

CS 486



Personal Notes

Marcus Chan

Taught by Pascal Poupart & Sriram Ganapathi
Subramanian

JW CS '25



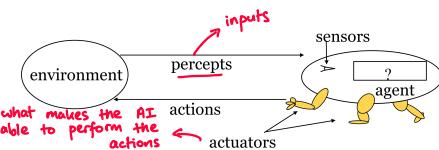
Chapter 1: Introduction

REINFORCEMENT LEARNING PROBLEM



Our goal is for the AI to learn to choose actions that maximize rewards.

AGENTS & ENVIRONMENTS



The "agent function" maps percepts to actions; ie $f: P \rightarrow A$.

The "agent program" runs on the physical architecture to produce f .

RATIONAL AGENTS

A "rational agent" chooses whichever action that maximizes the expected value of its performance measure given the percept sequence to date.

Note that rationality is not omniscience, but rather learning & autonomy.

PEAS

"PEAS" helps us specify the task environment:

- ① Performance measure;
eg safety, destination, etc
- ② Environment;
eg streets, traffic, etc
- ③ Actuators; &
eg steering, brakes, etc.
- ④ Sensors.
eg GPS, engine sensors, etc.

PROPERTIES OF TASK ENVIRONMENTS

Task environments can be:

- ① fully vs partially observable:
 - fully observable: agent knows state of the world from the stimuli
 - partially observable: agent does not directly observe the world's state

② deterministic vs stochastic;

- deterministic: next state is observable at any time
- stochastic: next state is unpredictable

③ episodic vs sequential;

- episodic: agent's current action will not affect a future action
- sequential: agent's current action will affect a future action

④ static vs dynamic;

- static: model is trained once
- dynamic: model is trained continuously

⑤ discrete vs continuous;

⑥ single agent vs multiagent.

*the former option is "easier" than the latter.

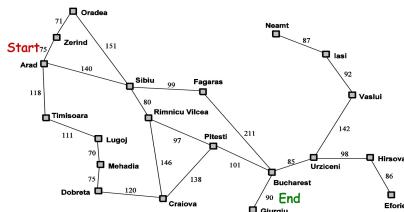
Chapter 2: Uninformed Search Techniques

SIMPLE PROBLEM SOLVING AGENT

```
function SIMPLE-PROBLEM-SOLVING-AGENT(percept) returns an action
  static: seq, an action sequence, initially empty
    state, some description of the current world state
    goal, a goal, initially null
    problem, a problem formulation
  state ← UPDATE-STATE(state, percept)
  if seq is empty then do
    goal ← FORMULATE-GOAL(state)
    problem ← FORMULATE-PROBLEM(state, goal)
    seq ← SEARCH(problem)
  action ← FIRST(seq)
  seq ← REST(seq)
  return action
```

- This can only tackle problems that are
- ① fully observable;
 - ② deterministic;
 - ③ sequential;
 - ④ static;
 - ⑤ discrete; &
 - ⑥ single agent.

EXAMPLE: TRAVELLING IN ROMANIA



- initial state: In Arad
- actions: drive between cities
- goal test: In Bucharest
- path cost: distance between cities

EXAMPLE: 8-TILE PUZZLE



Start State



Goal State

- states: locations of 8 tiles & blank
- initial state: any state
- actions: up, down, left, right
- goal test: does state match desired configuration
- path cost: # of steps

SEARCHING

Q: We can visualize a state space search in terms of trees or graphs;

- ① nodes correspond to states; &
- ② edges correspond to taking actions.

Q: These "search trees" are formed using "search nodes", which have

- ① the state associated with it;
- ② parent node & operator applied to the parent to reach the "current" node;
- ③ cost of the path so far; &
- ④ depth of the node.

EXPANDING NODES

Q: "Expanding a node" refers to applying all legal operators to the state contained in the node & generating nodes for all corresponding successor states.



GENERIC SEARCH ALGORITHM

Q: Algorithm:

- ① Initialize search with initial problem state.
- ② Then repeat:
 - if no candidate nodes can be expanded, return failure
 - otherwise, choose a leaf node for expansion according to our search strategy.
 - if the node contains a goal state, return the solution.
 - otherwise, expand the node by applying the legal operators to the state associated within the node, & add the resulting nodes to the tree.

EVALUATING SEARCH ALGORITHMS

We can use the following properties when evaluating search algorithms:

- ① "completeness" — is the algorithm guaranteed to find a solution (if it exists)?
 - ② "optimality" — does the algorithm find the optimal solution (ie lowest path cost)?
 - ③ time & space complexity.
- We consider the following variables:
- ① "branching factor" (b) — the # of children each node has
 - ② depth of shallowest goal node (d); &
 - ③ max length of any path in the state space (m).

