CE 273
Markov Decision Processes

Lecture 25

Partially Observable and Risk-Sensitive MDPs

Previously on Markov Decision Processes

The objective in the discounted cost MDP problem is

$$\lim_{N\to\infty} \mathbb{E}_w \sum_{k=0}^{N-1} \left\{ \alpha^k g(x_k, u_k, w_k) \right\}$$

Under most practical situations that we encounter, this limit exists and we can also exchange the limit and expectation and write

$$\mathbb{E}_w \sum_{k=0}^{\infty} \left\{ \alpha^k g(x_k, u_k, w_k) \right\}$$

Likewise, given a particular policy $\pi = \{\mu_0, \mu_1, \ldots\}$, the value function can be written as

$$J_\pi(x_0) = \lim_{N\to\infty} \mathbb{E}_w \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}$$

We will make appropriate assumptions (such as bounded costs) that will guarantee the existence of the above limit.

Lecture Outline

1. Partially Observable MDPs
2. Risk-Sensitive MDPs

Lecture Outline

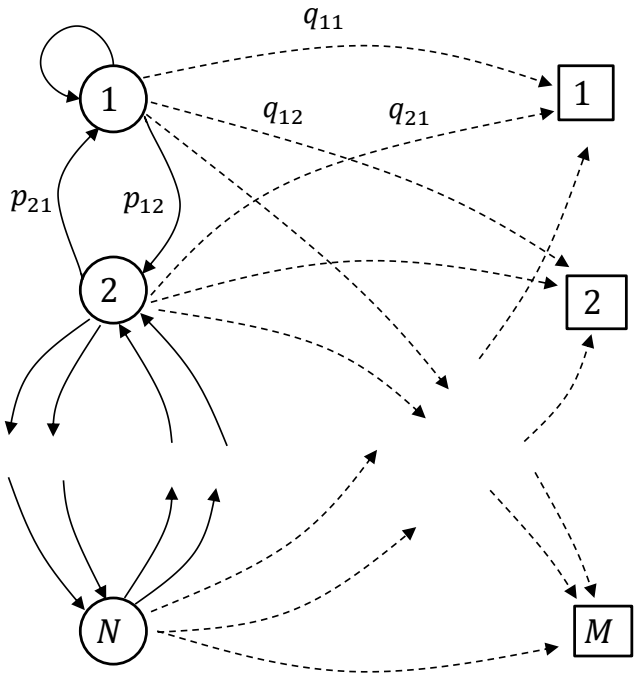
Partially Observable MDPs

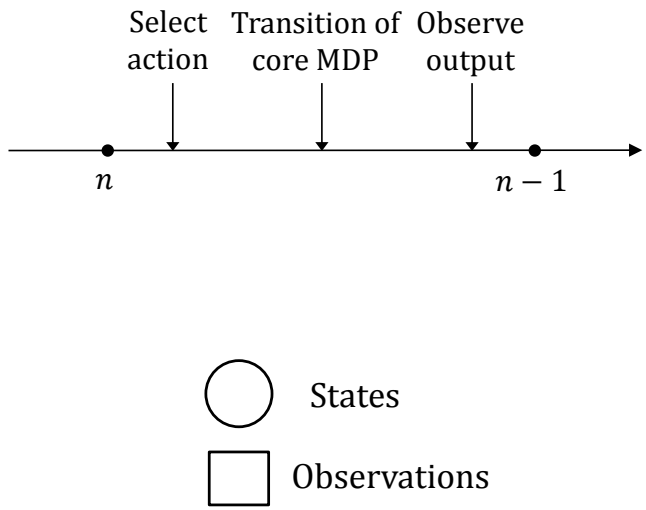
Partially Observable MDPs

Introduction

In many situations, decision makers may not have access to the state of the system but can only see some related observations.

