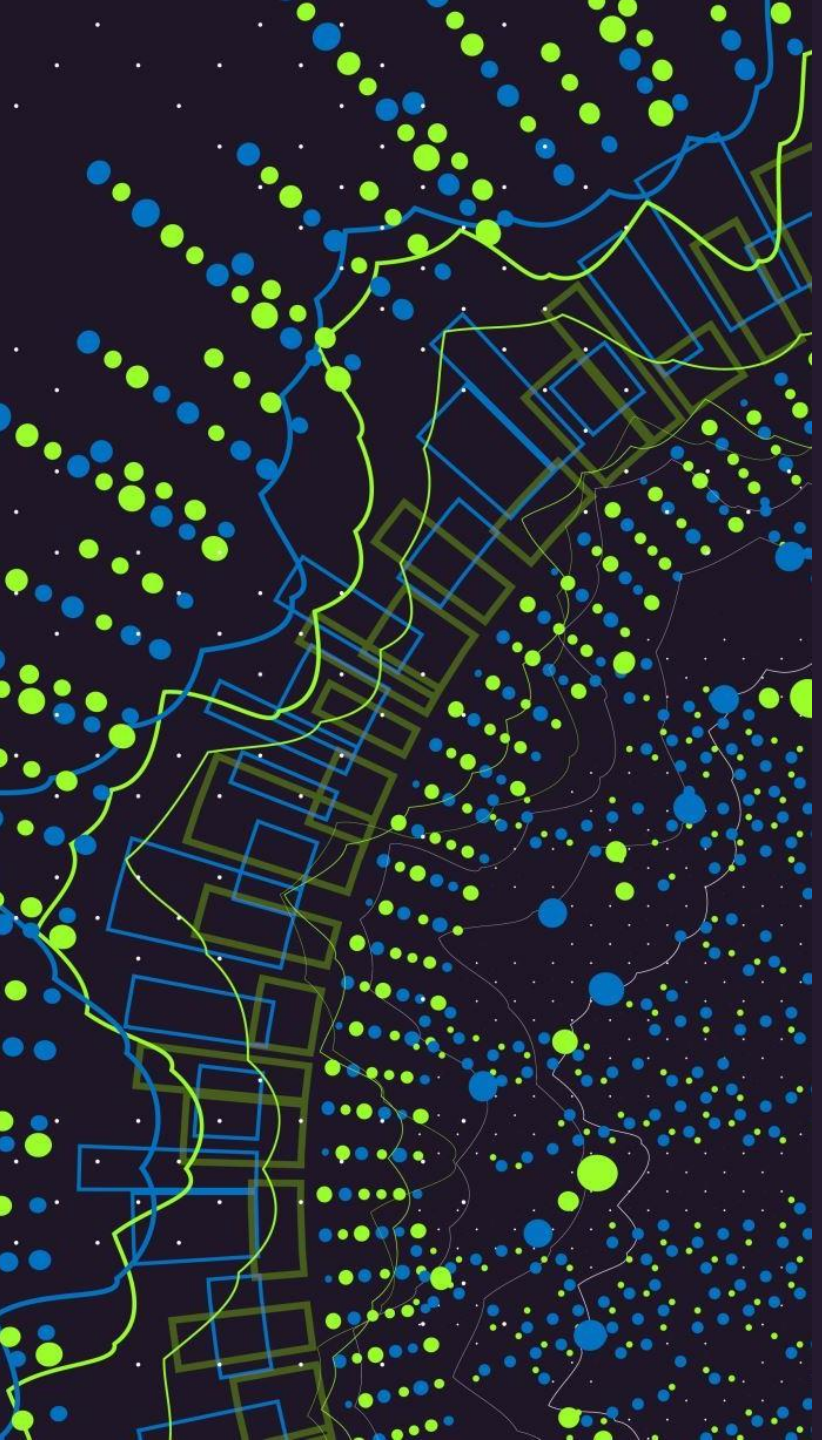


REINFORCEMENT LEARNING:
AN INTRODUCTION 2ED
SUTTON & BARTO. (2018).

