

Lecture 4

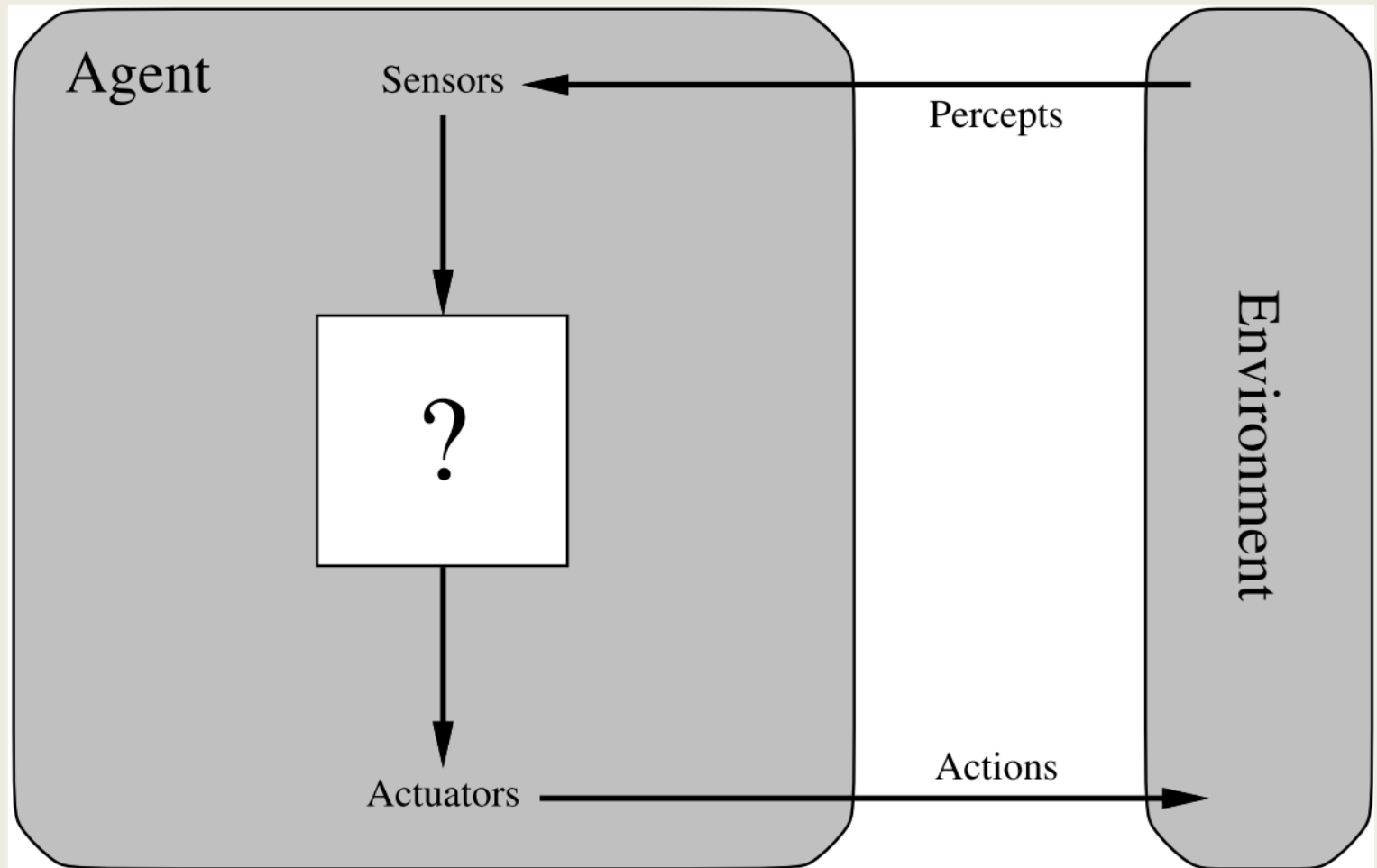
Statistical Artificial Intelligence

Michail G. Lagoudakis

**Intelligent Systems Laboratory
School of Electrical and Computer Engineering
Technical University of Crete
Chania, Crete, Greece**



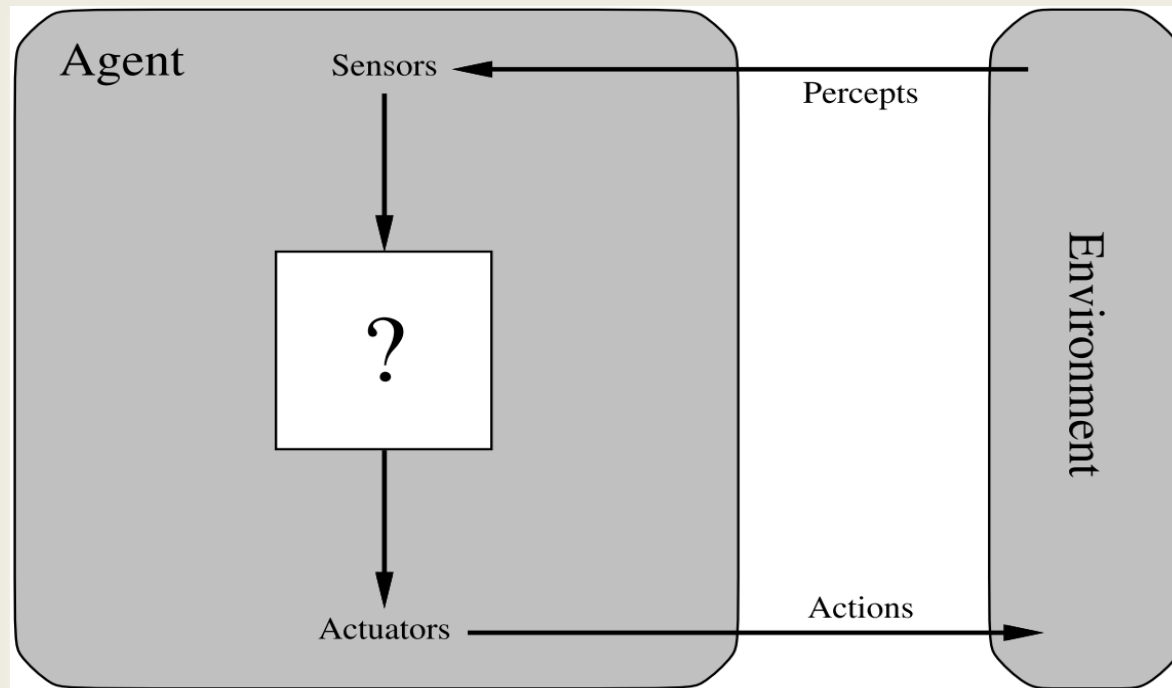
The Intelligent Autonomous Agent Paradigm



The Intelligent Autonomous Agent Paradigm

Symbolic AI (GOOF-AI)

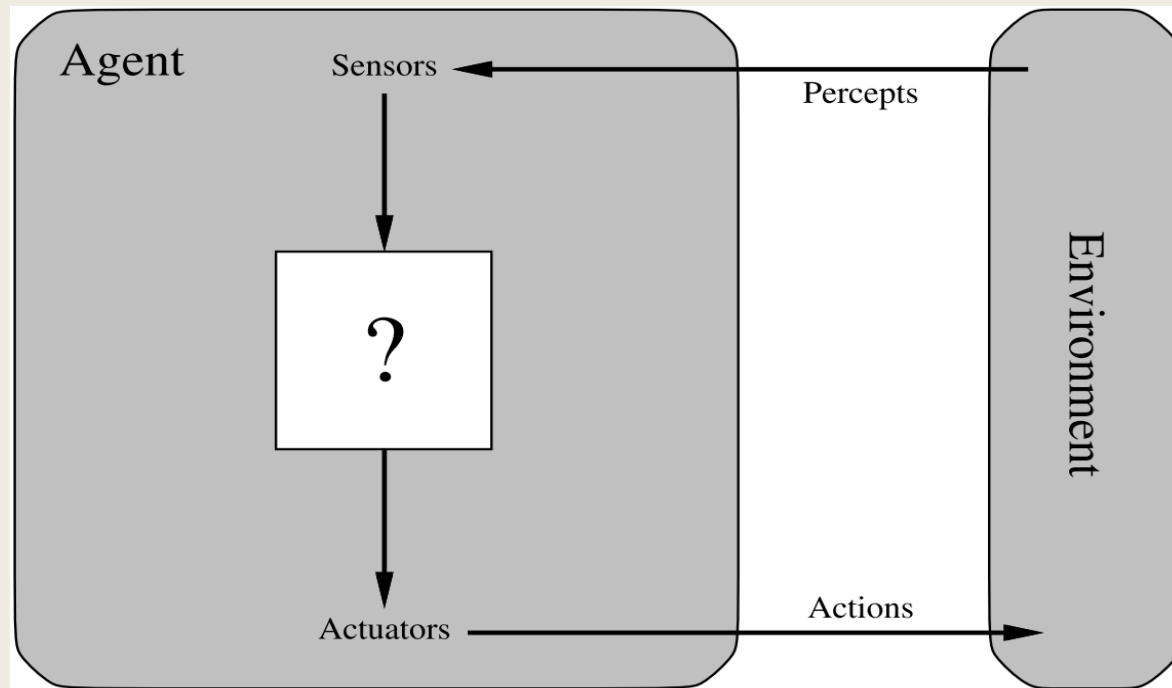
- since the very beginning in the 1960s
- everything can be represented using propositions and truth values
- reasoning is simply a procedure of logical inference
- ... using axioms of logic, logical inference, derivation of proofs, ...
- problems: huge needs of resources, no handling of uncertainty, ...



The Intelligent Autonomous Agent Paradigm

Statistical AI

- 1990s: let's embrace uncertainty + let's make AI work in the real world
- probability theory, utility theory, decision theory, information theory, statistics, numerical optimization, ...
- celebrated “tools”: Bayesian Networks, MDPs, HMMs, CNNs, LLMs, ...
- ...but also game theory, auctions, mechanisms for multi-agent systems



Bayesian Probabilities

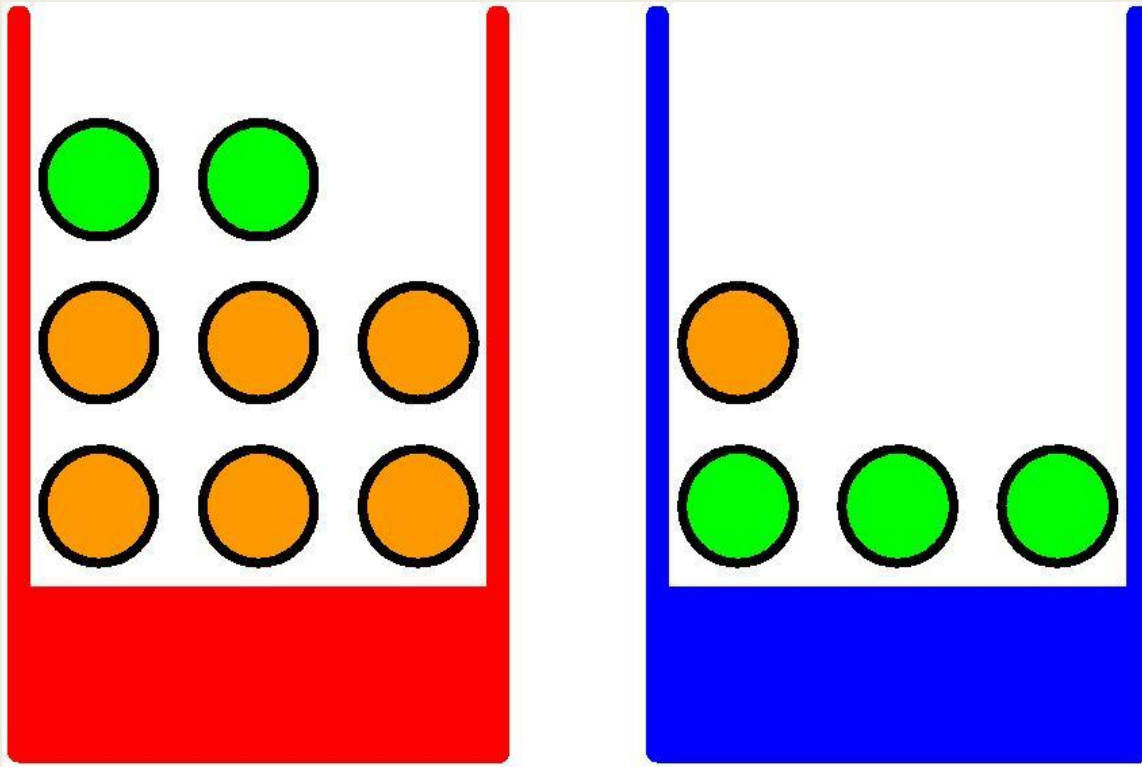
Bayes Rule

Bayesian Approach



Probability Theory

Apples and Oranges



Probability Theory

- **Procedure**

- choose a box: red (40%) or blue (60%)
- pick up a fruit: randomly (equally likely)
- replace the fruit and repeat