- Uncertainty regarding the state of a MDP
- Allows information acquisition
- Transition probabilities are known



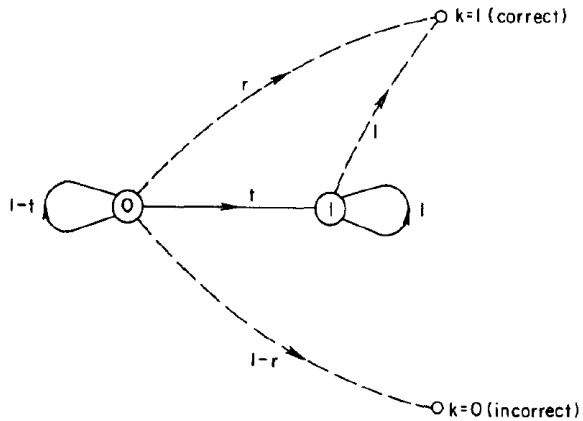


Partially Observable MDPs

Applications: Human Learning

- **States:** Unlearned (0) and learned (1)
- **Observations:** Correct or incorrect
- **Actions:** Simply present item, Present and observe students response, Remove the item

The objective is to minimize the instruction costs associated with the three actions.

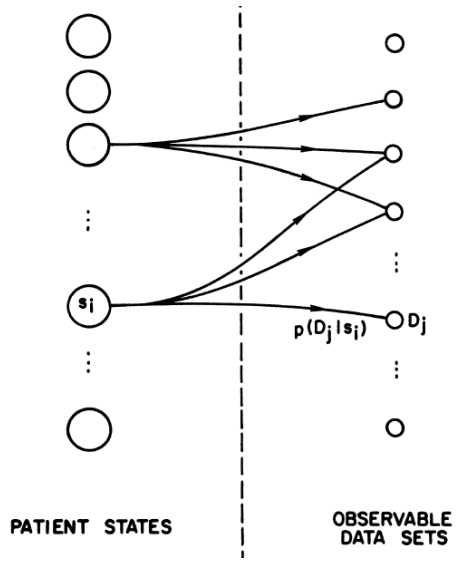


- Smallwood, R. D. (1971). The analysis of economic teaching strategies for a simple learning model. Journal of Mathematical Psychology, 8(2), 285-301.

Partially Observable MDPs

Applications: Medical Diagnosis

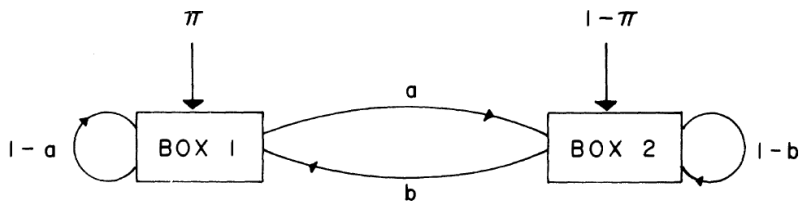
- **States:** Physiological or psychological states
- **Observations:** Physician state of information
- **Actions:** Treatments



Partially Observable MDPs

Applications: Searching for Moving Objects

- **States:** Object's location
- **Observations:** Whether the object is at a location
- **Actions:** Look at locations



Objectives:
What strategy will produce the minimum expected number of looks needed to detect the target, and what is the value of this minimum number of looks?

What strategy will produce the maximum probability of detecting the target within n looks available, and what is the value of this maximum probability?

- Pollock, S. M. (1970). A simple model of search for a moving target. Operations Research, 18(5), 883-903.

