes the expected value of a function f of the reward sequences. In general MDP optimizes the expected value of the function f that defines a way of aggregating the running costs related to all the time slots under consideration, i.e., to optimize $E[f(R_1, R_2, ...)]$. Linear MDPs consider expected value of either sum of all the running costs $E[\sum_t R_t]$, or sum of discounted values of all the running costs $E[\sum_t \beta^t R_t]$ with discount factor $\beta < 1$, or time average of the running costs $\lim_{T\to\infty} [E[\sum_{t\leq T} R_t]/T$ .They are respectively called total cost, discounted cost or average cost problems. Also MDP whether discounted or total cost, can either be modelled over a finite time horizon or over the infinite time horizon.

Linear MDPs, also called risk neutral MDPs, control the first moment (expected value) of the sum cost. Optimization of the expected cost of an MDP is well studied. Many approaches have been developed to solve the problem. Dynamic programming (DP) is very popular technique to find optimal policies for finite horizon problems. Dynamic programming involves finding the optimal expected cost using Backward Induction. It basically solves a big problem by dividing it into many sub problems and than optimizing these sub problems. The corresponding equations are called optimality equations [7]. This gives rise to what is known as the Principle of Optimality where no matter what the current state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. Others solution methods include Value Iterations, Policy iteration and their variants. These are iterative algorithms based on fixed point equations derived from the Optimality equations and the theory of contraction mappings on normed linear spaces. The relationship between discounted MDPs and Linear programs (LPs) is also pretty well defined. Though not efficient, LPs have elegant theory: ease in addition of additional constraints and facility of sensitivity analysis make it very attractable in

some cases. LPs are derived based on properties of the solutions to the optimality equations (see Theoram 6.2.2, [7]). Primal linear program finds the optimal value while its dual provides the optimal policy.

In some situations we may also want to control the fluctuations around the expected value, which calls for Risk Sensitive MDPs (RSMDPs). 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. One can also just optimize a function of mean and variance only as in [6]. The literature on RSMDPs is relatively limited though. In [3], Howard and Matheson have considered the maximisation of total reward with a constant risk sensitive parameter. They have discussed the application of iterative algorithms like Value Iteration and Policy Iteration to risk sensitive case. Recently in [[4]] ,authors have proposed an LP based approach to solve the RSMDPs for finite horizon problem. Also they have augmented the state space appropriately so as to make addition of more constrains simple to solve the constrained RSMDP. The relation between LPs and infinite horizon is even more complex and needs to be investigated.

MDPs can be applied to varied examples in real life situations. One such application is the classic inventory control problem. In INVENTORY CONTROL problem the decision maker must decide how much to order in each time period to meet demand for its products. The demand is considered to be stochastic, whose distribution function is known. The objective here is to find a policy of ordering at any time epoch so as to minimize the total cost incurred for total period of time T thus making it a finite horizon problem. The various cost incurred depends on the current state of inventory, amount ordered and demand in that time period. There are three basic costs considered in this situation. Ordering cost that includes total ordering cost of the quantity and also fixed order set-up cost. Holding cost for storing inventory and carrying it into the next period. Shortage cost is applicable when we cannot fulfil the demand on hand. All the costs are considered to be linear. Hence given the state and the amount to be ordered in the current period the next inventory level will only depend on the demand in the current state and will transition according to the demand distribution. First we model it into a sequential decision model defining appropriate state space, actions, transition probability and rewards at each time epoch. If we solve this problem using one of the techniques mentioned previously, the policy that the decision maker has to follow to optimize the expected reward only is of the type (s,S).In [8], Herbert Scarf has shown that when holding cost and shortage cost are linear then the optimal policy is of this type. This means that whenever the inventory level goes below the level s will order so as to increase are inventory level to S, if not than we will not order. The aim in my work is to find optimal policies for a risk averse decision maker. Authors in [4] has given LP model for risk sensitive MDPS which can be applied for this example.