- **Random Variables**

- B : r or b , red or blue box
- F : a or o , apple or orange fruit
- probability of event: fraction of times it occurs in the limit

- **Queries**

- what is the probability of picking an apple?
- what is the probability that the box was red, given an apple?
- ...



Frequentist Probability

- **Frequentist probability** is the interpretation of probability that defines an event's probability as the limit of its relative frequency in a large number of trials.



Bayesian Probability

- **Bayesian probability** is the interpretation of the probability that supports the idea that probability can be defined as the degree to which a person (or community) believes that a proposition is true.



Bayesian vs Frequentist Probability

Both agree with the statement that $P(\text{heads})=1/2$.



- The **Bayesian** would say that given there are two sides to the coin, and there is no reason **to believe** that one side is more favored than the other (or if one is favored, there is no way to know which side is favored), then I am completely **uncertain** as to which side will fall, so the probability of it coming up heads must be $1/2$.
- The **frequentist** would say that if this is a fair coin and it would be flipped a very large number of times, then **the long term ratio of frequencies** of heads and tails is $1/2$.



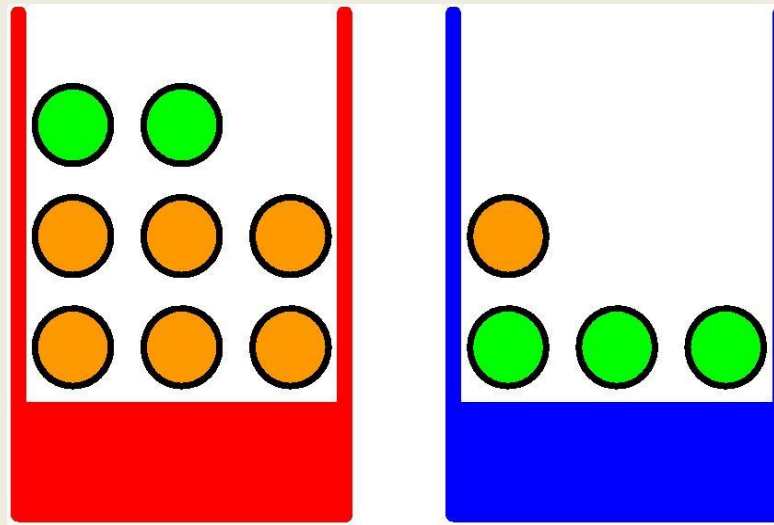
Towards Bayesian Probability

- **Frequentist probability**
 - frequencies of random, repeatable events
- **Considerations**
 - can we reason about rare, non-repeatable events?
 - example: will the Arctic ice cap disappear in 1000 years?
 - what if we have some idea about the rate of ice melting?
 - what if this knowledge is revised after 100 years?
 - need to quantify uncertainty
 - need to make revisions in light of new evidence
- **Bayesian probability**
 - quantification and revision of uncertainty



Bayesian Thinking

Apples and Oranges



- blue box most likely chosen, before looking at fruit (prior)
- red box most likely chosen, after picking an orange (posterior)

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$



Probability Theory as Extended Logic

- Framework for **plausible reasoning** developed by three physicists:
 - Harold Jeffreys (1891 – 1989)
 - Richard Threlkeld Cox (1898 - 1991)
 - Edwin Thompson Jaynes (1922 – 1998)



Probability Theory as Extended Logic

1. Degrees of **plausibility** are represented by real numbers.
2. Qualitative correspondence with common sense:
 - Increasing the **plausibility** of a statement, decreases the **plausibility** of its negative.
3. Consistency:
 - If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.
 - It must be possible to incorporate all the evidence relevant to a question into the reasoning process.
 - All equivalent states of knowledge are represented by equivalent **plausibility** assignments.



Probability Theory as Extended Logic

- These simple statements (known as Cox's Theorems) mean that the mathematical description of **plausible reasoning** is completely determined.
 - It turns out to be exactly the same as that of standard probability theory
 - Probability theory is a general form of logic where statements are not just true or false, but have degrees of certainty.
 - ...therefore, *use probability to measure beliefs*



Bayes Theorem

- **Rev. Thomas Bayes (1702 – 1761)**
 - a British mathematician and Presbyterian minister



Bunhill Fields Cemetery in London



Bayes Theorem

Likelihood
(the Bayes factor)

Prior

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

Posterior

Evidence



Bayes Theorem: Derivation

From conditional probabilities ...

$$P(X | Y) = P(X, Y) / P(Y)$$

$$P(Y | X) = P(X, Y) / P(X)$$

... we get:

$$P(X, Y) = P(X | Y) P(Y) = P(Y | X) P(X)$$

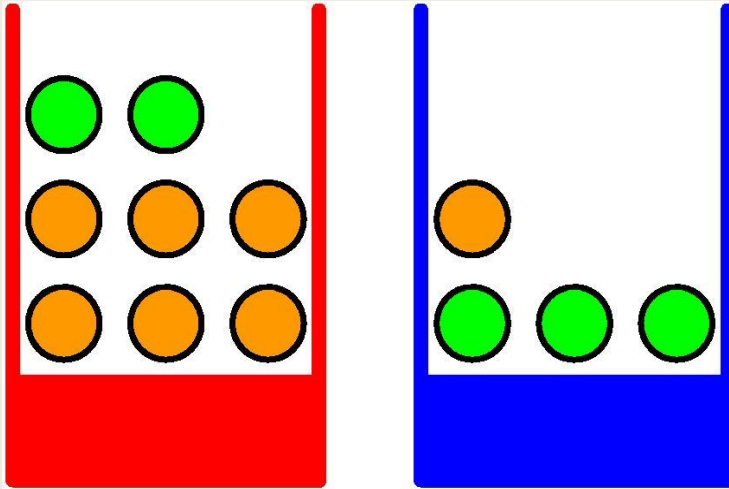
(this is known as the *product* or *chain rule*)

... and then, rearranging:

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$



Probability Examples



$$p(B = r) = 4/10 \quad p(B = b) = 6/10$$

$$p(F = a|B = r) = 1/4$$

$$p(F = o|B = r) = 3/4$$

$$p(F = a|B = b) = 3/4$$

$$p(F = o|B = b) = 1/4$$

$$p(F = a) = p(F = a|B = r)p(B = r) + p(F = a|B = b)p(B = b) = 11/20$$

$$p(F = o) = p(F = o|B = r)p(B = r) + p(F = o|B = b)p(B = b) = 9/20$$

$$p(B = b|F = a) = \frac{p(F = a|B = b)p(B = b)}{p(F = a)} = \frac{(3/4)(6/10)}{11/20} = \frac{9}{11}$$



Why is this useful?

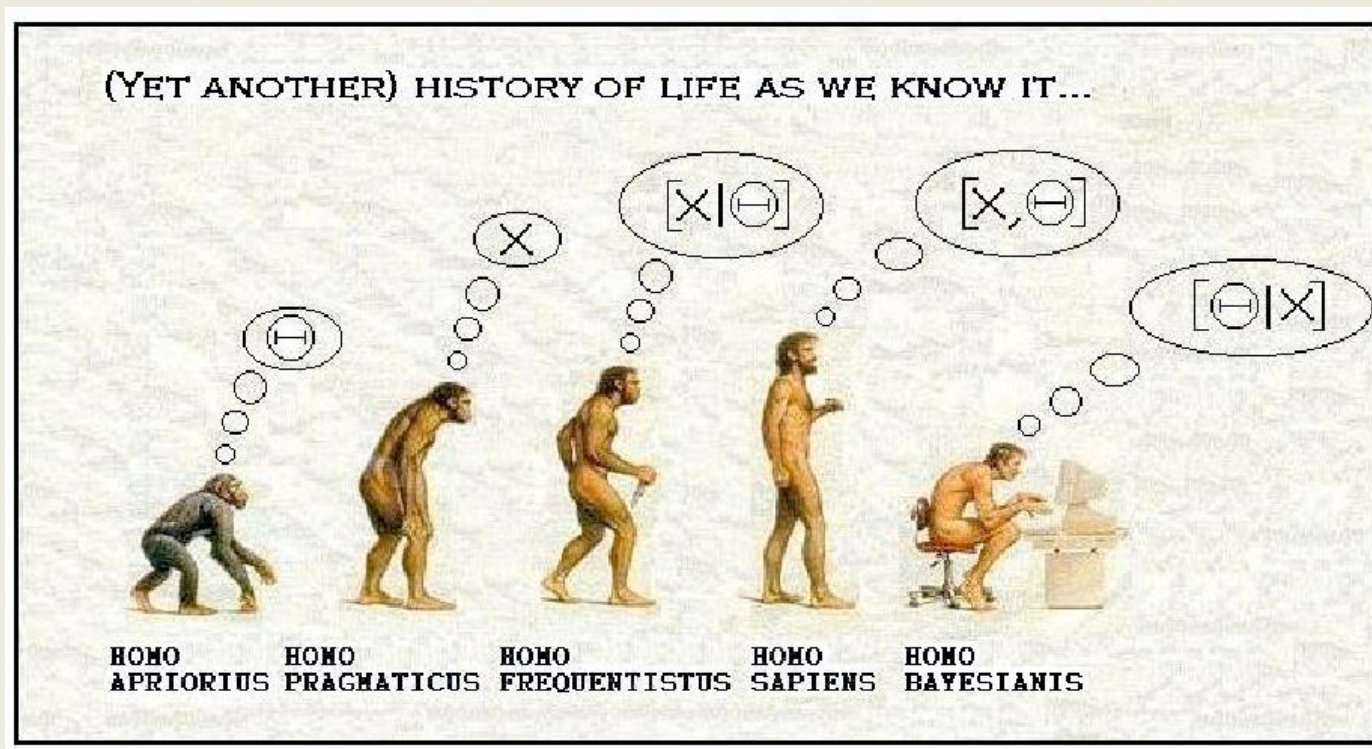
$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