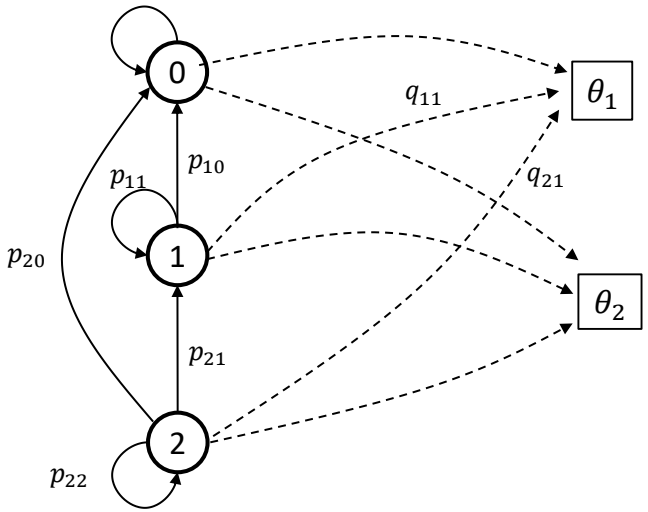
Partially Observable MDPs

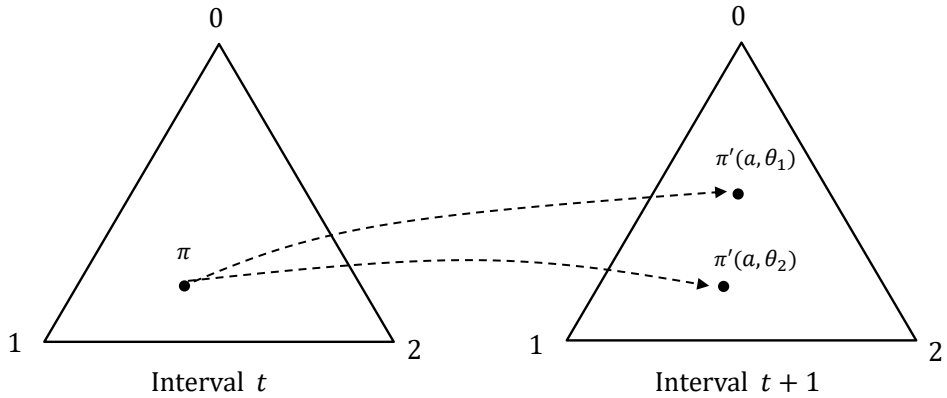
Example: Machine Replacement

Consider a machine with two components:

- **States:** 0, 1, 2 (Number of failed components)
- **Observations:** θ_1(non-defective), θ_2(defective)
- **Actions:** Manufacture, Examine, Inspect, Replace

Let $\pi = [\pi_0, \pi_1, \pi_2]$ be a probability distribution over the state space, also called as *information vectors*.

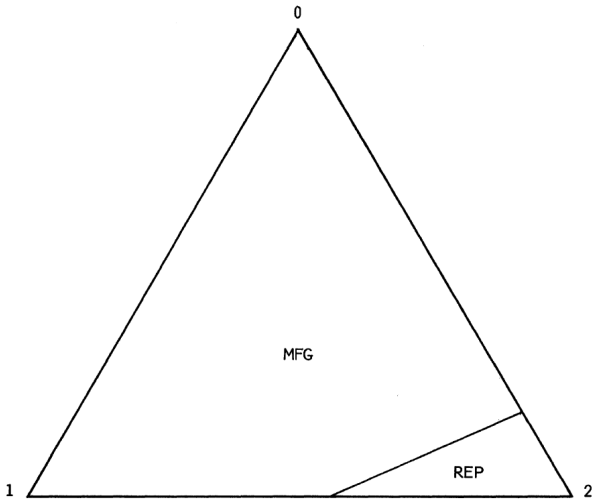




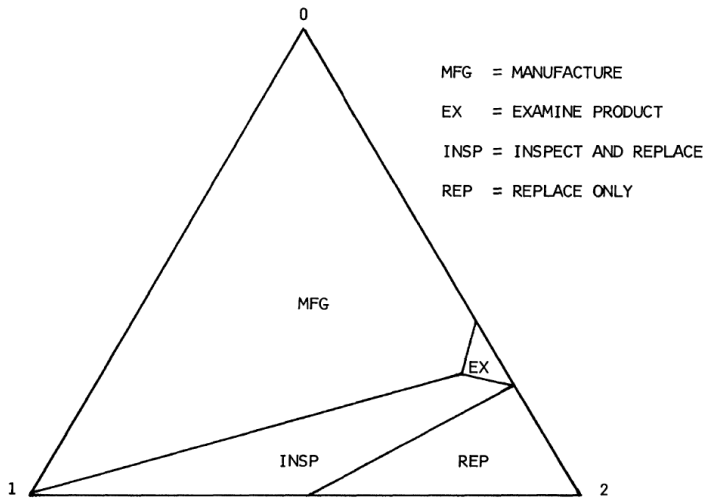
Partially Observable MDPs

Example: Machine Replacement

Actions	P(a)			Q(a)		R(a)
Manufacture	0.81	0.18	0.01	1	0	0.9025
	0	0.9	0.1	1	0	0.475
	0	0	1	1	0	0.25
Examine	0.81	0.18	0.01	1	0	0.6525
	0	0.9	0.1	0.5	0.5	0.225
	0	0	1	0.25	0.75	0
Inspect	1	0	0	1	0	-0.5
	1	0	0	1	0	-0.15
	1	0	0	1	0	-0.25
Replace	1	0	0	1	0	-2
	1	0	0	1	0	-2
	1	0	0	1	0	-2



Optimal policy for the control horizon with n = 3



Optimal policy for the control horizon with n = 4

Partially Observable MDPs

Finite Horizon Model Formulation

Time periods T: A finite number of time steps

Core Process $\{x_t\}_{t \in T}$: A finite state Markov chain with N states

Observation Process $\{y_t\}_{t \in T}$: Outputs observed

Action Space A: a_t represents an action taken at time t

Transition Probabilities:

$$p_{ij}(a_t) = \Pr[x_{t+1} = j | x_t = i, a_t]$$

The probability that the core process moves to j from i when a_t is chosen.

$$q_{jk}(a_t) = \Pr[y_{t+1} = k | x_{t+1} = j, a_t]$$

Probability that we observe k when the core process moves to j.