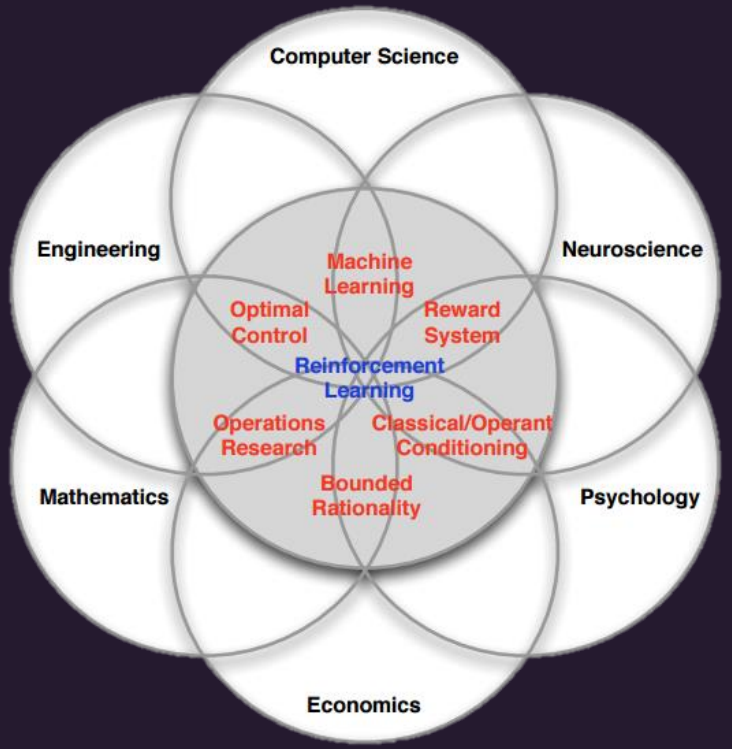
Introduction

Andrés F. Higuera

AGENDA

1. Reinforcement Learning
2. Examples
3. Elements of Reinforcement Learning
4. Limitations and Scope
5. An Extended Example: Tic-Tac-Toe

REINFORCEMENT LEARNING



Many Face of Reinforcement Learning (by David Silver)

→ Decision-making

REINFORCEMENT LEARNING

Rather than directly theorizing about how people or animals learn, we primarily explore idealized learning situations and evaluate the effectiveness of various learning methods.

We explore designs for machines that are effective in solving learning problems of scientific or economic interest, evaluating the designs through mathematical analysis or computational experiments.

Reinforcement Learning (image by Flat-Icons on IconScout under license to Chris Mahoney)

REINFORCEMENT LEARNING

All reinforcement learning agents have **explicit goals**, can **sense** aspects of **their environments**, and can **choose actions** to influence their environments. Moreover, it is usually assumed from the beginning that the agent has to operate despite significant uncertainty about the environment it faces.

Reinforcement Learning (image by Flat-Icons on IconScout under license to Chris Mahoney)

REINFORCEMENT LEARNING

Supervised learning is learning from a training set of labeled examples provided by a knowledgable external supervisor. Each example is a description of a situation together with a specification—the label—of the correct action the system should take to that situation, which is often to identify a category to which the situation belongs. The object of this kind of learning is for the system to extrapolate, or generalize, its responses so that it acts correctly in situations not present in the training set.

Unsupervised learning, which is typically about finding structure hidden in collections of unlabeled data.

We therefore consider *reinforcement learning* to be a third machine learning paradigm, alongside supervised learning and unsupervised learning and perhaps other paradigms.

EXAMPLES

- A master **chess** player makes a move.
- An adaptive controller adjusts parameters of a **petroleum refinery**'s operation in real time.
- A **gazelle** calf struggles to its feet minutes after being born. Half an hour later it is running at 20 miles per hour.
- A mobile **robot** decides whether it should enter a new room in search of more trash to collect or start trying to find its way back to its battery recharging station.
- Phil **prepares his breakfast**. Closely examined, even this apparently mundane activity reveals a complex web of conditional behavior and interlocking goal–subgoal relationships.