- Because usually $P(X | Y)$
 - is easier to obtain ... **X is data (evidence / effects / symptoms)**
 - provided **given an assumption (hypothesis / cause)**
- In many cases we have info on: $P(\text{effects} | \text{cause})$
- but we really care to find this: $P(\text{cause} | \text{effects})$
- Bayes rule turns a diagnostic probability to a causal one!

$$P(\text{cause} | \text{effects}) = \frac{P(\text{effects} | \text{cause})P(\text{cause})}{P(\text{effects})}$$



The Bayesian Approach



- The Homo Apriorius establishes the probability of a hypothesis, no matter what data says.
- The Homo Pragmaticus is interested in the probability of the data only.
- The Homo Frequentist measures the probability of the data, given the hypothesis.
- The Homo Sapiens is interested in the probability of the data and of the hypothesis.
- The Homo Bayesian measures the probability of the hypothesis, given the data.

with thanks to M. West and to S. Andreon

Frequentists vs. Bayesians

- **Frequentist approach**

- measures the probability of the data, given the hypothesis
- predictions may be misled (e.g. three coin tosses, all heads)
- being somewhat objective (just believe the data)
- parameters are fixed and unknown constants (point estimates)
- uncertainty is quantified by confidence intervals
- predictions using point estimates of the parameters

- **Bayesian approach**

- measures the probability of the hypothesis, given the data
- makes smoother predictions (e.g. three coin tosses, all heads)
- being somewhat subjective (dependence on prior belief)
- parameters are described as probability distributions
- uncertainty is directly included in the distribution
- predictions by marginalization over the space of parameters



Example 1



- Suppose that Bob can decide to go to work by one of three modes of transportation, car, bus, or train.
 - Because of high traffic, if he decides to go by car, there is a 50% chance he will be late.
 - If he goes by bus, which has special reserved lanes but is sometimes overcrowded, the probability of being late is only 20%.
 - The train is almost never late, with a probability of only 1%, but is more expensive than the bus.



Example 1

- Suppose that Bob is late one day, and his boss wishes to estimate the probability that he drove to work that day by car.
- Since he does not know which mode of transportation Bob usually uses, he gives a prior probability of 1 of 3 to each of the three possibilities.
- What is the boss' estimate of the probability that Bob drove to work?



Example 1

- First list all the facts available:

$$P(\text{car}) = \frac{1}{3}$$

$$P(\text{late}|\text{car}) = \frac{1}{2}$$

$$P(\text{bus}) = \frac{1}{3}$$

$$P(\text{late}|\text{bus}) = \frac{1}{5}$$

$$P(\text{train}) = \frac{1}{3}$$

$$P(\text{late}|\text{train}) = \frac{1}{100}$$



Example 1

- Then apply Bayes theorem:

$$P(\text{car}|\text{late}) = \frac{P(\text{late}|\text{car})P(\text{car})}{P(\text{late})}$$

$$= \frac{P(\text{late}|\text{car})P(\text{car})}{P(\text{late}|\text{car})P(\text{car}) + P(\text{late}|\text{bus})P(\text{bus}) + P(\text{late}|\text{train})P(\text{train})}$$

$$= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + \frac{1}{5} \times \frac{1}{3} + \frac{1}{100} \times \frac{1}{3}}$$

$$\approx 0.704$$

P(evidence) is given by the Law of Total Probability:
a “weighted average” that captures the overall probability of an outcome that can be realized via several distinct events



Example 2

- A clinical test, designed to diagnose a specific illness, comes out positive for a certain patient.
- We are told that:
 - The test is 79 percent reliable
(that is, it misses 21 percent of actual cases)
 - On average, this illness affects 1 percent of the population in the same age group as the patient
 - The test has a false positive rate of 10 percent.
- Taking this into account and assuming you know nothing about the patient's symptoms or signs, what is the probability that this patient actually has the illness?



Example 2

- First list all the facts available:

$$P(D) = 0.01$$

$$P(T|D) = 0.79$$

$$P(T|\neg D) = 0.1$$



Example 2

- Then apply Bayes theorem:

$$\begin{aligned}P(D|T) &= \frac{P(T|D)P(D)}{P(T)} \\&= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\neg D)P(\neg D)} \\&= \frac{0.79 \times 0.01}{0.79 \times 0.01 + 0.1 \times 0.99} \\&\approx 0.074\end{aligned}$$

A good argument against
random screening if the test
does not discriminate very well!



Optimal Learning: Overview

- Behavior that maximizes expected total reward, while interacting with an uncertain world.
- Behave well while learning, learn while behaving well.
- Optimal learning $\Leftrightarrow$ using a Bayesian approach



Optimal Learning: Bayesian Approach

Bayesian approach

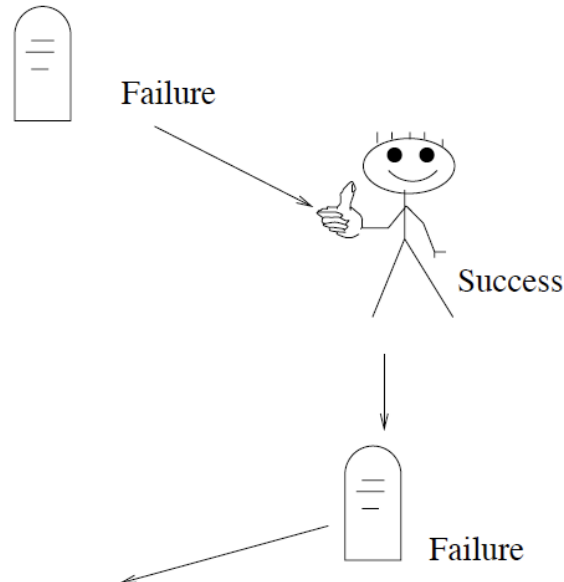
- Evolve uncertainty about *unknown process parameters* of the world model (e.g., transitions, rewards, ...)
- Uncertainty over the parameters is captured by *prior distributions*.
- The values of the (hyper-) parameters of these priors correspond to *information states* (describing our uncertainty about the world)
- Using the prior, the likelihood of observed data, and Bayes rule ...
- ... the belief is updated and a *posterior distribution* is formed,
- ... which serves as the new prior distribution in the next step!
- ... still capturing our (reduced) uncertainty about the world!



A simple example

$$\theta_1 = \Pr\{\text{Cure} \mid \text{Drug 1}\}$$

$$\theta_2 = \Pr\{\text{Cure} \mid \text{Drug 2}\}$$



example due to
M. O' Gordon Duff (2002)

- Bernoulli process parameters θ_1 , θ_2 describe the actual (but unknown) probabilities of success
- **Bayesian approach:** Uncertainty about parameters?
describe it by conjugate prior distributions



Conjugate Priors

- A prior is **conjugate** to a likelihood function, if the posterior is in the same family with the prior.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

- Prior in the family $\Rightarrow$ posterior in the family
- A simple update of the **hyper-parameters** is typically enough to get the posterior!

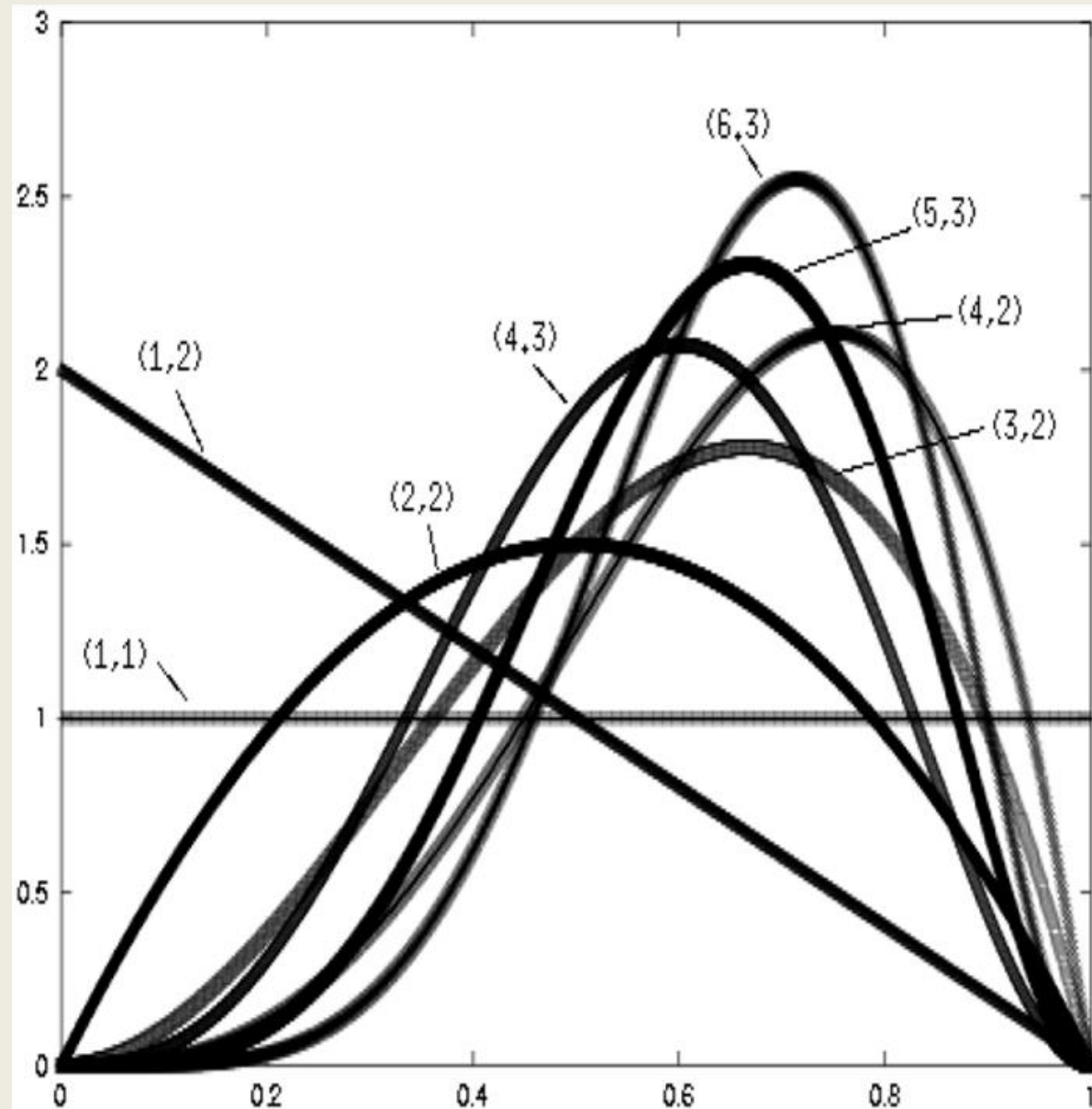
- **Remark:** The uncertainty over the parameters is captured by these *prior-turned-posterior* distributions.