Partially Observable MDPs

Finite Horizon Model Formulation

Data available for decision making at t: $d_t = (\pi(0), k_1, a_1, \ldots, a_{t-1}, k_t)$

Let $\pi(t) = [\pi_1(t), \pi_2(t), \ldots, \pi_N(t)]$ be a distribution over the state space, where $\pi_i(t) = \mathbb{P}[x_t = i \mid d_t]$

It can be shown using Bayes' formula that

$$\begin{aligned}\pi_j(t+1) &= \mathbb{P}[x_{t+1} = j \mid d_{t+1} = (d_t, a_t, y_{t+1} = k)] \\ &= \frac{q_{jk}(a_t) \sum_i p_{ij}(a_t)\pi_i(t)}{\sum_{j'} q_{j'k}(a_t) \sum_i p_{ij'}(a_t)\pi_i(t)}\end{aligned}$$

Note that $\pi(t)$ contains all the information that the decision maker can extract to choose a control at time t.

Proposition

For any fixed sequence of actions $a_1, a_2, \ldots, a_t \in A$, the probabilities $\{\pi(t)\}_{t \in T}$ is a Markov process.

Hence, a POMDP can be converted into a Markov decision process.

However, note that while the core process was defined on a finite state space, the modified Markov process is defined on an uncountable state space.

Partially Observable MDPs

Components of POMDP

States: $\pi(t)$

Control: $a_t \in A$

Dynamics: $\pi(t+1) = f_t(\pi(t), a_t, k)$

Disturbance: $\pi_j(t+1) = \dfrac{q_{jk}(a_t)\sum_i p_{ij}(a_t)\pi_i(t)}{\sum_{j'} q_{j'k}(a_t)\sum_i p_{ij'}(a_t)\pi_i(t)}$

Rewards: Suppose that a reward of $w_{ijk}(a_t)$ is received when action a_t is taken and the core process makes a transition from i to j and produces an output k.

Then, the expected one-step reward is $\pi(t)^{\mathsf{T}} g(a_t)$, where

$$g_i(a_t) = \sum_j \sum_k w_{ijk}(a_t) p_{ij}(a_t) q_{jk}(a_t)$$

Partially Observable MDPs

Bellman Equations

Suppose the terminal rewards are $J_N(\pi)$ are assumed. Then, the Bellman equations can be written as

$$J_t\left(\pi(t)\right) = \max_{a \in A} \left\{ \pi(t)^{\mathsf{T}} g(a) + \sum_k \mathbb{P}\left[k | \pi(t), a\right] J_{t+1}\left(f_t(\pi(t), a, k)\right) \right\}$$

$$= \max_{a \in A} \left\{ \pi(t)^{\mathsf{T}} g(a) + \sum_{i,j,k} \pi_i(t) p_{ij}(a) q_{jk}(a) J_{t+1}\left(f_t(\pi(t), a, k)\right) \right\}$$

Partially Observable MDPs

Sondik's Algorithm

Theorem

$J_t(\pi)$ is piecewise linear and convex, and can thus be written as $J_t(\pi(t)) = \max_s \{\sum_i \alpha_i^s(t)\pi_i(t)\}$ for set of vectors $\alpha^s(t)$, where $s = 1, 2, \ldots, S$.