BREADTH-FIRST SEARCH

Refer to CS341 notes for details; we expand all nodes on a given level before any node on the next level is expanded.

Evaluating the algorithm:

- ① Completeness: yes if $b \geq 0$
- ② Optimality: yes if all costs are same
- ③ Time: $1+b+\dots+b^d \in O(b^d)$
- ④ Space: $O(b^d)$.

* all uninformed search methods have exponential time complexity.

UNIFORM COST SEARCH

Idea: we expand the node with the lowest path cost.

We can implement this using a priority queue.

Let C^* = cost of optimal solution & $\epsilon = \min \text{action cost}$. Then

- ① Completeness: yes if $\epsilon > 0$
- ② Optimality: yes
- ③ Time: $O(b^{C^*/\epsilon})$
- ④ Space: $O(b^{C^*/\epsilon})$

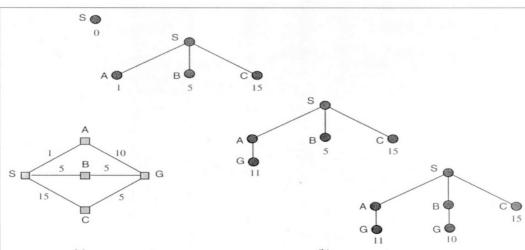


Figure 3.13 A route-finding problem. (a) The state space, showing the cost for each operator. (b) Progression of the search. Each node is labelled with $g(n)$. At the next step, the goal node with $g = 10$ will be selected.

DEPTH-FIRST SEARCH

Refer to CS341 notes for details; we expand the deepest node in the current fringe of the search tree first.

Evaluation:

- ① Complete: no
- may get stuck going down a long path
- ② Optimal: no
- might return a solution which is deeper (ie more costly) than another solution
- ③ Time: $O(b^m)$
- note we might have $m > d$
- ④ Space: $O(bm)$

DEPTH-LIMITED SEARCH

Idea: Treat all nodes at depth l as if they have no successors.
- try to choose l based on the problem

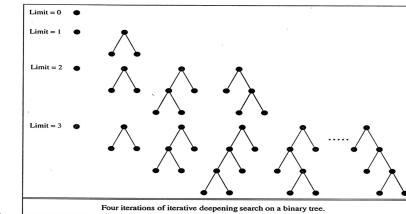
This avoids the problem of unbounded trees.

Evaluation:

- ① Time: $O(b^l)$
- ② Space: $O(b^l)$
- ③ Complete: no
- ④ Optimal: no

ITERATIVE-DEEPENING

Idea: repeatedly perform depth-limited search, but increase the limit each time.



Evaluation:

- ① Complete: yes
- ② Optimal: yes
- ③ Time: $O(b^d)$
- ④ Space: $O(b^d)$

Time:

$$\begin{aligned}
 \text{(limit=1)} & \quad 1 \\
 \text{(limit=2)} & \quad 1 + b \\
 \text{(limit=3)} & \quad 1 + b + b^2 \\
 & \vdots \\
 \text{(+) (limit=d)} & \quad 1 + b + b^2 + \dots + b^d \\
 & \quad d + (d-1)b + (d-2)b^2 + \dots \\
 & \quad + b^d \in O(b^d)
 \end{aligned}$$

Chapter 3: Informed Search Techniques

MOTIVATION

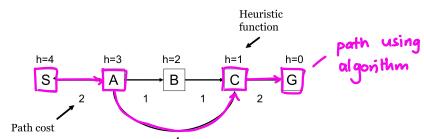
- In search problems, we often have additional knowledge about the problem.
eg with the "travelling around Romania", we know dist. bw cities
↳ so we can find the overhead in going the wrong direction
- Our knowledge is often about the "merit" of nodes.
- Notions of merit:
 - how expensive it is to get from a state to a goal;
 - how easy it is to get from a state to a goal.

HEURISTIC FUNCTIONS • $h(n)$

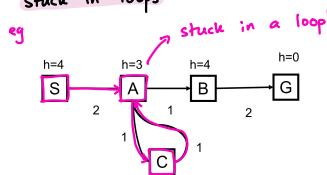
- We need to develop domain specific "heuristic functions" $h(n)$, which guess the cost of reaching the goal from node n .
- In general, if $h(n_1) < h(n_2)$, we guess reaching the goal is cheaper from n_1 than from n_2 . We also need
 - $h(n) = 0$ if n is a goal node
 - $h(n) > 0$ if n is not a goal node.

GREEDY BEST-FIRST SEARCH

- Idea: Use $h(n)$ to rank the nodes in the fringe & expand the node with the lowest h -value.
(ie "greedily" trying to find the least-cost solution).
- Example:



- Note greedy best-first is not optimal.
eg in the above example,
 - path found has cost=6.
 - but cheaper path is $S \rightarrow A \rightarrow B \rightarrow C \rightarrow G$ with cost=6.
- It is also not complete, as it can be stuck in loops.