Distribution Beta(α, β)

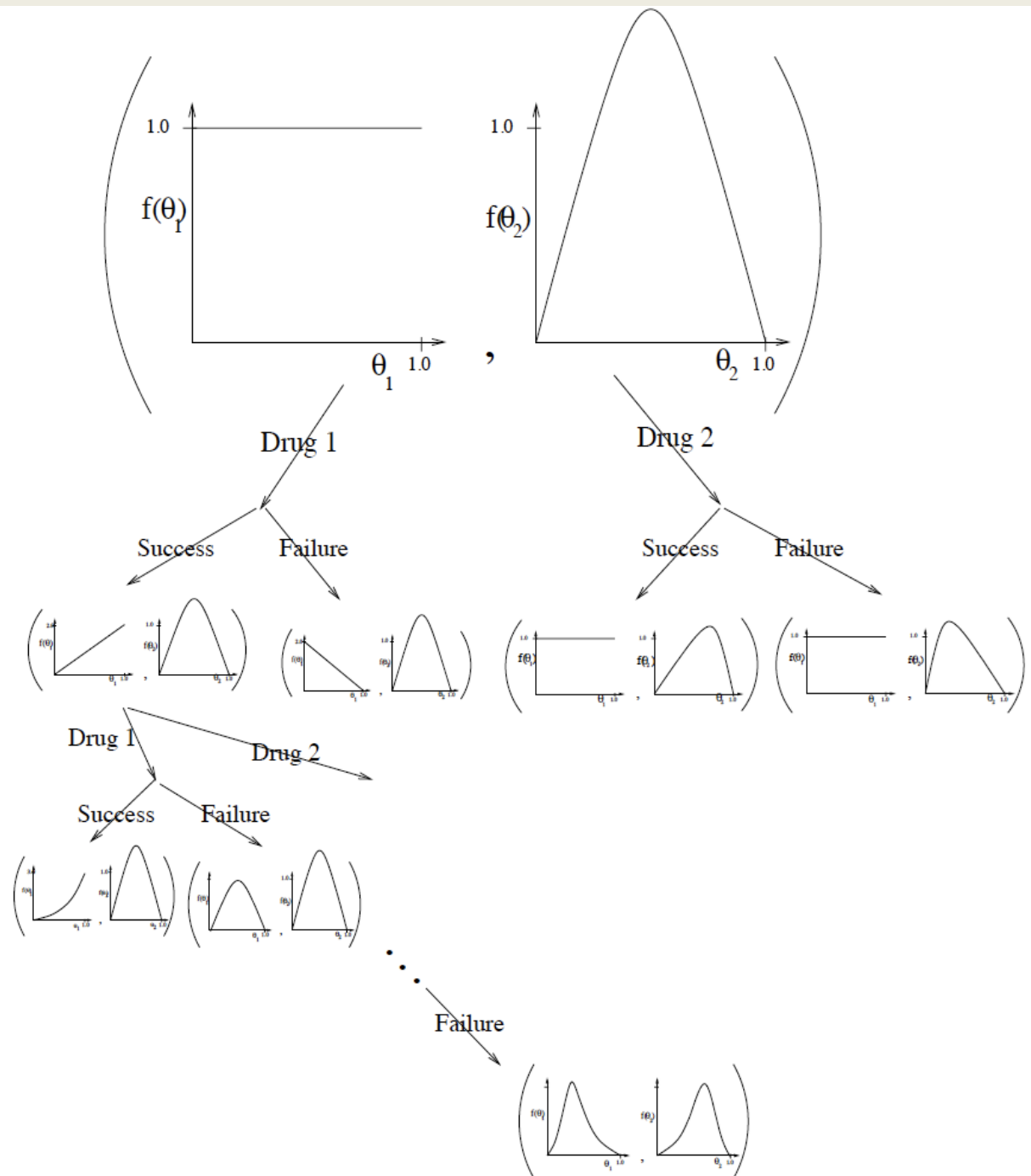
$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

$$\text{where } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$



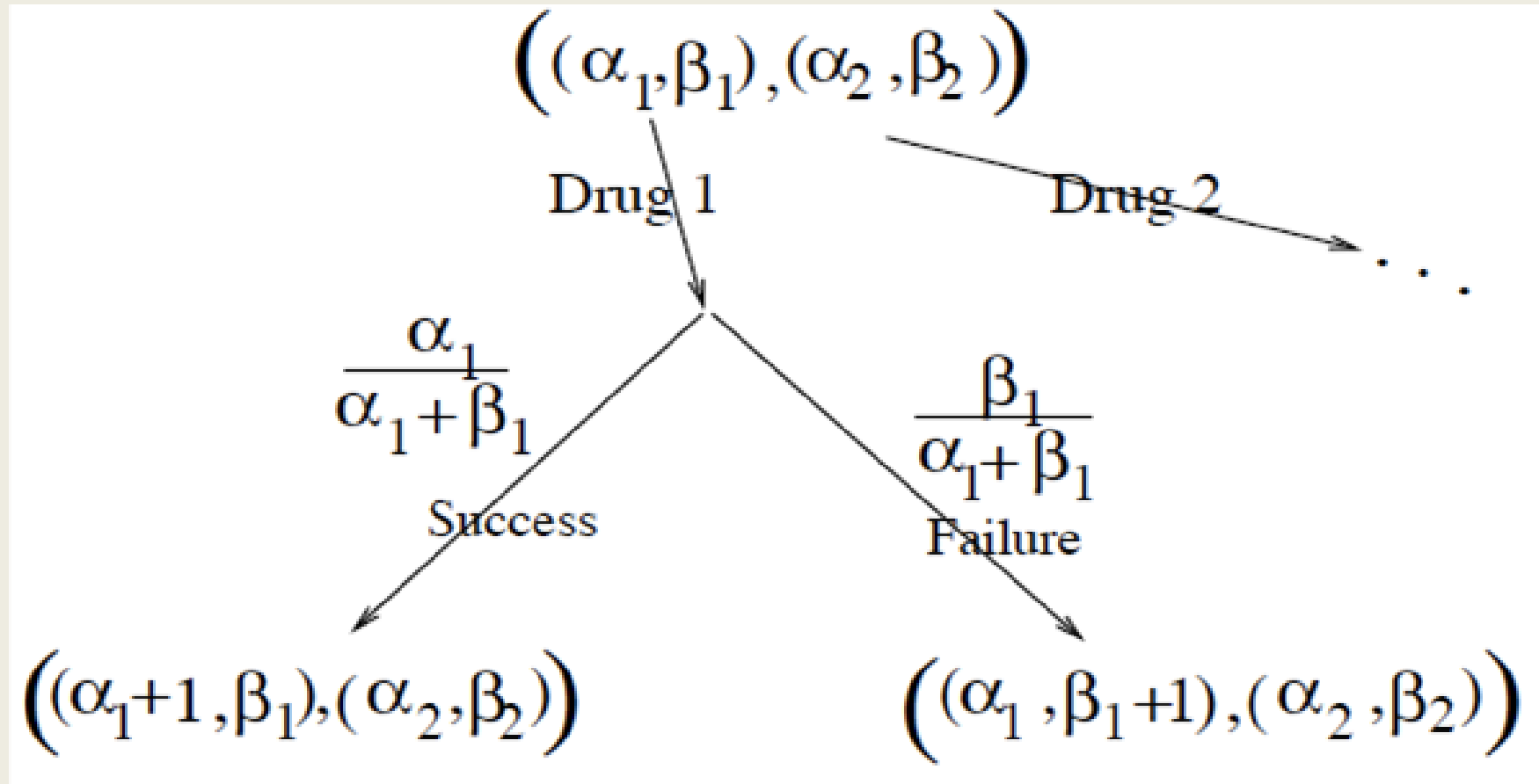
example due to
M. O' Gordon Duff (2002)

Information- State Transition Diagram



example due to
M. O' Gordon Duff (2002)

For Beta distribution, it simply becomes



example due to
M. O' Gordon Duff (2002)



Acting under Uncertainty



What should an agent do when operating under uncertainty? Decision theory to the rescue

Decision Theory = Utility Theory + Probability Theory

```
function DT-AGENT(percept) returns an action
  persistent: belief_state, probabilistic beliefs about the current state of the world
              action, the agent's action

  update belief_state based on action and percept
  calculate outcome probabilities for actions,
    given action descriptions and current belief_state
  select action with highest expected utility
    given probabilities of outcomes and utility information
  return action
```



A Definition of Rationality

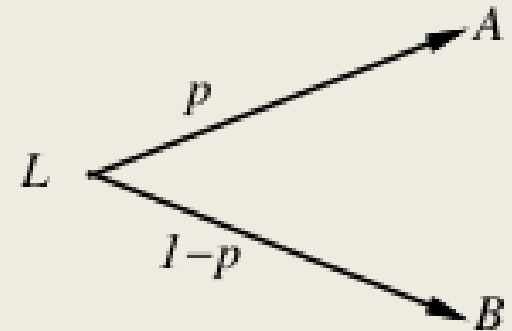
An agent is rational *if-f* it picks its actions adhering to the **maximum expected utility principle**:

*when choosing among (stochastic) actions,
always pick the action with maximum expected utility.*



Justification: Lotteries and Stochastic Actions

- A standard lottery L leads to a prize A with probability p , and to a prize B with probability $1-p$.



- Generally: $L = \{(p_1, X_1), (p_2, X_2), \dots, (p_n, X_n)\}$
- If an outcome X_i has utility $U(X_i)$...
- ... then Lottery L has expected utility $U(L) = \sum p_i U(X_i)$
- Due to the Ramsey-von Neumann and Morgenstern expected utility theorem, you can construct a real-valued utility function that corresponds to a lottery's expected utility, *such that:*

$$U(L1) \geq U(L2) \text{ if-} f \text{ } L1 \text{ preferable than } L2$$

- Thus, if $U(L1) > U(L2)$, which lottery would you prefer?**
- Now, think of a lottery as an action with uncertain outcomes
- Hence, the *Maximum Expected Utility Principle*



Inference and Decision

- **Inference**

- input $\mathbf{x}$, target output $\mathbf{t}$
- determine either $p(\mathbf{x}, \mathbf{t})$ or $p(\mathbf{t} | \mathbf{x})$
- *example*: probability of cancer ($\mathbf{t}$), given an x-ray image ($\mathbf{x}$)

- **Decision**

- for given $\mathbf{x}$, determine optimal decision based on predicted $\mathbf{t}$
- *example*: give cancer treatment, given an x-ray image
- the decision step is easy, if we have solved the inference step

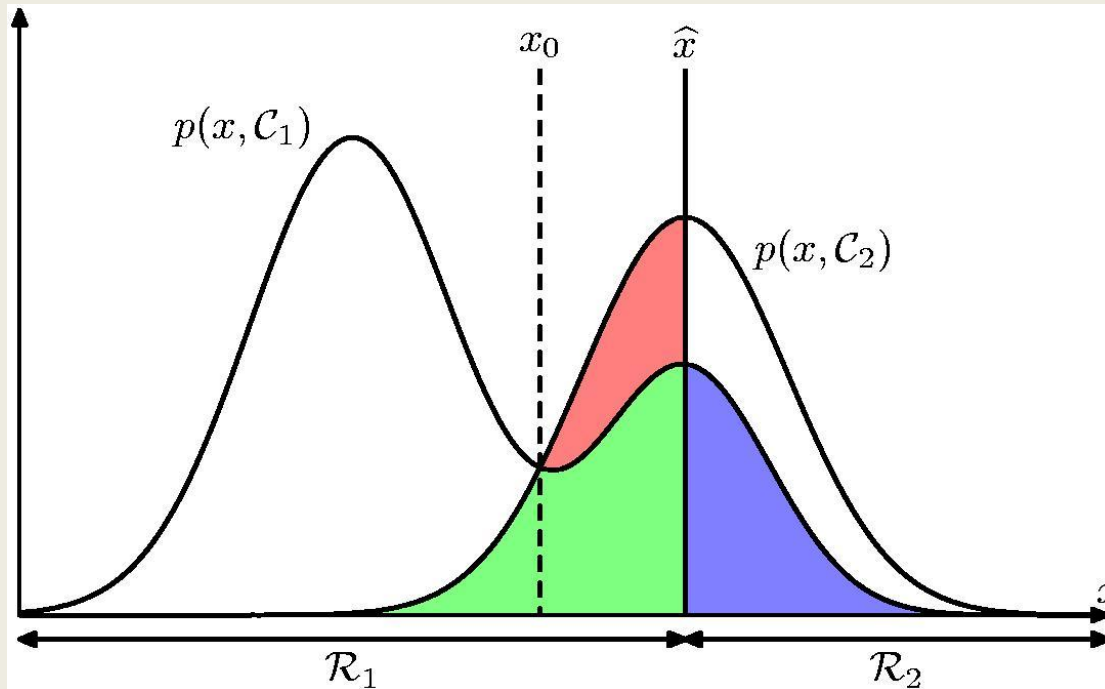
- **Example: Classification Rule**

- decision region R_k over $\mathbf{x}$
- corresponding to class C_k (decision)

$$p(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k) p(C_k)}{p(\mathbf{x})}$$