EXAMPLES

All involve *interaction* between an active decision-making agent and its environment, within which the *agent seeks to achieve a goal despite uncertainty* about its environment. The agent's actions are permitted to affect the future state of the environment (e.g., the next chess position, the level of reservoirs of the refinery, the robot's next location and the future charge level of its battery), thereby affecting the actions and opportunities available to the agent at later times. Correct choice requires taking into account indirect, delayed consequences of actions, and thus may *require foresight or planning*.

The knowledge the agent brings to the task at the start—either from previous experience with related tasks or built into it by design or evolution—influences what is useful or easy to learn, but *interaction with the environment is essential for adjusting behavior to exploit specific features of the task*.

ELEMENTS OF REINFORCEMENT LEARNING

Policy ~ defines the learning agent's way of behaving at a given time. Roughly speaking, a policy is a **mapping from perceived states of the environment to actions** to be taken when in those states (Psychology: stimulus-response).

Reward Signal ~ defines what are the good and bad events for the agent (Biology: experiences of pleasure or pain). **The reward signal is the primary basis for altering the policy.**

Value Function ~ the value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state. Whereas rewards determine the immediate, intrinsic desirability of environmental states, values **indicate the long-term desirability of states** after taking into account the states that are likely to follow and the rewards available in those states (farsighted judgment).

Model of the environment (optional) ~ This is something that mimics the behavior of the environment, or more generally, that allows inferences to be made about **how the environment will behave**.

LIMITATIONS AND SCOPE

We assume that the state signal is produced by some preprocessing system that is nominally part of the agent's environment. We do not address the issues of constructing, changing, or learning the state signal in this book. We take this approach not because we consider state representation to be unimportant, but in order to focus fully on the decision-making issues. In other words, our concern in this book is not with designing the state signal, but with deciding what action to take as a function of whatever state signal is available.

Most of the reinforcement learning methods we consider in this book are structured around estimating value functions, but it is not strictly necessary to do this to solve reinforcement learning problems.

AN EXTENDED EXAMPLE: TIC-TAC-TOE

How might we construct a player that will find the imperfections in its opponent's play and learn to maximize its chances of winning?

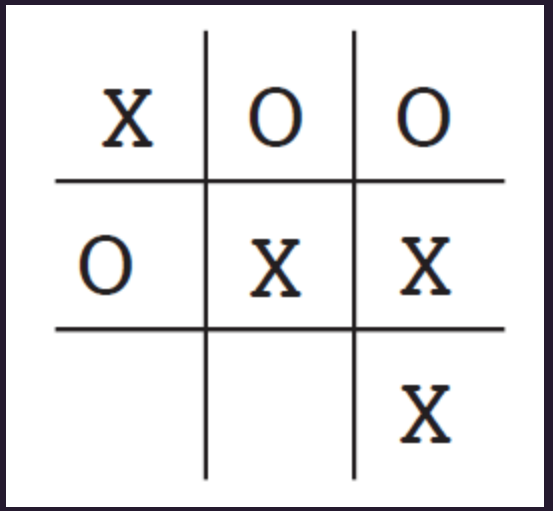
1. We would set up a table of numbers, one for each possible state of the game. Each number will be the latest estimate of the probability of our winning from that state.

   Assuming we always play Xs, then for all states with three Xs in a row the probability of winning is 1, because we have already won. Similarly, for all states with three Os in a row, or that are filled up, the correct probability is 0, as we *cannot win* from them. We set the initial values of all the other states to 0.5, representing a guess that we have a 50% chance of winning.

2. We then play many games against the opponent. To select our moves we examine the states that would result from each of our possible moves and look up their current values in the table. (Exploration vs Explotation).