In Chapter III the model for inventory control is defined with the used notations. The rewards pertaining to the linear MDP have been identified. There were two ways in which we could model the rewards for this problem. One way was that given the current state of the inventory and the action we can find the average reward at that point. Other was to shift the holding and shortage cost from the current period to the next period. Both the cases were solved using LP formulation given in [4] for optimising the linear expected cost and also by using dynamic program from [7] and the results were validated. Furthermore in Chapter V the model for risk sensitive MDPs is presented.

Chapter 2

Literature Survey

Classical MDPs deal with the maximisation of cumulative reward, however a decision maker may focus on several other distributional properties. Most decision makers may be risk averse i.e. they would like to reduce their risk and thus this forms the basis of a risk sensitive model. One way is to focus on mean and variance of the cumulative reward. Mean-variance optimisation has varied applications from finance to manufacturing fields.

2.1 Mean Variance Optimization

In [6], authors consider mean variance optimization, that is they consider optimizing

$$E[\sum_t R_t] - \theta Var(\sum_t R_t)^{**} \tag{2.1}$$

With the introduction of variance in the cost function, we can optimize the fluctuations around the mean. They also deal with various computational complexities of mean-variance optimisation. There many problems associated with Mean-Variance MDPs.

1. The main problem with mean-variance optimisation is the absence of principle of optimality which could have led to simple recursive algorithms. For example consider a situation where the decision maker has received unexpected high rewards in the initial stages and has got to a state. He than may incur losses to keep the variance small. In case the user is at the same state and at the same stage, however received low rewards before this stage, then it would receive higher rewards in the later stages. Thus the principle of optimality fails.

2. Also 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.

3. These MDPs are typically NP hard but may admit a pseudo polynomial time solution.

They have defined their MDP model and compared different types of policies classes. Given the MDP model M (defined by the time horizon, state space, set of actions, set of rewards, transition probability) and rational numbers λ, ν the question they are addressing is whether there exist a policy in the set $\prod$ such that $J = E(W) \geq \lambda$ and $V = Var(W) \leq \nu$. They and established NP-hardness of this problem for all type of policies. They have addressed this using constrained Markov Decision Process to develop pseudo polynomial exact or approximate algorithms. To develop exact algorithms they augment the state space , to include the a term that contains the sum reward until the given stage. Considering bounded integer rewards, they have identified cases for the existence of exact pseudo polynomial time algorithms. LP approach cannot be used directly used for the problem due to convexity issues in Mean Variance MDPs. Hence they have concentrated on the pair J and $Q = E[W^2]$ and for integer rewards established the existence of pseudo

polynomial solution. In developing approximate algorithms the authors precisely consider the following:

- Minimize V while J $\geq \lambda$
- Maximize J while V $\leq \nu$

On similar lines as previously they consider integer rewards the authors have proved existence of a polynomial time algorithm. For any general reward case we can discretize the rewards and obtain new MDP which will be equivalent to one with integer rewards and the same algorithm can be applied.