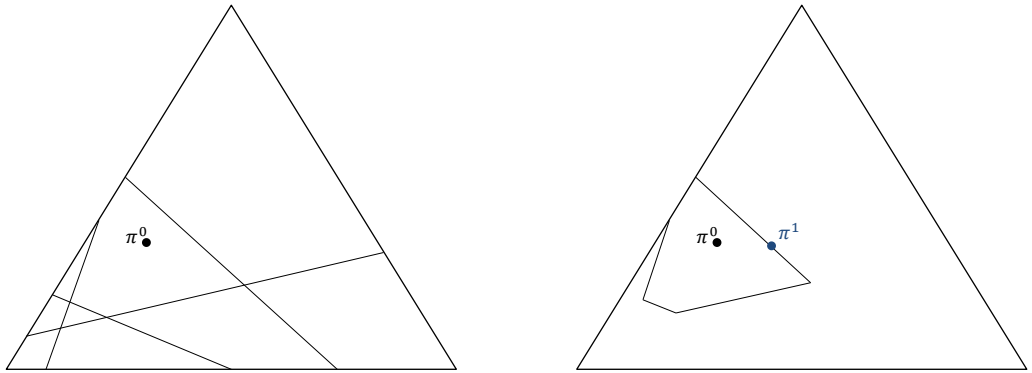
Suppose we are given the vector of α's at time step $t+1$, then the Bellman's equations can be re-written as follows:

$$J_t(\pi(t)) = \max_{a \in A} \left\{ \sum_i \pi_i(t) \underbrace{\left\{ g_i(a) + \sum_{j,k} p_{ij}(a) q_{jk}(a) J_{t+1}\big(f_t(\pi(t), a, k)\big) \right\}}_{\alpha \text{ values}} \right\}$$

Partially Observable MDPs

Sondik's Algorithm

The information vector space can thus be partitioned into convex regions where a certain action is optimal.



The algorithm proceeds by selecting an arbitrary π^0 and solving the Bellman equations to first get the optimal α^* vector.

A system of linear equations are then constructed for which $J_t(\pi(t)) = \pi(t)^{\mathsf{T}}\alpha^*$. The same action is optimal in the state space enclosed by these equations.

Another vector π^1 is selected on the boundary of this region and this process is repeated till the state space is fully explored.

Partially Observable MDPs

Additional Reading

- Monahan, G. E. (1982). State of the art-a survey of partially observable Markov decision processes: theory, models, and algorithms. Management Science, 28(1), 1-16.
- Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable Markov processes over a finite horizon. Operations research, 21(5), 1071-1088.

Risk-Sensitive MDPs

Risk-Sensitive MDPs

Introduction

The objectives in all the models we saw so far minimized or maximized the expected value of a random variable.

Consider the following experiment. Imagine an unbiased coin is tossed and if it lands on H, we get ₹2. If we see T, the game ends.

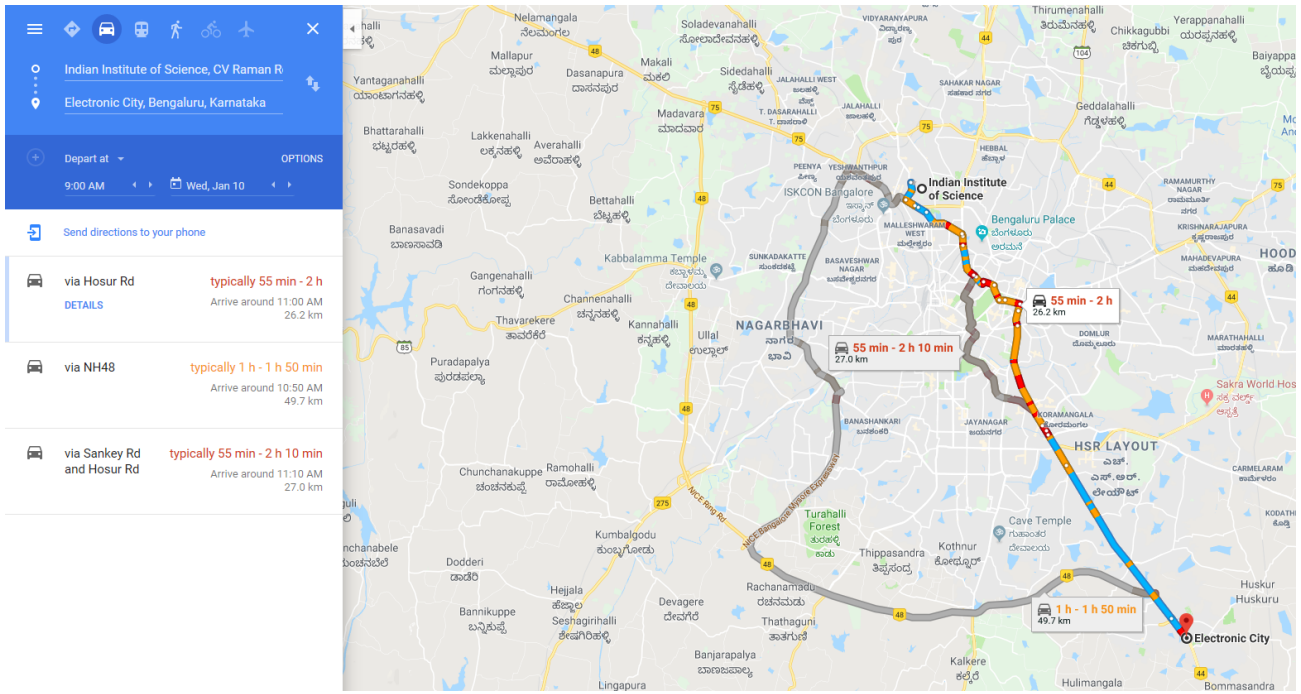
Every time we see H, the coin is tossed one more time and the winnings are doubled. Would you pay ₹100 to enter this game? What is the expected amount you could win?