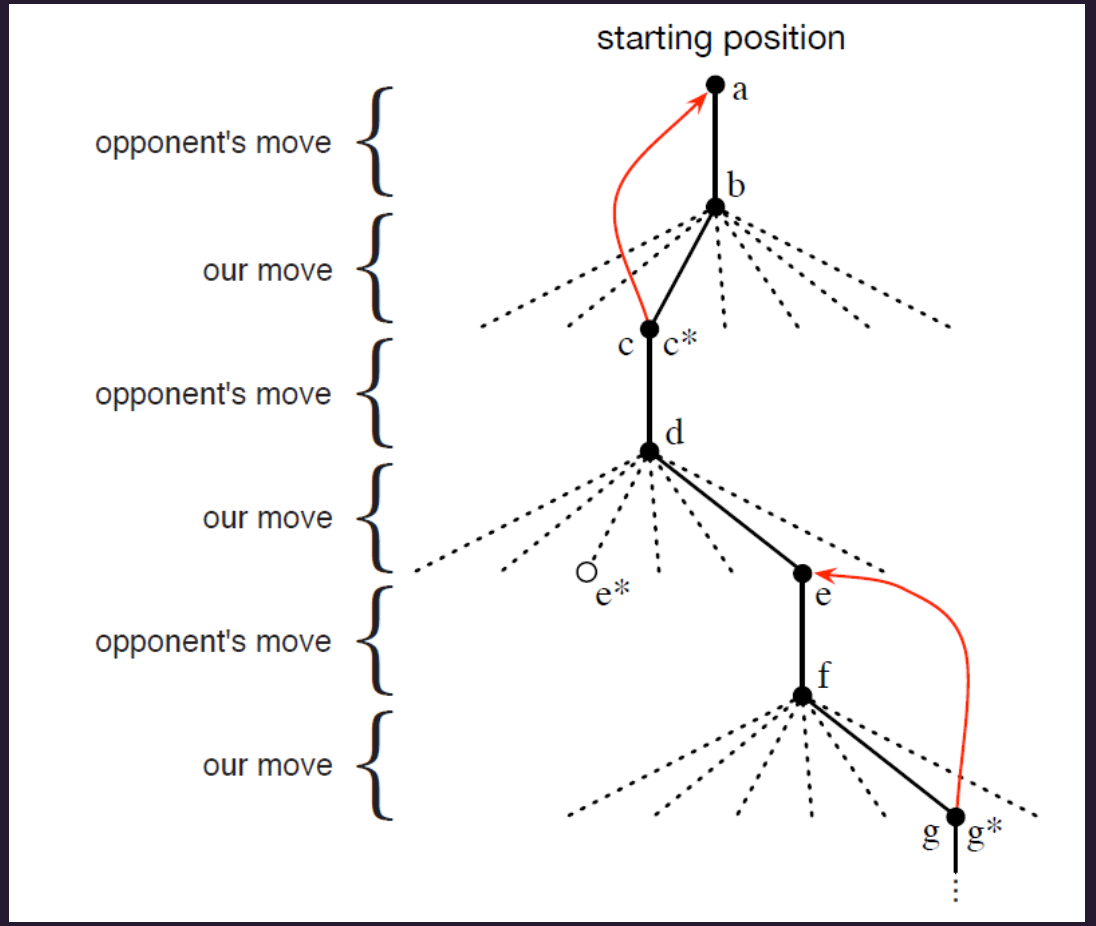
3. While we are playing, we change the values of the states in which we find ourselves during the game. We attempt to make them more accurate estimates of the probabilities of winning:

$$V(S_t) \leftarrow V(S_t) + \alpha[V(S_{t+1}) - V(S_t)]$$ (*temporal-difference*)



AN EXTENDED EXAMPLE: TIC-TAC-TOE

The solid black lines represent the moves taken during a game; the dashed lines represent moves that we (our reinforcement learning player) considered but did not make. Our second move was an exploratory move, meaning that it was taken even though another sibling move, the one leading to e*, was ranked higher. Exploratory moves do not result in any learning, but each of our other moves does, causing updates as suggested by the red arrows in which estimated values are moved up the tree from later nodes to earlier nodes as detailed in the text.



A sequence of tic-tac-toe moves

AN EXTENDED EXAMPLE: TIC-TAC-TOE

This simple example illustrates some of the key features of reinforcement learning methods. First, there is the emphasis on learning while interacting with an environment, in this case with an opponent player. Second, there is a clear goal, and correct behavior requires planning or foresight that takes into account delayed effects of one's choices. For example, the simple reinforcement learning player would learn to set up multi-move traps for a shortsighted opponent. It is a striking feature of the reinforcement learning solution that it can achieve the effects of planning and lookahead without using a model of the opponent and without conducting an explicit search over possible sequences of future states and actions.