In [1], Filar et al has given a Quadratic Program (QP) to directly solve Mean-Variance MDP ($_{**}$). The constraints of the QP are similar to the usual LP constraints, but the objective is quadratic in nature [1, equations(2.9)]. This is convex constrained QP. In [2],the author again considers the Mean-Variance problem($_{*}*$) to design plans for total productive maintenance (TPM). TPM focuses on improving the overall effectiveness of a manufacturing facility by eliminating the waste of time and resources. Their aim was to find optimal productive maintenance (PM) time. PM help reduce the frequency of unexpected repairs when failure rate is of increasing nature. Traditional approaches in TPM only used expected value of the long run cost and overlooked the risk associated with the high cost. In actual practice risk sensitive managers would use the expected value of long-run cost and then use a factor of safety for the same optimal time. This is a heuristic approach and is more conservative. They give mean variance model where the objective function is $g(\tau) = \mu_C + \theta\sigma^2$ with $\theta > 0$, where μ_C and σ^2 denote the long-run mean and the long-run variance, respectively, of the net cost per unit time incurred from following a preventive maintenance plan that prescribes τ as the time for PM.An alternative formulation in terms of rewards, in which the objective function is maximized,is $g(\tau) = \mu_R - \theta\sigma^2$ with $\theta > 0$, where μ_R and σ^2 denote the long-run mean and variance of the net reward per unit time, respectively. Risk neutral models have $\theta = 0$. Typically θ is selected experimentally. TPM plans for the production line in its entirety tend to be distinct from those for individual units that operate independently of the line. Hence they have developed separate models for the individual-unit scenario and the production-line scenario. For the case of the individual unit, a renewal-theory model is used. He considers a sub-optimal problem by considering the optimal among the age replacement policies. A typical age replacement policy replaces the item when its age reaches τ or when the failure occurs. Therefore the MDP crumble downs to a single variable optimization problem. Here every failure or a maintenance triggers a so-called renewal event. Using Renewal Reward Theory he has derived an expression for $g(\tau)$ in terms of τ, failure distribution and the individual cost of repair and maintenance. For the case of the production line a more involved model based on MDPs is used. He first develops the MDP identifies a Quadratic Program given in [] to solve the MDP. But solving the QP this is computationally expensive. He present an approach based on linearising the quadratic objective function by using a surrogate form for variance. Optimization could then be performed via linear programming. The existence of a deterministic policy is then proved for the surrogate objective function using DP approach to solve the problem by deriving the necessary optimality conditions. Using this he proposes a PI algorithm and shows its convergence. Computational results showed that the surrogate function did mimic the exact function reasonably well and also proves convergence in relatively short period of time.

2.2 Variants of Risk Sensitive MDPs

Mean Variance MDP reduces the variance trying to optimize the mean cost. However several other authors have considered many alternative ways of optimizing the fluctuations around the mean. In

[5], authors talk about risk sensitive planning with one switch utility functions. One switch utility functions model a decision maker whose decision change with their wealth level. For example a decision maker with low wealth levels may be risk averse, but may become risk neutral as the wealth level increases. These function model the risk attitude of decision maker.A decision maker is risk-neutral if their utility function is linear, risk-averse if their utility function is concave, and risk-seeking if their utility function is convex. Here one need to maximise the expected utility of a MDP for a given one switch utility function. This is difficult since the resulting planning problem is not decomposable. They model a probabilistic planning problem as a fully observable Goal-Directed Markov Decision Problems (GDMDPs) and investigate how to maximize the expected utility. The optimal course of action now depends not only on the current state of the GDMDP but also the wealth level. Thus the authors first gives an approach to transform the risk sensitive GDMDP (RS-GDMDP) into a risk neutral one by augmenting the state space of the RS-GDMDP with possible wealth levels. The authors than relate the values and policies of the original and augmented GDMDP. The resulting RS-GDMDP has an infinite state space but its properties allow us to generalize the standard risk sensitive version of Value Iteration (VI) which manipulates to a risk sensitive version of VI, which manipulates the functions that map wealth level to values.

2.3 Exponential form of Risk Sensitive MDPs

As seen above Mean-Variance MDPs have several issues. However by considering exponential cost, one can also reduce the fluctuations around the mean. Exponential cost cover some of the drawbacks that are present in Mean-Variance MDPs. In comparison with linear MDPs which control the first moment (expected value) of the sum cost, the risk sensitive MDPs also control the variability/fluctuations around the expected value, by considering the higher moments. In [4], the authors model finite horizon risk sensitive constrained MDPs using LP technique. The aim is to give varying importance to sample path trajectories and the expected value, as controlled by a parameter γ. Depending upon γ, called the risk parameter, it provides importance to higher moments of the sum cost. The linear MDPs are viewed as risk neutral MDPs with $\gamma = 0$. The aggregation function in this case equals the exponential of the sum of the running costs, $E[e^{-\gamma \sum_{t \leq T} R_t}]$. Note that $E[e^{-\gamma \sum_{t \leq T} R_t}] = E[\sum_k \gamma^k (\sum_{t \leq T} R_t)^k / k!]$, thus considering the higher moments of the sum cost. The authors have considered the following transformation of the risk neutral objective function to capture the higher moments.

$$\tilde{J}_o = (1/\gamma)\log(J_o) \text{ where} \qquad J_o = E[e^{\gamma \sum_t r_t(S_t, A_t)}]$$

This the risk sensitive objective function. For smaller values γ the objective takes the form:

$$\tilde{J}_o \simeq E[\sum_t r_t(S_t, A_t)] + (\gamma/2) Var[\sum_t r_t(S_t, A_t)] \tag{2.2}$$

As one can see as $\gamma \to 0$ the objective function approaches risk neutral cost. In this paper the authors provide connection between DP equations and two appropriate LPs for finite horizon case. The primal LP provides the value function while the dual LP provides the optimal policy. For LPs in risk sensitive MDPs the authors have circumvented this problem by incorporating the multiplicative cost term into the mapping that converts any given Markov policy to a feasible point of the LP. This due to the fact that it is possible to construct the linear objective function of the relevant LP using the running costs of all time slots but the cost accumulates in a multiplicative manner for risk sensitive MDPs and hence the same approach cannot be adapted directly for constructing an LP for risk MDP. Primal and dual LPs for finite horizon RSMDP are efficiently designed. By substituting $\gamma = 0$ in the Primal and Dual LP we get back to a risk neutral MDP which will maximise the expected cost. However there is inconsistency between those LP and they do

not point to same optimal policies. They have identified this disparity between the LPs and have corrected the inconsistency. However when we need to add extra constraints to this model, the fact that costs are multiplicative poses a problem. The authors have proposed a technique to add more constraints to the LP by augmenting the state space with an extra component, which is representative of this multiplicative cost. The addition of this extra component makes it easy to add new constraints .They have also discussed two applications of the problem. First application includes delay tolerant networks where a message has to be transferred from a source to a faraway destination with the help of occasional contacts between the freely moving nodes (that are willing to become the relays) and the source/destination. Second application includes a lossy (finite buffer) queuing system with two server modes and with a constraint on the utilization of fast server mode. They have also compared the risk neutral policies with risk sensitive policies for the constrained MDP problem.

2.4 Inventory Control

In [8], Herbert Scarf considers a dynamic inventory problem where he considers ordering costs, holding costs and shortage costs. The underlying assumption here is that the holding and shortage cost are linear. He shows that when this is true the optimal policy in each period is always of the type (S,s). This means that whenever the inventory level goes below level s will order so as to increase are inventory level to S, if not than we will not order. He builds a total cost function, composed of the three costs described above. This function can be used to construct a dynamic program equation. He shoes the total cost function follows K-convexity. Due to this the optimal policy the optimal policy turns out to be of such simple form.

Chapter 3

Inventory Control - Linear And Risk Sensitive MDP Framework

3.1 Inventory Control

Inventory refers to idle goods or materials that are held by an firm for use sometime in the future. Items stored in inventory can vary from raw materials, purchased parts, finished goods etc. One reason firms maintain inventory is that it is rarely possible to predict sales levels, production times, demand, and usage needs exactly. Thus, inventory serves as a buffer against uncertain and fluctuating usage and keeps a supply of items available in case the items are needed by the firm or its customers. While inventory serves an important and essential role, the expense associated with financing and maintaining inventories is a substantial part of the cost of doing business. In large organizations, the cost associated with inventory can run into the millions of dollars. Two important questions that must be answered in order to effectively manage inventories are as follows:

1. When should the inventory be replenished?
2. How much should be ordered when the inventory is replenished?

Sequential decision models have been widely applied to inventory control problems and are one its earliest applications. These models can be applied not only to find reorder points of single product, but also to complex multi product supply network. In this work, I consider a single product Inventory control system and try to find optimal reorder points. Under various scenarios it is proved that $(\mathcal{S}, \int)$ policy is optimal (e.g., [8]) when one considers optimizing the expected value of the costs in Inventory control. As a first part of my work, using numerical methods i.e dynamic programming and LPs I proved the optimality of $(\mathcal{S}, \int)$ in one example scenario.I have also considered risk sensitive cost (expected value of exponent of total costs) and studied the differences in the two optimal policies.