- but if we check for repeated states then we are okay
- This algorithm uses exponential space & worst-case time.

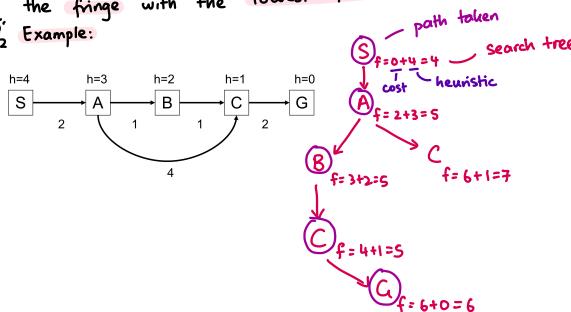
A* SEARCH

Θ_1 Idea: Define $f(n) = g(n) + h(n)$, where

- ① $g(n)$ = cost of path to node n
- ② $h(n)$ = heuristic estimate of cost of reaching goal from node n .

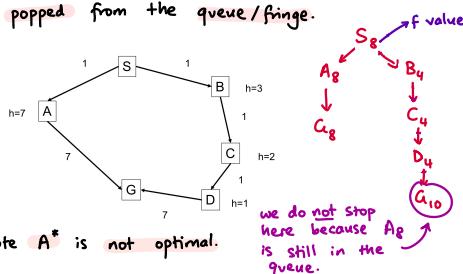
We then iteratively expand the node in the fringe with the lowest f value.

Θ_2 Example:

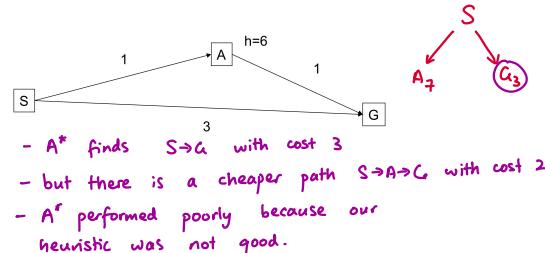


Θ_3 A^* should terminate only when the goal state is popped from the queue/fringe.

eg



Θ_4 Note A^* is not optimal.



- A^* finds $S \rightarrow G$ with cost 3
- but there is a cheaper path $S \rightarrow A \rightarrow C$ with cost 2
- A^* performed poorly because our heuristic was not good.

ADMISSIBLE (HEURISTICS)

Θ_1 Let $h^*(n)$ be the true minimal cost to the goal from node n .

Then, we say the heuristic $h(n)$ is "admissible" if

$$h(n) \leq h^*(n) \quad \forall n.$$

Θ_2 In particular, admissible heuristics never overestimate the cost to the goal.

$h(n)$ IS ADMISSIBLE $\Rightarrow A^*$ IS OPTIMAL

Θ_1 If $h(n)$ is admissible, then A^* with tree-search is optimal.

Proof. Let G be an optimal goal state, & $f(G) = f^* = g(G)$.

Let G_2 be a suboptimal goal state, ie $f(G_2) = g(G_2) > f^*$. (since $h(G_1) = h(G_2) = 0$)

Assume, for a cont'n, that A^* selects G_2 from the queues; ie A^* terminates with a suboptimal solution.

Let $n =$ node currently a leaf node on an optimal path to G .



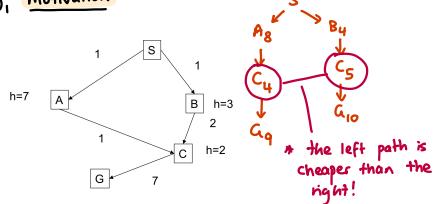
G_1

$\bullet G_2$

Since h is admissible, $f^* \geq f(n)$. If n is not chosen for expansion over G_2 , then $f(n) \geq f(G_2)$. Thus $f^* \geq f(G_2)$. Since $h(G_2) = 0$, thus $f^* \geq g(G_2)$, a cont'n.

REVISITING STATES IN A*

Θ_1 Motivation:



Θ_2 If we allow states to be expanded again, we might get a better solution!

CONSISTENT (HEURISTICS)

Θ_1 We say $h(n)$ is "consistent" if

$$h(n) \leq \text{cost}(n, n') + h(n') \quad \forall n, n'.$$

Θ_2 Note that A^* graph-search with a consistent heuristic is optimal.

PROPERTIES OF A*

Θ_1 Note that A^* is

- ① Complete if $h(n)$ is consistent; - f always increases along any path
- ② Has exponential time complexity in the worst-case; & - but a good heuristic helps a lot - $O(\log m)$ if heuristic is perfect
- ③ Has exponential space complexity.