Minimum Misclassification Rate

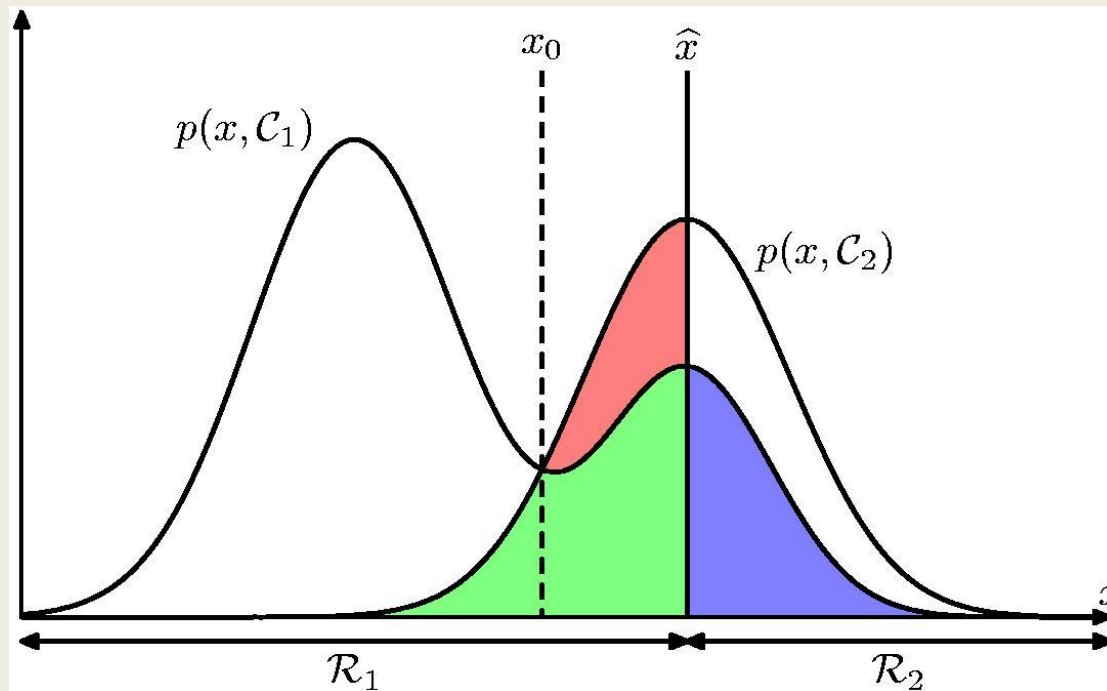


Choose the class
with the
highest
probability

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, C_2) + p(\mathbf{x} \in \mathcal{R}_2, C_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, C_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, C_1) d\mathbf{x}. \end{aligned}$$



Maximum Correct Classification Rate



$$\mathcal{C}^* = \arg \max_k p(\mathbf{x}, \mathcal{C}_k)$$

$$p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k | \mathbf{x}) p(\mathbf{x})$$

$$\mathcal{C}^* = \arg \max_k p(\mathcal{C}_k | \mathbf{x})$$

Choose the class
with the highest
posterior probability

$$p(\text{correct}) = \sum_{k=1}^K p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) = \sum_{k=1}^K \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}$$



Minimum Expected Loss

Decision

cancer normal

Truth	cancer	0	1000
	normal	1	0

- **Loss Function** L_{kj}
 - associates decisions with costs
 - depends on decision and truth

- **Minimum Expected Loss**

- expected cost due to mistakes

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}$$

- choose regions that minimize the expected loss

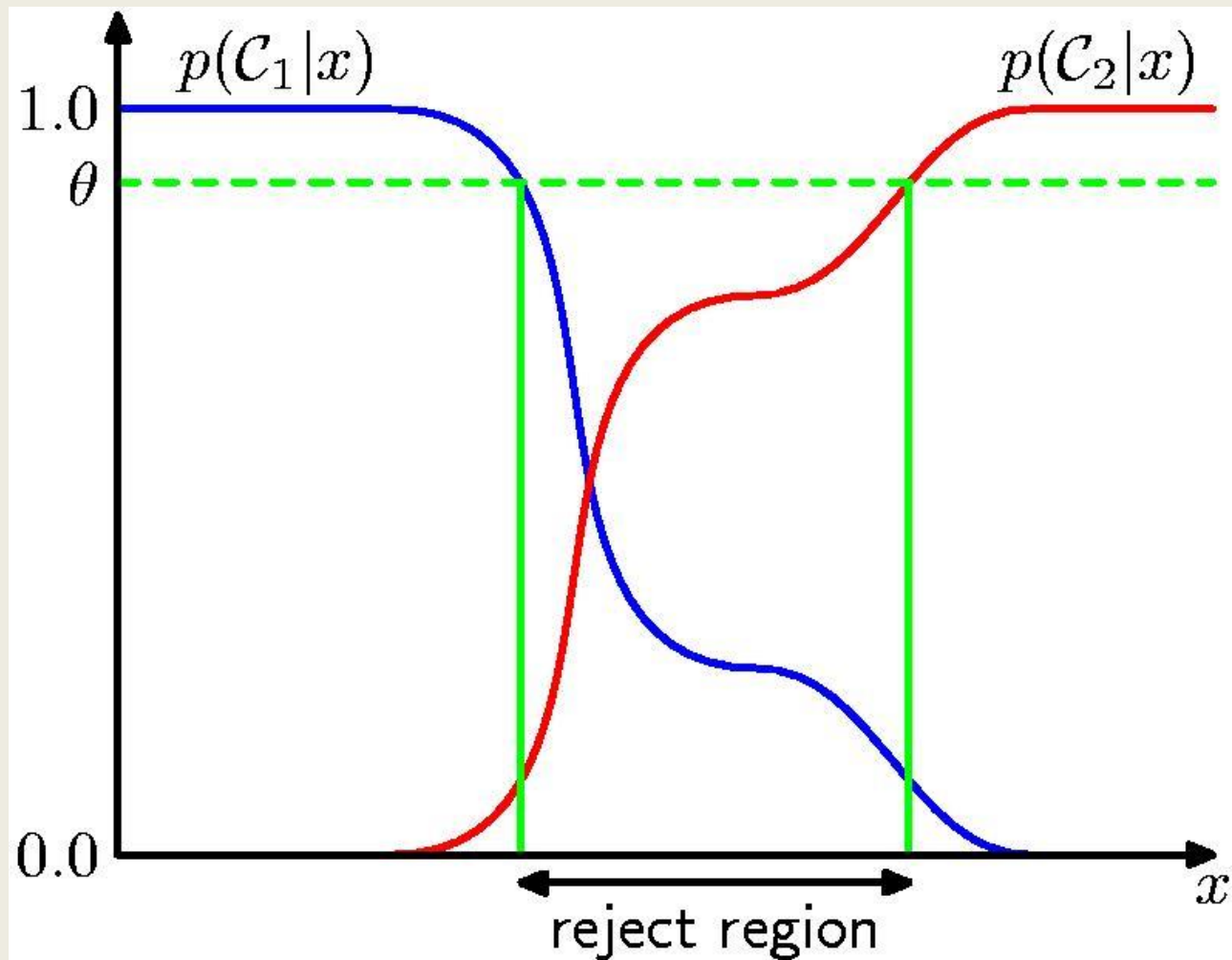
- **Decision**

- choose class that minimizes loss

$$C^* = \arg \min_j \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$$



Reject Option



Approaches to Classification Decision Problem

- **Approach I: generative models**
 - determine the class-conditional densities $p(\mathbf{x}|C_k)$ (and priors)
 - infer the posterior class probabilities $p(C_k|\mathbf{x})$ (through Bayes)
 - use decision theory to make decision
- **Approach II: discriminative models**
 - determine the posterior class probabilities $p(C_k|\mathbf{x})$
 - use decision theory to make decision
- **Approach III: discriminant functions**
 - determine a function that maps inputs to classes directly
 - use the discriminant function to make decision

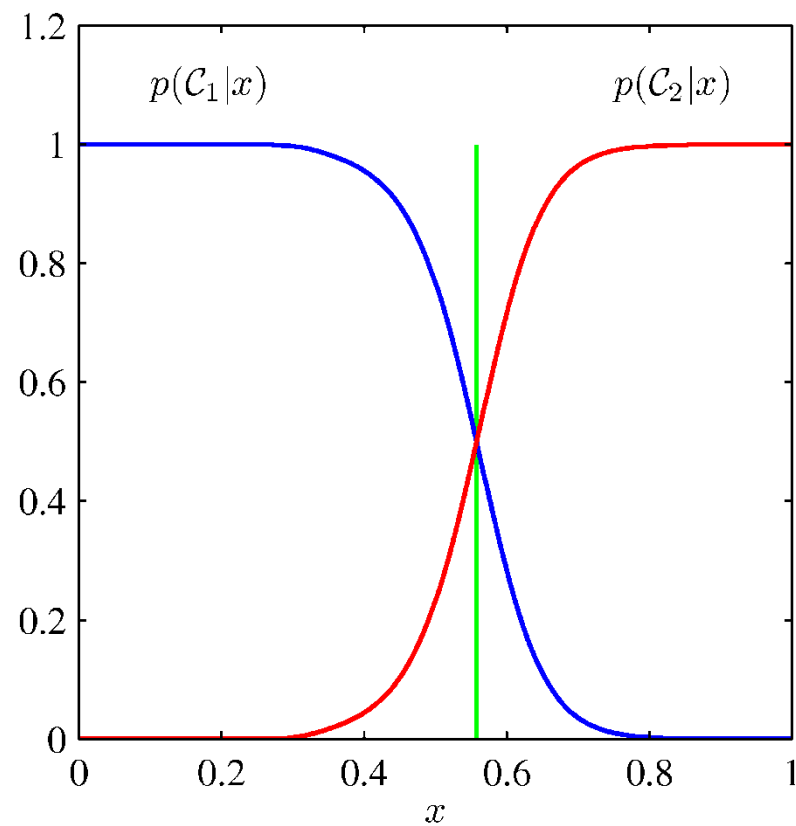
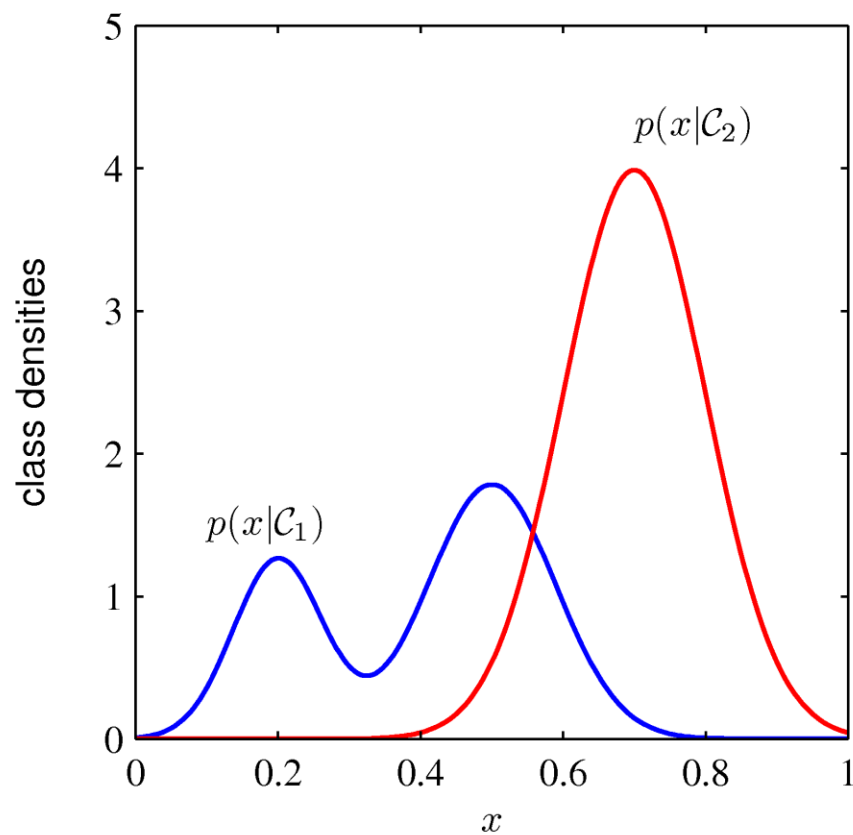