This example, called the St. Petersburg paradox, indicates that individuals don't always care about the expected value of an outcome and can have some risk preferences.

Risk-Sensitive MDPs

Introduction

The objectives in MDPs have been extended to incorporate risk measures using some transformations of the rewards or using the variance.



Risk-Sensitive MDPs

Introduction

Consider a lottery in which you can win a reward of ₹0 or ₹100 with equal probability.

Suppose you are to choose between entering the lottery and receiving ₹5 for sure, which option would you choose? What if I give you ₹20 for sure?

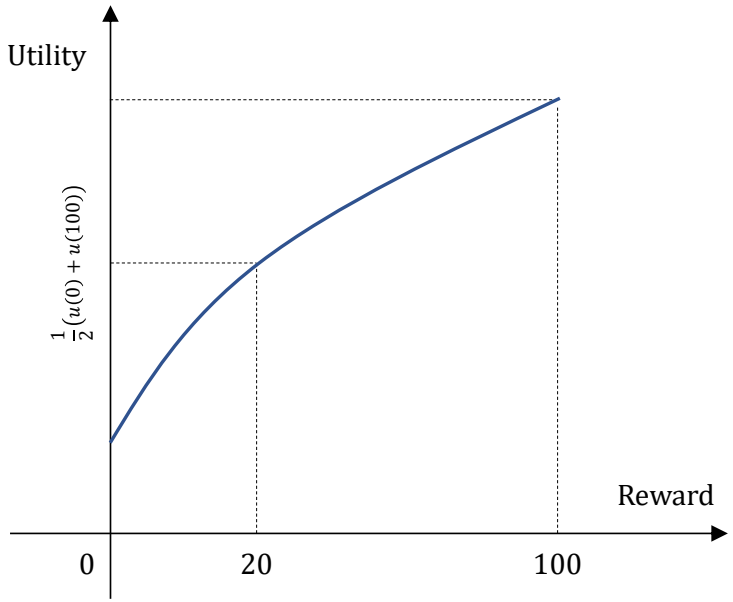
The smallest amount for which you would be willing to pass up the lottery is called **certainty equivalent**.

Risk-averse behavior is commonly modeled using this idea and concave utility functions in behavioral economics.

Risk-Sensitive MDPs

Introduction

One can also model risk-seeking behavior using convex utility functions.



Suppose the earlier example is modified as follows. The lottery gives you a reward of ₹10,000 or ₹10,100 with equal probability or ₹10,020 for sure. Which option would you choose?

The answer is likely to change for most individuals, but for mathematical reasons, we will assume that the certainty equivalent is independent of the current wealth.

Risk-Sensitive MDPs

Exponential Utilities

Risk attitudes of individuals can be captured using exponential utility functions of the following form

$$u(g) = -(\text{sgn}\gamma)e^{-\gamma g}$$

where sgn is the signum function, g is the reward, and u denotes the utility. These type of functions satisfy the property that the certainty equivalent is independent of current wealth.

A positive γ implies that the utility function is concave and represents risk aversion. A negative γ on the other hand implies risk seeking attitude.

One can find the γ value of an individual using some questions like the ones presented before and by fitting the above curve to match the data.