3.2 Model

Consider an Inventory storage that can be stocked with goods to a maximum size of M. Results proving optimality of (S,s) policy (e.g., [8]) often do not consider a limit on inventory size, however due to limitation of numerical computations we assume M as the maximum inventory size possible. This is also a realistic scenario. Let N be the number of planning horizons. A sequence of ordering decisions is made at the beginning of number of equally spaced intervals. The inventory builds upon ordering stock and depletes on fulfilling demand. Following assumptions apply to the model

1) Demand $\xi_1, \xi_2 ..... \xi_N$ are independent and identically distributed by a common discrete distribution function $P(\xi)$.

2) All orders are placed at the start of the period and received immediately and lag is zero.

3) Demands arrive immediately after the orders arrive and are fulfilled instantly.

4) Holding and shortage costs are charged linearly.

5) Unsatisfied demand is lost forever.

Now to define the different costs associated with the inventory process.

1) Order cost (C_O) : This contains two components:

a) Fixed order cost (K) : This the fixed cost for an order irrespective of the number of units ordered.

b) Variable cost : This is a linear cost cost directly proportional to number of units ordered.Thus the order cost when quantity a is ordered equals:

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

where c_u is the unit price.

2) Holding cost (C_H) : Cost associated with storing the inventory that remains unsold. This cost will depend on the demand.

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

where c_k is the unit holding cost.

3) Shortage cost(C_S) : Cost associated with not being able to fulfil demand from the stock.

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

where c_s is the unit shortage cost.

Hence now the total expected cost to be charged for any period $t$will be:

$$= (K + c_u a)\mathbf{1}_{\{a>0\}} + \sum_{\xi=0}^{s+a-1} c_H(s + a - \xi)P(\xi) + \sum_{\xi=s+a}^{\infty} c_S(\xi - s - a)P(\xi) \tag{3.1}$$

We now describe a sequential model to determine the optimal reorder points for single product inventory system. Decision epoch will be the review of the inventory level after each interval. The system state is the inventory level at the time of review. In a given state the action correspond to the amount of stock the order is placed for. Note that in our case we maximum order quantity so that are total inventory would not be greater than M. Once the order has been placed for a given state, the transition probabilities will depend on the quantity ordered and the random customer demand for the product in the period. After the demand is fulfilled we incur a cost (ordering + shortage + holding) depending on the order, stock on hand and the random demand. A decision rule will provide us the quantity to be ordered, given the state of the inventory at the time of review.

Thus we seek a policy that consists of such decision rules to minimize the total cost of the entire planning period. A Markov decision process formulation follows below:

Decision Epochs:

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

State space:

$$S = \{0, 1, 2, ...M\}$$

Actions:

$$A_s = \{0, 1, 2, ...M - s\}$$

where $s \in S$
Expected Rewards:

$$r_t(s, a) = (C_O(a) + C_H(s, a) + C_S(s, a))$$

Terminal reward $r_T(s) = 0$.
Transition Probability:

$$p(j|s, a) = \begin{cases} P(\xi = s + a - j), & s + a > j \geq 0 \\ P(\xi \geq s + a), & j = 0 \\ 0, & \text{else} \end{cases}$$

3.3 Linear Cost - Standard MDP Model

3.3.1 MDP with cost considered in the same period

Our objective is to minimize the expected value of the sum of all running costs, that is $E[\sum_t r_t(s, a)]$. This equivalent to optimizing:

$$E[\sum_t r_t(s, a)] = \sum_t E[r_t(s, a)] \tag{3.2}$$

$$= \sum_t E\Big[E\Big[R_t|(S_1, A_1), (S_2, A_2), ..., (S_{T-1}, A_{T-1}), (S_T)\Big]\Big]. \tag{3.3}$$

Hence given the current state s and action to be taken a we can find the cost any period t. The cost r_t for the period t will depend only on the state s_t and the actiona_t in that period and can be calculated as follows:

$$E[R_t(s, a)|(S_t = s, A_t = a)] = E[C_O(a) + C_H(s, a) + C_S(s, a)]$$

This quantity is easy to compute, given s, a and the distribution of demand ξ. Hence now we can compute $E[r_t(s, a)]$ by simply multiplying with the probabilities $\{P(S_t = s, A_t = a)\}$

In [4], authors have given a Primal and Dual LPs to solve finite horizon linear program. The primal will give us the optimum cost, while the dual gives the optimum policies. The dual LP is

given below.