Pros and Cons

- **Approach I: generative models**
 - pros: can be used to generate synthetic data
 - cons: solves a much bigger problem, quite demanding
 - pros: can be used for outlier/novelty detection
- **Approach II: discriminative models**
 - pros: makes no waste of resources, avoids complex models
 - cons: solves a slightly bigger problem
- **Approach III: discriminant functions**
 - pros: a single (learning) problem
 - pros (cons?): may not use probabilities at all!
 - cons: black-box approach, provides no intuition



Comparison: $p(x|C_k)$ vs. $p(C_k|x)$ vs. x^*



Separating Inference and Decision

- **Minimizing risk**
 - loss matrix may change over time
 - changes can be easily introduced in the minimum risk criterion
- **Reject option**
 - posterior probabilities allow for a rejection criterion
- **Unbalanced class priors**
 - posterior probabilities can be transformed to other populations
- **Combining models**
 - split the problem, solve separately, model for each sub-problem
 - combine sub-models using probability theory to solve problem
 - exploit conditional independence



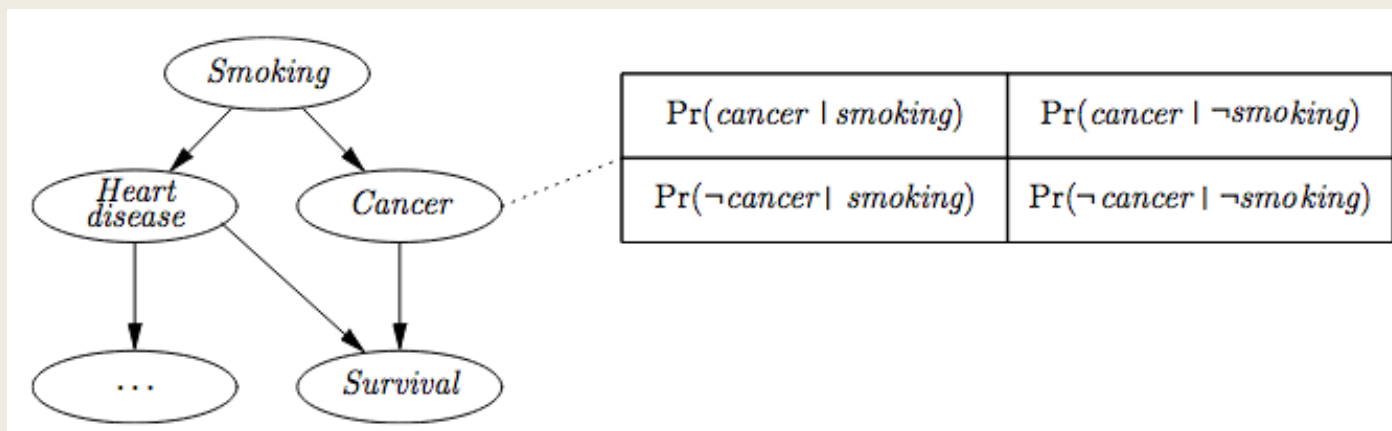
Bayesian Networks

A celebrated decision-making under uncertainty paradigm and tool



Bayesian Networks (BNs) or Probabilistic Graphical Models (PGMs)

- Events can be characterized by/linked to random variables.
How to represent/easily calculate their probability?
- There can be (in-)dependencies among the variables
- Graphical representation of the (in-)dependencies (via **DAGs**):



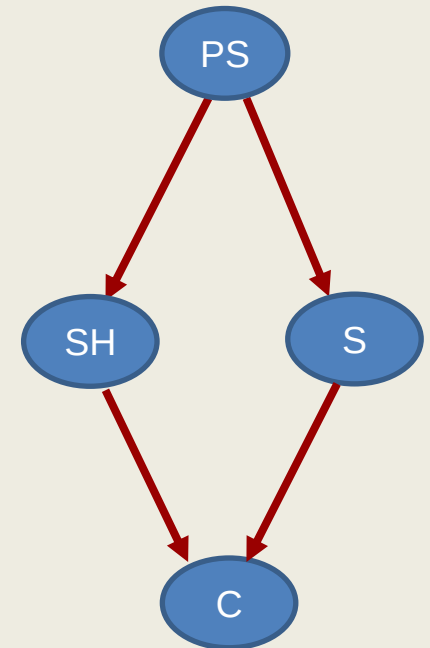
- *This can help provide structure to a problem!*
- *Makes it easier to reason about*
- *Makes it easier to elicit information from experts*
- *Could require far fewer numbers to represent the probability*



Example: reasoning about smoking causing cancer - using Bayesian networks for inference

- C: “has cancer”
- SH: “exposed to second-hand smoking”
- PS: “at least one parent smokes”
- S: “smokes”

- a qualitative representation of independencies
 - Whether or not the patient has cancer is directly influenced by S and SH
 - S and SH are influenced by PS
 - PS influences C also... However,
 - ***once values of SH and S are known, finding out about PS provides no additional information***
- We say that “C is independent of PS, given SH and S”



Why do we care about independence?

We want to represent probability distributions compactly

- State spaces characterized by $X_1, \dots, X_n$ random variables
- States have form $(x_1, \dots, x_n) : X_1 = x_1, \dots, X_n = x_n$
- Example: suppose $X_1, \dots, X_5$ independent *binary* variables
 - We can completely characterize a distribution by 5 numbers!
 - $\Pr(X_i = 0)$ for $i = 1, \dots, 5$
 - If $\Pr(X_i = 0) = a_i$, then $\Pr(X_i = 1) = 1 - a_i$
 - Independence: $\Pr(1, 0, 0, 1, 0) = (1 - a_1) * a_2 * a_3 * (1 - a_4) * a_5$
 - Can compute probability of a set of states by adding things

Thus, we can compactly describe the distribution using only n numbers (the $\Pr(X_i = 0)$ ones) instead of 2^n



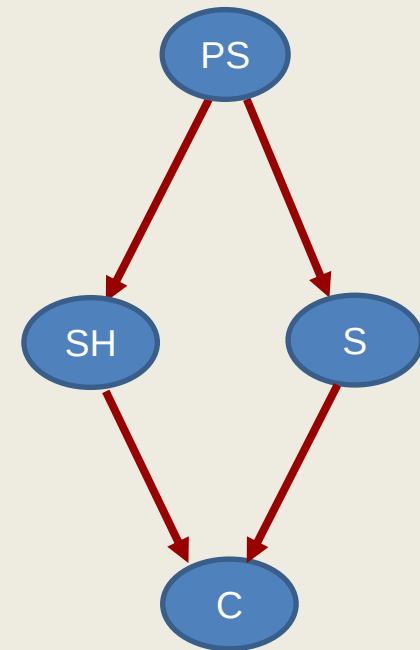
Quantitative BNs: Qualitative BNs + conditional probability tables (CPTs)

S	SH	C=1
1	1	0.6
1	0	0.4
0	1	0.1
0	0	0.01

PS	S=1
1	0.4
0	0.2

PS	SH=1
1	0.8
0	0.3

PS = 1
0.3



- The tables include only values for $\Pr(C=1 \mid S, SH)$, $\Pr(S=1 \mid PS)$, $\Pr(SH=1 \mid PS)$, $\Pr(S=1)$, but:
 - $\Pr(C=0 \mid SH) = 1 - \Pr(C=1 \mid SH)$
 - Can similarly compute other values of interest

via the chain rule and exploiting independencies from non-descendant nodes given parent nodes...

$$\Pr(PS = 0, S=0, SH=1, C=1) =$$

$$\Pr(C=1 \mid S=0, SH=1) * \Pr(S=0 \mid PS=0) * \Pr(SH=1 \mid PS=0) * \Pr(PS=0) = 0.1 * 0.8 * 0.3 * 0.7 = 0.0168$$



Recap: What do BNs buy us?

- **Conciseness of representation**
 - If a node has $< k$ parents, then we need $< 2^k n$ numbers to represent the distribution
 - if k is not too large, then $2^k n \ll 2^n$
- **Information tends to be easier to elicit**
 - experts are better in giving information about dependencies, rather than about numbers
- **Information easier to understand**
- **Many computational tools**
 - for computing probabilities given information
 - for learning Bayesian networks
 - many commercial packages exist
- **Numerous applications**
 - notably for medical decision making and diagnosis