Risk-Sensitive MDPs

Bellman Equations for Risk-Sensitive MDPs

One way to model an MDP for a risk-sensitive decision maker is to transform the problem from finding the expected rewards to finding the expected utilities using the certainty equivalents.

Imagine a risk-averse agent (positive γ) in a finite horizon setting who follows a policy $\{\mu_k\}_{k=0}^{N-1}$. Fixing the policy induces a time-varying Markov chain for which we can write

$$u_k(J_k(i)) = \sum_{j=1}^{n} p_{ij}(\mu_k(i)) u_{k+1}\Big(g_k(i, \mu_k(i), j) + J_{k+1}(j)\Big)$$

J values here represent the certainty equivalents. Recall that for risk-sensitive agents $u(g) = -e^{-\gamma g}$. Hence, the above equation can be written as

$$u_k(J_k(i)) = \sum_{j=1}^{n} p_{ij}(\mu_k(i)) e^{\gamma g_k(i, \mu_k(i), j)} u_{k+1}\Big(J_{k+1}(j)\Big)$$

Defining $p_{ij}(\mu_k(i))e^{\gamma g_k(i,\mu_k(i),j)}$ as $q_{ij}(\mu_k(i)$, the disutility contribution matrix, we get

$$u_k(J_k(i)) = \sum_{j=1}^{n} q_{ij}(\mu_k(i)) u_{k+1}\Big(J_{k+1}(j)\Big)$$

Thus, one can solve for the utilities using backward induction and recover the certainty equivalents using

$$J_k(i) = -\frac{1}{\gamma} \ln[-\gamma u_k(i)]$$

So far, we have found the value functions associated with a policy. We can use this as a subroutine in a PI-like algorithm.

Extensions to infinite horizon discounted and average cost MDPs also exist.

Risk-Sensitive MDPs

Alternate Approaches

The reason for using exponential utilities is that we can derive DP algorithms to address the problem.

However, in many transportation and finance applications, we wish to minimize an objective consisting of some function the expected value and variance of a random variable.

- Mean-variance tradeoff: $J - \lambda\sqrt{V}$
- Sharpe Ratio: $J/\sqrt{V}$

These objectives are not DP friendly but we can find still find the variance of a policy.

Risk-Sensitive MDPs

Variance of MDPs

Consider a random variable G_μ which represents the discounted rewards of an MDP over an infinite horizon assuming some initial distribution (ignored in the notation for brevity)

$$G_\mu = \sum_{k=0}^{\infty} \alpha^k g(x_k, u_k, x_{k+1})$$

Let F represent the CDF of the above random variable. That is,

$$F_\mu(y) = \mathbb{P}[G_\mu \leq y]$$

Let the mth moment of G_μ be denoted as $\Lambda_\mu^{(m)}$, i.e.,

$$\Lambda_\mu^{(m)} = \int_0^\infty y^m dF_\mu(y)$$

What is $\Lambda_\mu^{(1)}$? J_μ. Thus, the variance of the policy can be written as

$$V_\mu = \Lambda_\mu^{(2)} - (J_\mu)^2$$

Risk-Sensitive MDPs

Variance of MDPs

Define a vector $\theta_\mu \in \mathbb{R}^n$ as

$$\theta_\mu(i) = \sum_{j=1}^{n} p_{ij}(\mu(i))\Big(g(i,\mu(i),j) + \alpha J_\mu(j)\Big)^2 - (J_\mu)^2$$

Proposition

The vector of variances solves the following system of equations

$$V_\mu = \theta_\mu + \alpha^2 P_\mu V_\mu$$

Hence, variances can be derived from a policy evaluation-like step using $(I - \alpha^2 P_\mu)^{-1}\theta_\mu$.

Risk-Sensitive MDPs

Variance of MDPs

It turns out that variances do not satisfy the monotonicy property of DP.

Some studies have tried to formulate alternate frameworks in which the objective is to optimize the expected value subject to a constraint on the variance.

RL methods for solving this problem using policy gradients also exists.

Risk-Sensitive MDPs

Additional Reading

- Howard, R. A., & Matheson, J. E. (1972). Risk-sensitive Markov decision processes. Management science, 18(7), 356-369.
- Sobel, M. J. (1982). The variance of discounted Markov decision processes. Journal of Applied Probability, 19(4), 794-802.
- Sobel, M. J. (1994). Mean-variance tradeoffs in an undiscounted MDP. Operations Research, 42(1), 175-183.

Your Moment of Zen