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

subject to:

$$\sum_{a} y(0, s', a) = \alpha(s') \qquad \text{for all } s' \in S$$

$$\sum_{a} y(t, s', a) = \sum_{a} \sum_{s} p(s'|s, a) y(t-1, s, a) \qquad \text{for all } 1 \leq t \leq T-1 \text{ and } s' \in S$$

and $y(t, s, a) \geq 0$ for $a \in A_s$ and $s \in S$.

Using the above LP we solve the inventory problem and results are given in chapter 5. The Above MDP model is applicable for any discrete distribution of demand. But if the demand was to be continuous than we would need to change the MDP model as state space becomes infinite in that case and we need to come up with a idea. Then we would need to discretize the demand to keep the state space finite.

3.3.2 MDP after shiting the cost terms

There is another way which can be used to represent the rewards in this MDP model. We consider shift of few cost terms (from any time slot to its next), which would not change the actual problem and which would pave way for risk MDP. In formulation given in eq [], the holding and shortage costs depend upon the current inventory, order quantity and demand. In actual it depends only on the next state. 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. 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. So one can shift the holding and shortage costs to the next period where we would know the state. If $s = 0$ than we for sure know that there was no holding cost and only shortage cost would apply. Hence in this case we need to model the excess demand, when the state is zero. The reason for modelling the rewards this way is highlighted in the next section where we model Risk Sensitive MDPs. Considering this we can use the following:

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

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. Note that there exist terminal reward in this case. Hence with these new rewards, we can again solve the LP to get the optimal policy. Section 4.2 provides the results for this way of modelling. 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.

3.4 Risk Sensitive MDP

In risk sensitive MDPs we want to give importance to higher moments of the total rewards. Hence our objective is to optimize $E[e^{\sum_{t=1}^{T}\gamma r_t}]$

$$\begin{aligned}E[e^{\sum_{t=1}^{T}\gamma r_t}] &= E[\Pi_{t=1}^{T}e^{\gamma r_t}] \\ &= E\Big[E\Big[\Pi_{t=1}^{T}e^{\gamma R_t}|(X_1,X_2,...X_T)(A_1,A_1,...A_{T-1})\Big]\Big]\end{aligned}$$

As compared to linear cost MDPs we cannot simplify this easily. If one notices in linear MDPs **??**, we can simplify the total in terms of sum of expectations of the costs which is easy to compute as in **??** .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. Hence using we need to use the technique where we shift the rewards to the next time period. We will consider a the transformed risk reward as below: $E[e^{\sum_{t=1}^{T}\gamma r_t}]$

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

substitute the shifted reward from equation

$$\begin{aligned}&= E[E[\Pi_{t=1}^{T-1}e^{c_oA_t}\Pi_{t=1}^{T}e^{(c_hS_t)\mathbf{1}_{\{S_t>0\}}+(c_s\tilde{\xi})\mathbf{1}_{\{S_t=0\}}}|(X_1,X_2,...X_T)(A_1,A_2,...A_{T-1})]] \\ &= E[\Pi_{t=1}^{T-1}e^{c_oA_t}\Pi_{t=1}^{T}e^{(c_hS_t)\mathbf{1}_{\{S_t>0\}}}E[\Pi_{t=1}^{T-1}e^{(c_s\tilde{\xi})\mathbf{1}_{\{S_t=0\}}}|(X_1,X_2,...X_T)(A_1,A_2,...A_{T-1})]]\end{aligned}$$