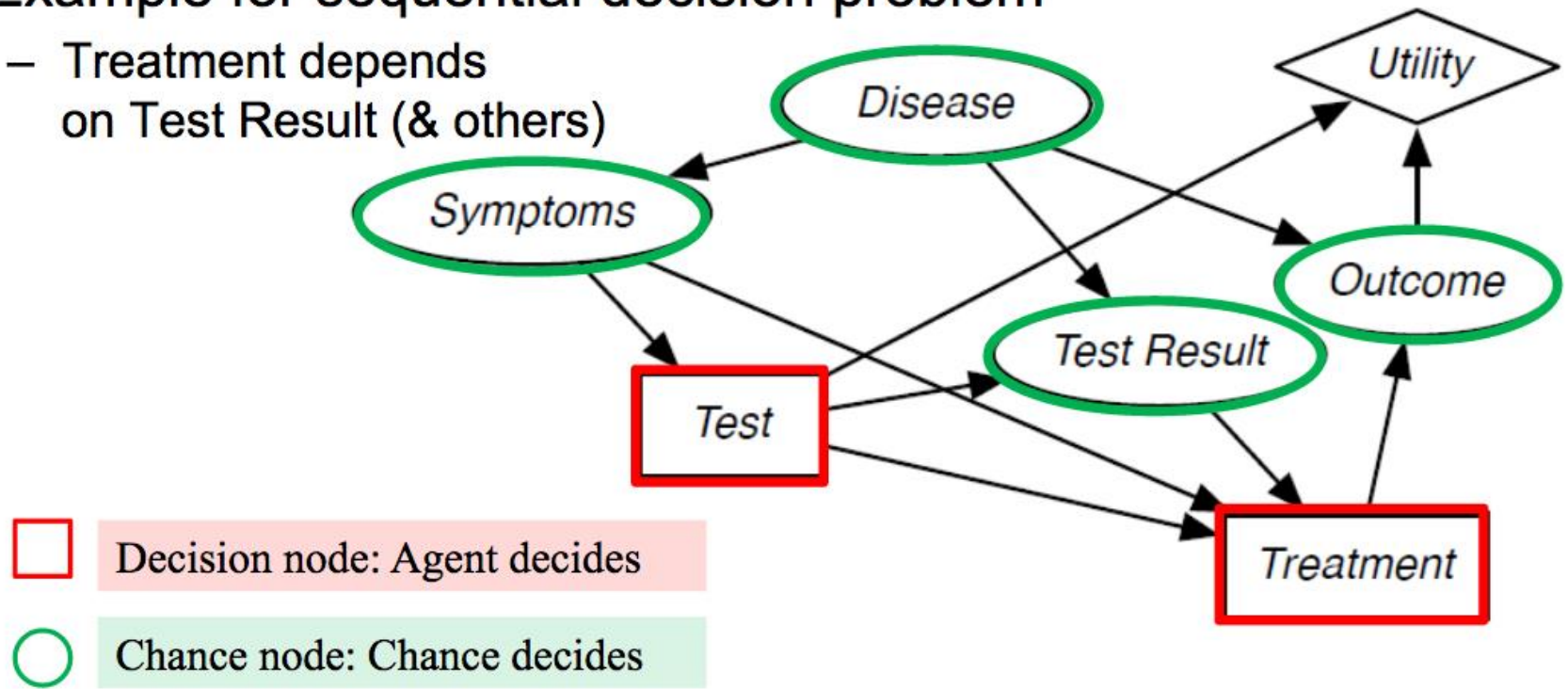
Making Sequentially Optimal Decisions



Decision Networks

Example for sequential decision problem

- Treatment depends on Test Result (& others)



- Each decision has an *information set* of parent variables (pa), whose value will be known at the time the decision is to be made
- $pa(\text{Test}) = \{\text{Symptoms}\}$, $pa(\text{Treatment}) = \{\text{Test}, \text{Symptoms}, \text{Test Result}\}$

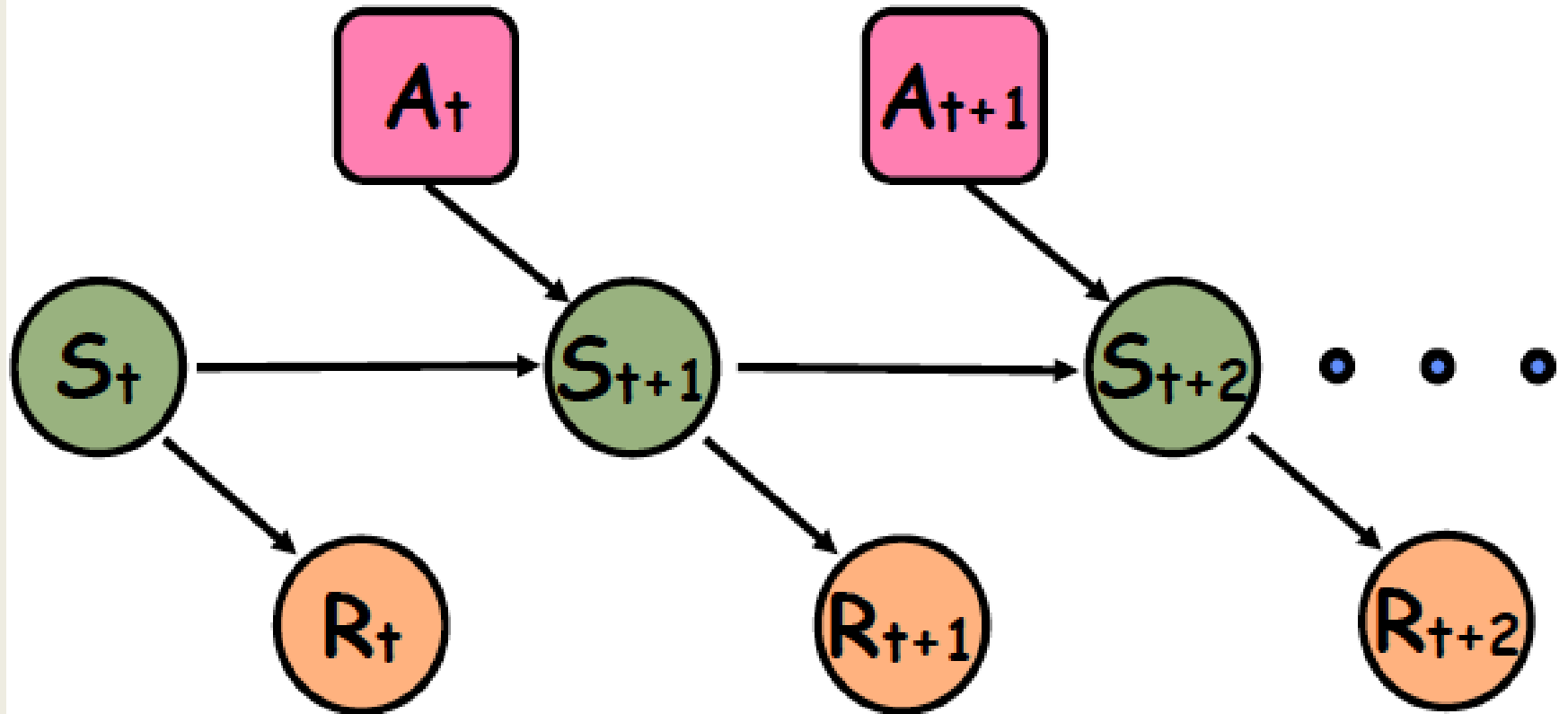


Markov Decision Processes (MDPs)

- MDPs are a family of nondeterministic search problems
- An MDP has four components: **S , A , R , Pr**
 - (finite) state set S ($|S| = n$)
 - (finite) action set A ($|A| = m$)
 - transition function $Pr(s' | s, a)$
 - each $Pr(s' | s, a)$ is a distribution over S
 - represented by set of $n \times n$ stochastic matrices
 - probability of transition to next state s' , when choosing action a in current state s
 - bounded, real-valued reward function $R(s)$
 - represented by an n -vector
 - can be generalized to include action costs: $R(s, a)$
 - can be stochastic (replaceable by expectation)



Graphical view of MDPs



What is Markov about MDPs

- Andrey Markov (1856-1922)
- “Markov” chains: given present state, future is independent from the past
- In MDPs: past states/actions are irrelevant, when taking decisions at present



Assumptions

- Markovian dynamics (history independence)
 - $\Pr(S^{t+1} \mid A^t, S^t, A^{t-1}, S^{t-1}, \dots, S^0) = \Pr(S^{t+1} \mid A^t, S^t)$
- Markovian reward process
 - $\Pr(R^{t+1} \mid A^t, S^t, A^{t-1}, S^{t-1}, \dots, S^0) = \Pr(R^{t+1} \mid A^t, S^t)$
- Stationary transition and reward dynamics
 - $\Pr(S^{t+1} \mid A^t, S^t) = \Pr(S^{t'+1} \mid A^{t'}, S^{t'})$ for all t, t'
 - $\Pr(R^{t+1} \mid A^t, S^t) = \Pr(R^{t'+1} \mid A^{t'}, S^{t'})$ for all t, t'
- Full observability
 - though we can't predict what state we will reach when we execute an action, once it is realized, we know what it is



Policies

- **Non-stationary policy**
 - $\pi: S \times T \rightarrow A$
 - $\pi(s,t)$ is action to do at state s **with t-steps-to-go**
- **Stationary policy**
 - $\pi: S \rightarrow A$
 - $\pi(s)$ is action to do at state s (regardless of time)
 - analogous to reactive or universal plan
- These assume or have these properties:
 - full observability
 - history-independence
 - deterministic action choice

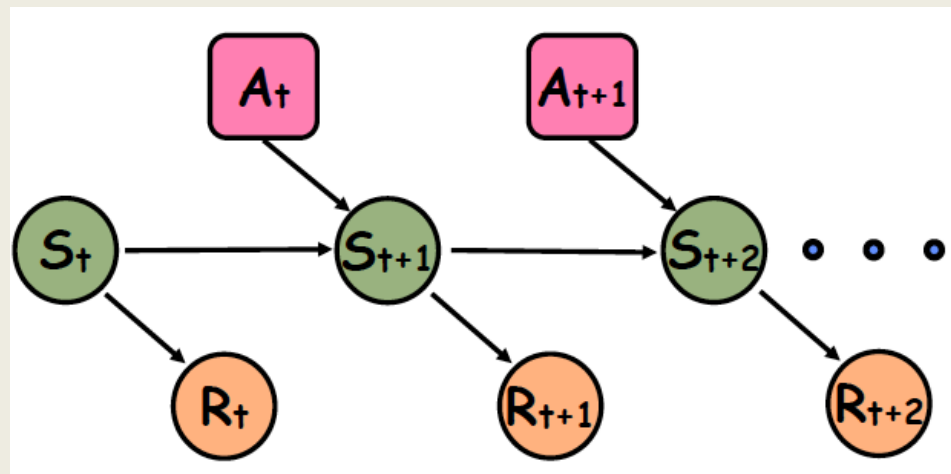


Value (utility) of a policy

- How good is a policy π ?
- Does it choose actions that bring a lot of reward in the long run?
- How do we measure “accumulated” reward?
- **Value function $V: S \rightarrow R$ (real-valued function)**
 - Associates each state with a value
- **$V^\pi(s)$ denotes value of policy π at state s**
 - **expected accumulated reward over horizon of interest**
 - Note that $V^\pi(s) \neq R(s)$: it measures (long-term) utility



MDPs are commonly "solved" via Dynamic Programming



What is Dynamic Programming?

Dynamic Programming (DP)

- methods for successively approximating optimal evaluation functions and decision rules

Used for:

- optimization problems with **objectives=sequences of states**
- DP solves these by solving recurrence relations
- **Basic DP step:**
backing-up state evaluations to solve the recurrence relations
- **Bootstrapping:** *updating estimates based on other (previous) estimates (of other quantities)*



(cons) requires repeated generation and expansion of all possible next states from each current state

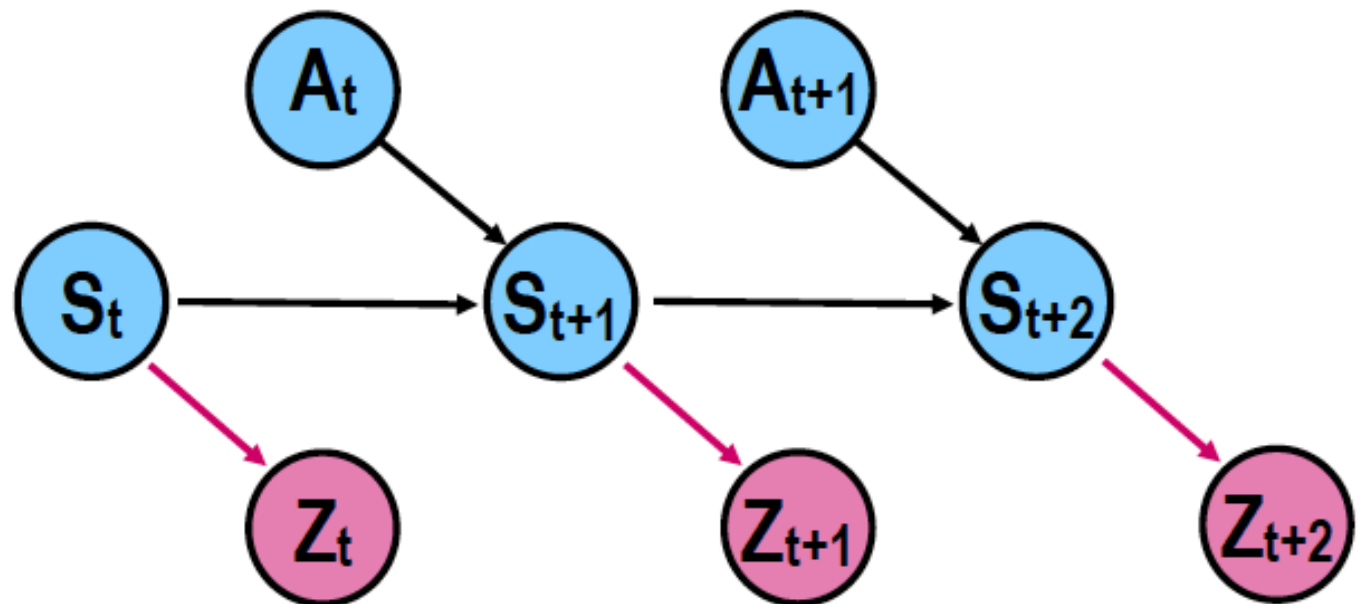
Enter Partial Observability

- Classical planning: observable states, deterministic actions
- MDPs: observable states, stochastic actions
- POMDPs: **partially observable states**, stochastic actions
 - *“The world is a huge POMDP. You just have to solve it.”*
[Craig Boutilier, circa 2002]

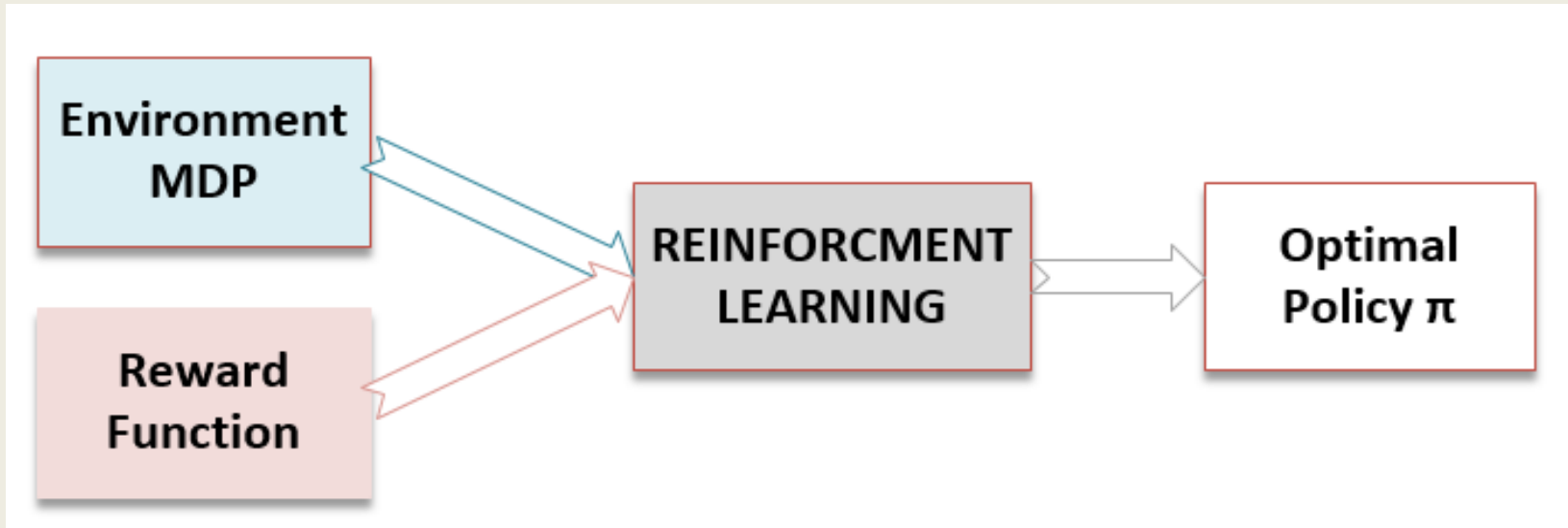


POMDP

- As in MDPs: $S, A, p_{ij}^a, r_i^a, r_i^T$
- Observation space: Z (or Z_a)
- Observation probabilities: p_{ijz}^a for $z \in Z_a$



Reinforcement Learning



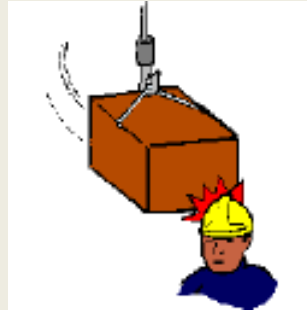
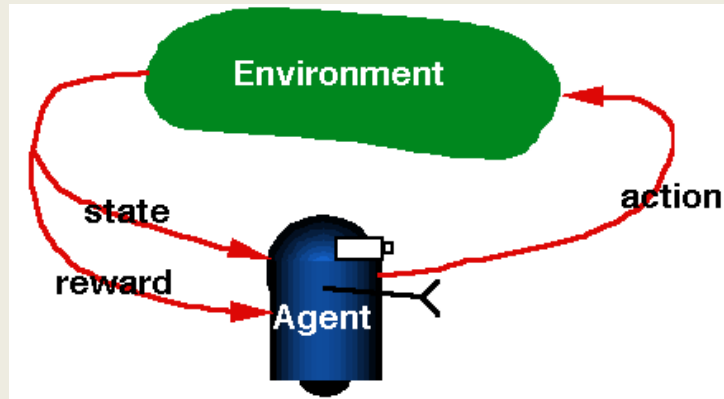
- Agent interacts with the environment (Markov Decision Process): chooses action, transitions to new state, and obtains a reward.
- The action choice in each situation (state) is based on its *value*: the (estimated or learned) long-term reward from similar situations.

Goal : maximize the expected reward obtained over long term

Goal of RL: policy π that specifies an action for each possible state



Reinforcement Learning



“ You do something, the world bangs you on the head; then you do something else, the world bangs you on the head again...”

Sam Roweis, 2002

✦ RL is learning by trial and error, i.e. learns how to map situations to actions, so as to maximize a sequence of rewards.

Approaches

- **model-based methods:** use observed rewards+transitions to build models of the world and use these to derive policies
- **model-free methods:** approximate the quality of choices directly from sampled reward-transition observation (e.g. “Q-learning”)



The Exploration vs. Exploitation Dilemma

- To explore or to exploit?
 - Stay / be content with what I know, or
 - ... attempt to test other states-action pairs?



Exploration vs. Exploitation++

- **RL**: the agent should explicitly **explore** the environment in order to acquire knowledge about it
- Act to **improve the estimate of the value function** (exploration) or to **get high (anticipated) payoffs** (exploitation) ? **Dilemma!**
- Rewards maximization ***requires*** exploration; but **too much exploration of irrelevant parts can waste time!**
 - Choice dependent on particular learning problem, as well as the particular learning technique employed
 - famous way to treat exploration vs exploitation problem:
 - GLIE: greedy in the limit exploration methods
 - better: the Bayesian approach



Tetris: an example of RL Results

Handcoded Player

Average Score ≈ 675

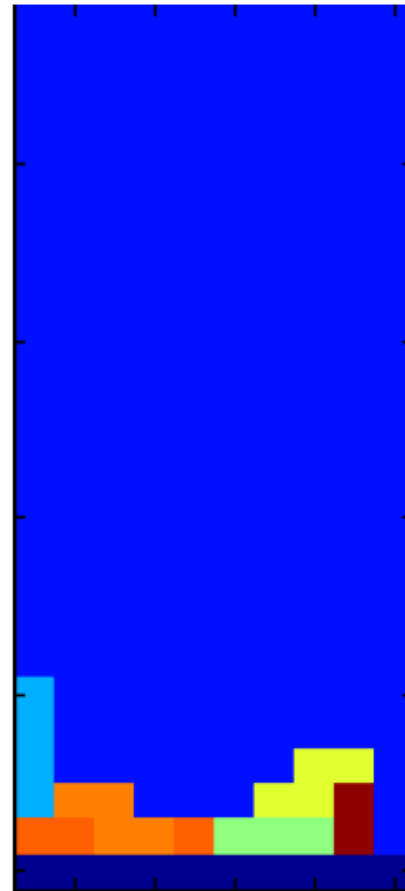
Score = 49
(after 150 steps)



Learned Player

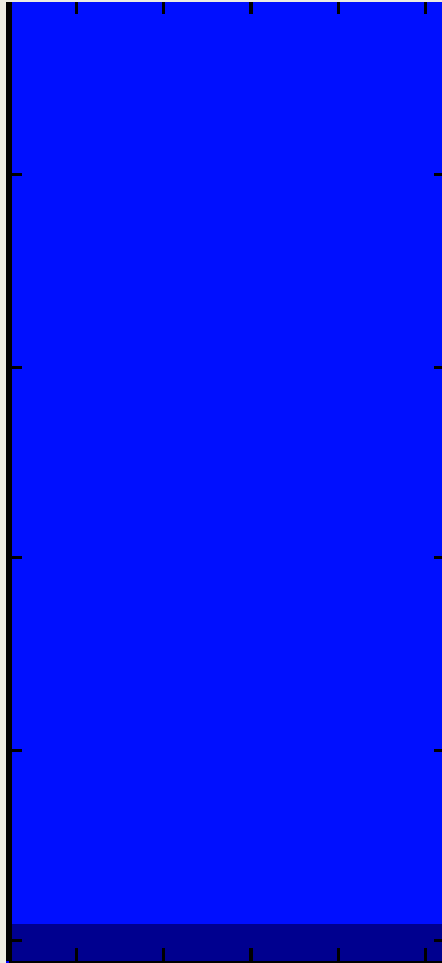
Average Score ≈ 3000

Score = 58
(after 150 steps)

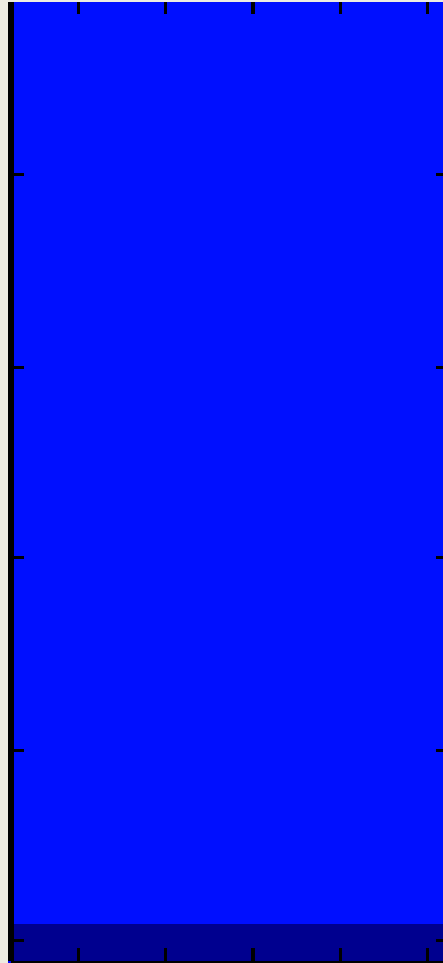


Tetris: Handcoded player and LSPI

Handcoded
Video



LSPI
Video



Suggested readings

- AI: A Modern Approach, 4th Edition
 - Chapters 12, 13, 16, and 17

