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 a reward and the system state evolves to possibly different state at the next epoch depending on the controlled 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 (). 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 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 a function of these reward sequences. In general MDP optimizes the expected value of the function that deﬁnes a way of aggregating the running costs related to all the time slots under consideration, i.e., to optimize . Linear MDPs consider expected value

of either sum of all the running costs , or sum of discounted values of all the running costs with discount factor or time average of the running costs .They are respectively called total cost, discounted cost or average cost problems. Also MDP problem can either be modelled over a ﬁnite time horizon or over the inﬁnite time horizon.

Linear MDPs control the first moment (expected value) of the sum cost. Suppose if we want to control the fluctuations around the expected value than we need to turn to Risk Sensitive MDPs (RSMDPs).

Linear MDPs are very well studied and one can use Dynamic programming, Linear programming, and Value Iteration approaches to solve them. In [], Howard and Matheson have discussed the application of Value Iteration and Policy Iteration to risk sensitive case. However in my work I have used LP to solve the required MDP. This is based on the work in []. The problem I have selected is the classic inventory control case.

The **Inventory Control problem** is the problem faced by a decision maker who must decide how much to order in each time period to meet demand for its products. This is stochastic demand problem, where the demand follows a known distribution. The objective here is to find a policy of ordering at any time epoch so as to minimize the cost incurred for total period of T which makes it a finite horizon problem. The various cost incurred depends on the current state of inventory, amount ordered and demand. There three basic cost applicable in general situations. Ordering cost that includes unit ordering price and also fixed order setup cost. Holding cost for carrying inventory into the next period. Shortage cost for not fulfilling any demand. This is problem first has to be modelled by into a sequential decision model defining appropriate state space, actions, transition probability and rewards at each time epoch. In section 3 I have defined the model and the notations. The decision policy that the decision maker has to follow if we optimize the expected reward only is known and is 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. In [], Herbert Scarf has shown that when holding cost and shortage cost are linear then the optimal policy in each period is always of this type. The aim in my work is to find optimal policies for a risk averse decision maker. [] has given LP for risk sensitive MDPS which can be applied for this example.

Literature Survey

In the present world, optimization of the expected cost of an MDP is well studied. Optimality equations are very familiar and solution to these equations correspond to optimal value functions and provide basis to determine optimal policies. Dynamic programming (DP) is very popular technique to find optimal policies for finite horizon problems. Basically one can use backward induction algorithms to find these optimal policies. However for infinite horizon one needs to consider the discounting factor. Using optimality equations and their properties and the theory of equations on normed linear spaces one can find the existence of a solution to the MDP. Many algorithms like Value Iteration and Policy Iteration have been defined to find optimal policies for infinite horizons. Also the relationship between discounted MDP’s and Linear programs (LP’s) is pretty well defined. Though not efficient LP’s have elegant theory, ease in addition of additional constraints and facility of sensitivity analysis make it very attractable in some cases.

Classical MDP’s deal with the maximisation of cumulative reward, however a decision maker may focus on several other distributional properties. Most decision makers maker 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 field to manufacturing fields. In [], authors deal with various computational complexities of mean-variance optimisation. The main problem with mean-variance optimisation is the absence of principle of optimality which could have led to simple recursive algorithms. Also variance is not a linear measure of the probability measure of the process. This is addressed using constrained Markov Decision Process to develop pseudo polynomial exact or approximate algorithms. They have defined their MDP problem first. The question they are addressing is that given a MDP and rational numbers λ, ν does there exist a policy in the set π such that and They compared different types of policies classes and established NP-hardness of the problem. Than they have identified cases for the existence of pseudo polynomial time algorithms. Approximate algorithms are also given where they approximate mean and variance.

In [], authors applied risk sensitive approach to total productive maintenance (PM). Their aim was to find optimal productive maintenance time so as to minimize 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 heuristic approach. A mean variance model where the objective function is with where and 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 as the time for PM. They have developed separate models for the individual-unit scenario and the production-line scenario. For the case of the individual unit renewal-theory model is used and for the case of the production line a more involved model based on MDPs is used. This involves forming a closed form equation for or solving it numerically. For the MDP it initially gives a Quadratic Program to solve the MDP. But this is computationally expensive. Therefore a surrogate objective function was developed to represent the original objective function. Using this function they developed a computationally lighter DP approach and showed its convergence. Results showed that this was the case as the surrogate function did mimic the exact function reasonably.

In [], 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. These function model the risk attitude of decision maker. 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 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 give 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 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.

In [], the authors give varying importance to sample path trajectories and the expected value, as controlled by a parameter. Depending upon, called the risk parameter, it provides importance to higher moments of the sum cost. The aggregation function in this case equals the exponential of the sum of the running costs, . Note that. While the linear MDPs 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. The linear MDPs are also viewed as risk neutral MDPs with. 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 LP’s 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. The authors also have discussed how to add more constraints to the LP by augmenting the state space with an extra component, which is representative of this multiplicative cost. 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.

In [], 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.

Problem Description

Consider an Inventory system that can be stocked to a maximum size of M. 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 ξ are independent and identically distributed by a common distribution function f(ξ).
2. All orders are placed at the start of the period and received immediately and lag is zero.
3. Demands arrive at the start of the period and are fulfilled instantly.
4. Holding and shortage costs are charged linearly.
5. Unsatisfied demand is lost forever.

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

1. Order cost – This contains two components:
2. Fixed order cost (K) – This the fixed cost for an order irrespective of the number of units you order.
3. Variable cost – This is a linear cost cost directly proportional to number of units you order.

Where a is the amount of quantity ordered and is the unit price.

1. Holding cost – Cost associated with storing the inventory that remain unsold. This cost will depend on the demand.

Where is the unit holding cost.

1. Shortage cost – Cost associate with not being able to fulfil demand from the stock.

Where is the unit shortage cost.

A Markov decision process formulation follows below:

Decision Epochs:

T = { 1,2,3……N }, N ≤ infinity.

State Space:

S = { 0,1,2,…..M }.

Actions:

As = { 0,1,2,…..M-s }.

Expected Rewards:

Transition Probability: