Reinforcement Learning

Introduction

Subramanian Ramamoorthy
School of Informatics

17 January 2017

Admin

Lecturer: Subramanian (Ram) Ramamoorthy

IPAB, School of Informatics

s.ramamoorthy@ed (*preferred* method of contact)

Informatics Forum 1.41, x505119

Teaching Assistant: Svetlin Penkov, SV.Penkov@ed.ac.uk

Class representative?

Mailing list: Are you on it – I will use it for announcements!

Admin

Lectures:

Tuesday and Friday 12:10 - 13:00 (Teviot Lecture Theatre)

Assessment: Homework/Exam 10+10% / 80%

- HW1: Out 3 Feb, Due 17 Feb
  - Some pen+paper questions & setup for HW2
- HW2: Out 28 Feb, Due 28 Mar
  - Implement an agent using ALE

Admin

Webpage: www.informatics.ed.ac.uk/teaching/courses/rl

- Lecture slides to be uploaded, typically, the day before

Main Textbook

- R. Sutton and A. Barto, Reinforcement Learning, 2nd ed.

Other Suggested Books

- See list on course web page: going into more detail on specific topics (e.g., dynamic programming, POMDPs, active sensing) than possible or necessary in lectures. These books are FYI and optional.

Other readings, e.g., papers, to be suggested for specific lectures.

Background: Mathematics, Matlab, Exposure to machine learning?

Problem of Learning from *Interaction*

- with environment
- to achieve some goal

- Baby playing. No teacher. Sensorimotor connection to environment.
  - Cause <-> effect
  - Action <-> consequences
  - How to achieve goals
- Learning to drive a car, hold conversation, etc.

- Environment's response affects our subsequent actions
- We find out the effects of our actions later

Rough History of RL Ideas

- Psychology – learning by trial and error

  … actions followed by good or bad outcomes have their tendency to be reselected altered accordingly
  - Selectional: try alternatives and pick good ones
  - Associative: associate alternatives with particular situations

- Computational studies (e.g., credit assignment problem)
  - Minsky's SNARC, 1950
  - Michie's MENACE, BOXES, etc. 1960s

- Temporal Difference learning (Minsky, Samuel, Shannon, …)
  - Driven by differences between successive estimates over time

Rough History of RL, contd.

- In 1970-80, many researchers, e.g., Klopf, Sutton & Barto,…, looked seriously at issues of "getting results from the environment" as opposed to supervised learning
  - Although supervised learning methods such as backpropagation were sometimes used, emphasis was different

- Stochastic optimal control (mathematics, operations research)
  - Deep roots: Hamilton-Jacobi → Bellman/Howard
  - By the 1980s, people began to realize the connection between MDPs and the RL problem as above…

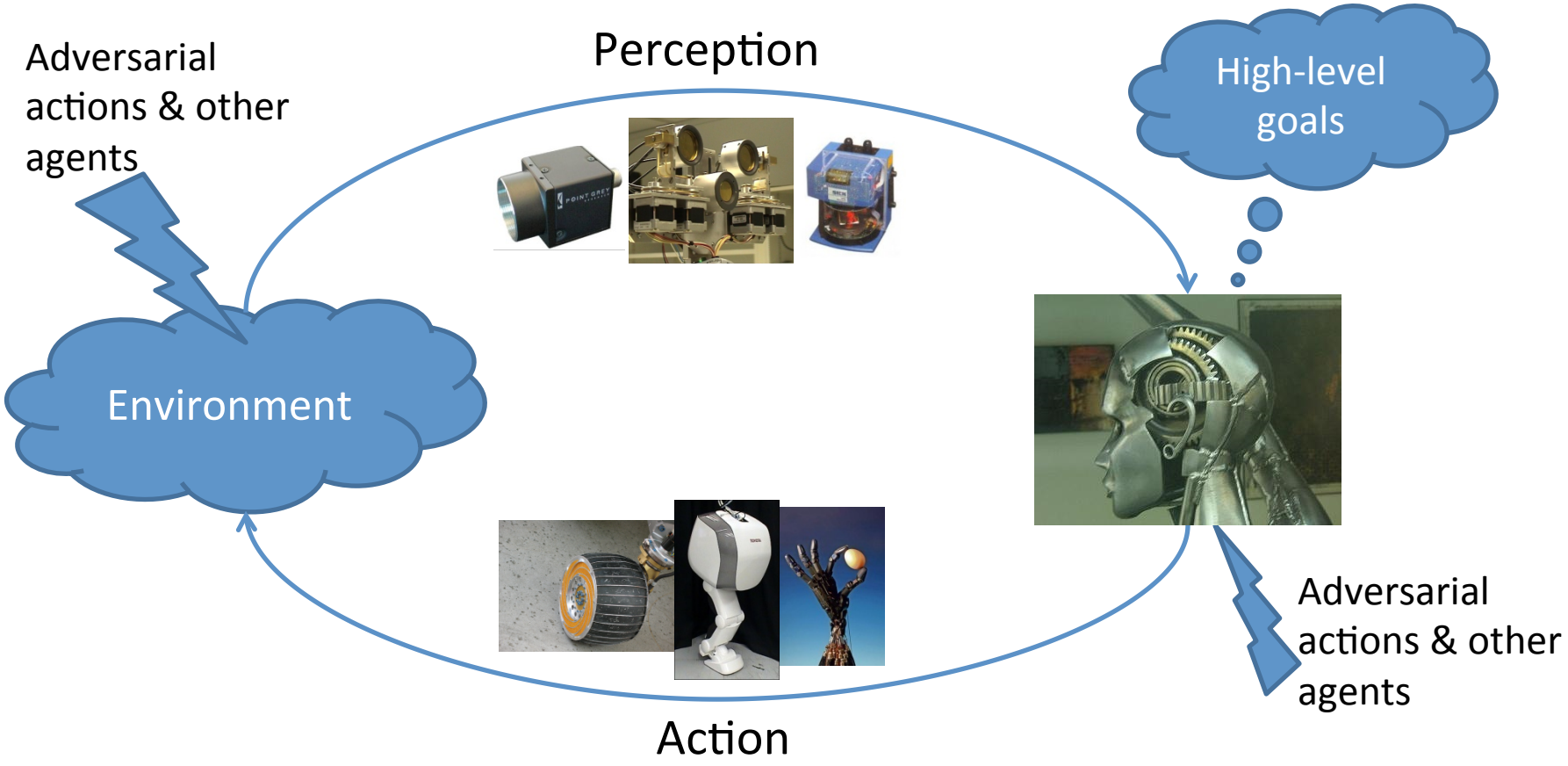
What is the Nature of the Problem?

- As you can tell from the history, many ways to understand the problem – you will see this as we proceed through course

- Unifying perspective:
  - Sequential decision making
  - Stochastic optimization over time

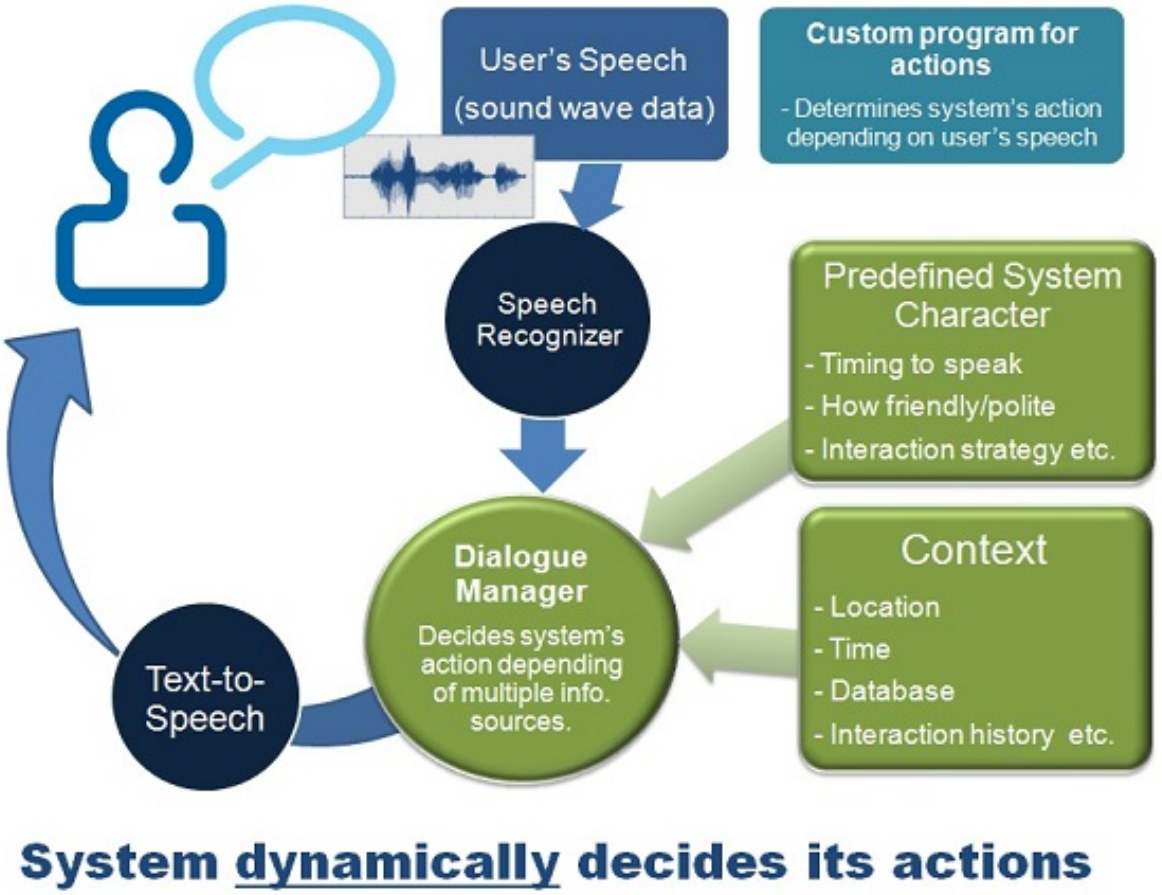
- Let us unpack this through a few application examples…

Example Domain: Robotics



Problem: How to generate actions, to achieve high-level goals, using limited perception and incomplete knowledge of environment?

Example Domain: Natural Language Dialogue



[Source: http://www.agilingua.com/images/DM_EN.jpg]

How to Model Decisions, Computationally?

- Who makes them?
  - Individual
  - ‘Group’
- What are the conditions?
  - Certainty
  - Risk
  - Uncertainty

For most of this course, we’ll take the ‘individual’ viewpoint and we’ll be somewhere in between risk and uncertainty

How to Model Decision under *Certainty*?

- Given a set of possible acts
- Choose one that maximizes some given index

If **a** is a generic act in a set of feasible acts **A**, f(**a**) is an index being maximized, then

Problem: Find **a*** in **A** such that f(**a***) > f(**a**) for all **a** in **A**.

The index f plays a key role, e.g., think of buying a painting.

Essential problem: How should the subject select an index function such that her choice reduces to finding maximizers?

An Operational Way to Find Index Function

- Observe subject's behaviour in restricted settings and predict purchase behaviour from that:
- Instruct the subject as follows:
  - Here are ten valuable reproductions
  - We will present these to you in pairs
  - You will tell us which one of the pair you prefer to own
  - After you have evaluated all pairs, we will pick a pair at random and present you with the choice you previously made (it is to your advantage to remember your true tastes)
- The subject's behaviour is **as though** there is a ranking over all paintings, so each painting can be summarized by a number

Some Desiderata of this Ranking

- *Transitivity*: Previous argument only makes sense if the rank is transitive – if A is preferred in (A, B) and B is preferred in (B, C) then A is preferred in (A, C); and this holds for all triples of alternatives A, B and C
- *Ordinal nature of index*: One is tempted to immediately turn the ranking into a latent measure of 'satisfaction' but that is premature as utilities need not be unique.

  e.g., we could assign 3 utiles to A, 2 utiles to B and 1 utile to C to explain the choice behaviour

  Equally, 30, 20.24 and 3.14 would yield the same choice

While it is OK to compare indices, it isn't (yet) OK to add/multiply

What Happens if we Relax Transitivity?

- Assume Pandora says (in the pairwise comparisons):
  - Apple < Orange
  - Orange < Fig
  - Fig < Apple
- Why is this a problem for Pandora?

- Assume a merchant who transacts with her as follows:
  - Pandora has an Apple at the start of the conversation
  - He offers to exchange Orange for Apple, if she gives him a penny
  - He then offers an exchange of Fig for Orange, at the price of a penny
  - Then, offers Apple for the Fig, for a penny
  - **Now, what is Pandora's net position?**

Decision Making under *Risk*

- Initially appeared as analysis of fair gambles, needed some notions of utility
- Gamble has n outcomes, each worth $a_1, \ldots, a_n$
- The probability of each outcome is $p_1, \ldots, p_n$
- How much is it worth to participate in this gamble?

$$b = a_1 p_1 + \ldots + a_n p_n$$

One may treat this monetary expected value as a fair price

Is this a sufficient description of choice behaviour under risk?

St. Petersburg Paradox of D. Bernoulli

- A fair coin is tossed until a head appears
- Gambler receives 2^n if the first head appears on trial n
- Probability of this event = probability of tail in first (n-1) trials and head on trial n, i.e., $(1/2)^n$

Expected value $= 2.(1/2) + 4.(1/2)^2 + 8.(1/2)^8 + \ldots = \infty$

- Are you willing to bet in this way? Is anyone?

Defining Utility

- Bernoulli went on to argue that people do not act in this way
- The thing to average is the 'intrinsic worth' of the monetary values, not the absolute values

  e.g., intrinsic worth of money may increase with money but at a *diminishing rate*

- Let us say utility of m is $\log_{10} m$, then expected value is,

  $$\log_{10} 2.(1/2) + \log_{10} 4.(1/2)^2 + \log_{10} 8.(1/2)^8 + \ldots = b < \infty$$

  Monetary fair price of the gamble is a where $\log_{10} a = b$.

Some Critiques of Bernoulli's Formulation

von Neumann and Morgenstern (vNM), who initiated the formal study of game theory, raised the following questions:

- The assignment of utility to money is arbitrary and *ad hoc*
  - There are an infinity of functions that capture 'diminishing rate', how should we choose?
  - The association may vary from person to person

- Why is the definition of the decision based upon expected value of the utility?
  - Is this actually descriptive of a single gambler, in one-shot choice?
  - How to define/constrain utility?

von Neumann & Morgenstern Formulation

- If a person is able to express preferences between every possible pair of gambles

  where gambles are taken over some basic set of alternatives

- Then one *can* introduce utility associations to the basic alternatives in such a manner that

- If the person is guided solely by the utility expected value, *he is acting in accord with his true tastes*.
  - provided his tastes are consistent in some way

Constructing Utility Functions

- Suppose we know the following preference order:
  - A < b ~ c < d < e

- The following are utility functions that capture this:

	a	b	c	d	E
U	0	1/2	1/2	3/4	1
V	-1	1	1	2	3
W	-8	0	0	1	8

  - So, in situations like St Petersburg paradox, the revealed preference of any realistic player may differ from the case of infinite expected value
  - Satisfaction at some large value, risk tolerance, time preference, etc.

Certainty Equivalents and Indifference

- The previous statement applies equally well to certain events and gambles or lotteries
- So, even attitudes regarding tradeoffs between the two ought to be captured
- Basic issue – how to compare?
- Imagine the following choice (A > B > C pref.) : (a) you get B for certain, (b) you get A with probability p and C otherwise
- If p is near 1, option b is better; if p is near 0, then option a: there is a single point where we switch
- Indifference is described as something like

$$(2/3)\,(1) + (1 - 2/3)\,(0) = 2/3$$

A C B

Decision Making under *Uncertainty*

- A choice must be made from among a set of acts, $A_1, \ldots, A_m$.
- The relative desirability of these acts depends on which state of nature prevails, either $s_1, \ldots, s_n$.
- As decision maker we know that one of several things is true and this influences our choice

- **Key point about decision making: whether or not you have a probabilistic characterization of alternatives has a big impact on how to approach the problem**

Example: Savage's Omelet Problem

Your friend has broken 5 good eggs into a bowl when you come in to volunteer and finish the omelet.

A sixth egg lies unbroken (you must use it or waste it altogether).

Your three acts: break it into bowl, break it into saucer – inspect and pour into bowl, throw it uninspected

Decision in Savage's Omelet Problem

Table 1. Savage's example illustrating acts, states, and consequences

Act	State	
	Good	Rotten
Break into bowl	six-egg omelet	no omelet, and five good eggs destroyed
Break into saucer	six-egg omelet, and a saucer to wash	five-egg omelet, and a saucer to wash
Throw away	five-egg omelet, and one good egg destroyed	five-egg omelet

- To each outcome, we could assign a utility and maximize it
- What do we know about the state of nature?
  - We may act *as though* there is one true state, we just don't know it
  - If we assume a probability over s_i, this is *decision under risk*
  - If we do not assume a probability over s_i, what might one do?