In the above rewards the shortage cost depends on the random quantity $\tilde{\xi}$ (excess demand). 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.

$$\begin{aligned}E[\Pi_{t=2}^{T}e^{(c_s\tilde{\xi})\mathbf{1}_{\{S_t=0\}}}|(X_1,X_2,...X_T)(A_1,A_2,...A_{T-1})] &= E[\Pi_{t=1}^{T-1}e^{(c_s\tilde{\xi})\mathbf{1}_{\{S_t=0\}}}] \\ &= \Pi_{t=2}^{T}E[e^{(c_s\tilde{\xi})\mathbf{1}_{\{S_t=0\}}}]\end{aligned}$$

Hence now we can write the transformed formed of the rewards as below:

$$r_t(s_t,a_t) = \begin{cases} c_oa_t & ,t=1 \\ c_oa + c_h(s_t)\mathbf{1}_{\{s_t>0\}} + \tilde{c})\mathbf{1}_{\{s_t=0\}} & ,1<t<T \\ c_h(s_t)\mathbf{1}_{\{s_t>0\}} + \tilde{c}\mathbf{1}_{\{s_t=0\}} & ,t=T \end{cases}$$

where $\tilde{c} = log(E[e^{(c_s\tilde{\xi})}])$.
Under this assumption one can apply directly use the risk sensitive LP to solve the MDP model from [4]. The dual LP to get the policy.

$$min \quad \sum_a\sum_s b_{(x,a)}y(T-1,s,a)$$

subject to:

$$\sum_a y(0,s',a) = \alpha(s') \qquad \text{for all } s' \in S$$

$$\sum_a y(t,s',a) = \sum_a\sum_s e^{\gamma r_{t-1}(s,a)}p(s'|s,a)y(t-1,s,a) \quad \text{for all } 1 \le t \le T-1 \text{ and } s' \in S$$

$$\text{with } b_{(x,a)} := e^{\gamma r_{T-1}(s,a)} \sum_{s'} p(s'|s,a) e^{r_T(s')} \text{ and } y(t,s,a) \geq 0 \text{ for } a \in A_s \text{ and } s \in S.$$

If we let $\gamma \to 0$, authors in [4] have shown that the risk sensitive cost converges to risk neutral cost. Similarly in this model the LP will converge to risk neutral LP. In the appendix the convergence of the risk sensitive cost is shown. This is an important step since $\tilde{c}$ is a function of gamma.

Chapter 4

Results

4.1 Risk Neutral MDP

4.1.1 Average Cost Model

Let us consider maximum Inventory level $M = 20$. The total planning horizons $N = 10$. Let us consider demand distribution to be geometric with parameter $p = 0.3$. Let $K = 0.2, c_u = 0.4, c_h = 0.2, c_s = 1$. Given these we can compute the rewards $r_t(s, a)$ and the necessary transition probability $p(s'|s, a)$ for the problem. $\frac{P(z<\xi\leq q+z)}{P(\xi>z)}$ $\int$

Bibliography

[1] Jerzy A Filar, Lodewijk CM Kallenberg, and Huey-Miin Lee. Variance-penalized markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.

[2] Abhijit Gosavi. A risk-sensitive approach to total productive maintenance. *Automatica*, 42(8):1321–1330, 2006.

[3] Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972.

[4] Atul Kumar, Veeraruna Kavitha, and N Hemachandra. Finite horizon risk sensitive mdp and linear programming. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 7826–7831. IEEE, 2015.

[5] Yaxin Liu and Sven Koenig. Risk-sensitive planning with one-switch utility functions: Value iteration. In *AAAI*, pages 993–999, 2005.

[6] Shie Mannor and John Tsitsiklis. Mean-variance optimization in markov decision processes. *arXiv preprint arXiv:1104.5601*, 2011.

[7] Martin L Puterman. Market decision process. *Chapter*, 8:5–1, 1990.

[8] Herbert Scarf. The optimality of (s, s) policies in the dynamic inventory problem. *Mathematical Methods in the Social Sciences*, 1959.