ITERATIVE DEEPENING A*

(IDA*)

- Idea: Like iterative deepening search, but change f-cost rather than depth in each iteration.
- This reduces the space complexity.

SIMPLIFIED MEMORY-BOUNDED A*

(SMA*)

- Idea: Proceeds like A* but when it runs out of memory it drops the worst leaf node (ie one with highest f-value).
- If all leaf nodes have the same f-value, then drop the oldest & expand the newest.
- This is
 - ① optimal;
 - ② complete if depth of shallowest goal node < memory size.

OBTAINING HEURISTICS

- One approach to get heuristics is to think of an easier problem & let $h(n)$ be the cost of reaching the goal in the easier problem.

We can also

- ① precompute solution costs of subproblems & store them in a pattern database; or
- ② learn from experience with the problem class.

EXAMPLE: 8-PUZZLE GAME

We can relax the game in 3 ways:

- ① We can move tile from position A \rightarrow B if A is next to B (ignore whether position is blank)
- ② We can move tile from position A \rightarrow B if B is blank (ignore adjacency)
- ③ We can move tile from position A \rightarrow B regardless.

③ leads to the "misplaced tile heuristic". (h_3)

- to solve this problem we need to move each tile into its final position.
- # of moves = # of misplaced tiles
- admissible

① leads to the "manhattan distance heuristic". (h_1)

- to solve this we need to slide each tile into its final position

- admissible

④ Note h_1 "dominates" h_3 ; ie $h_3(n) \leq h_1(n) \forall n$.

Chapter 4:

Constraint Satisfaction

INTRODUCTION

Q: These are useful for problems where the state structure is important.

Q: In many problems, the same state can be reached independent of the order in which the moves are chosen.

Q: So, we can try to solve problems efficiently by being smart about the action order.

4-QUEENS CONSTRAINT PROPAGATION

Q: Idea: Remove conflicting squares from consideration when we put a queen down.



CONSTRAINT SATISFACTION PROBLEM (CSP)

Q: A "constraint satisfaction problem" is defined by some $\{V, D, C\}$, where

① $V = \{V_1, \dots, V_n\}$ is a set of variables;

② $D = \{D_1, \dots, D_n\}$ is a set of domains, where D_i is the set of possible values for each V_i ;

&

③ $C = \{C_1, \dots, C_m\}$ is the set of constraints.

STATE

Q: A "state" is an assignment of values to some or all of the variables;

ie $V_i = x_i, V_j = x_j, \dots$.

CONSISTENT [ASSIGNMENT]

Q: We say an assignment is "consistent" if it does not violate any constraints.

SOLUTION

Q: A "solution" is a complete, consistent assignment.

EXAMPLE: 8 QUEENS AS A CSP

Q: 8 queens as a CSP:

- variables: $V_{ij}, i, j = 1, \dots, 8$

- domain of each var: $\{0, 1\}$

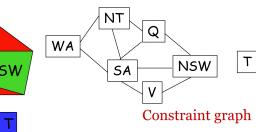
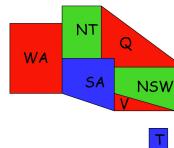
- constraints: $V_{ij} = 1 \Rightarrow V_{ik} = 0 \quad \forall k \neq j$

$V_{ij} = 1 \Rightarrow V_{kj} = 0 \quad \forall k \neq i$

similar constraint for diagonals

$$\sum_{i,j} V_{ij} = 8$$

EXAMPLE: MAP COLORING



Constraint graph

WA \neq NT, WA \neq Q, WA \neq SA, WA \neq NSW, WA \neq T

- variables: WA, NT, ..., T (the regions)
- each var has the same domain: {red, green, blue}
- no 2 adjacent variables have the same value
(ie $WA \neq NT$, $WA \neq SA$, etc)

PROPERTIES OF CSPS

Q: Types of variables:

① Discrete & finite;

eg 8-queens, map coloring

* we focus on this in this course.

② Discrete with infinite domains; &

eg job scheduling

③ Continuous domains.

Q: Types of constraints:

① "Unary constraint": relates a single variable to a value

- eg Queensland = blue

② "Binary constraint": relates two variables

③ "Higher order constraints": relates ≥ 3 variables.

CSPs & SEARCH

Q: We can formulate CSPs as a search problem:

① we have N variables V_1, \dots, V_n ;

② a valid assignment is $\{V_i = x_i, \dots, V_n = x_n\}$, $0 \leq i \leq n$ where values satisfy the variable constraints.

③ states: valid assignments

④ initial state: empty assignment

⑤ successor: $\{V_i = x_i, \dots, V_n = x_n\} \rightarrow \{V_i = x_i, \dots, V_n = x_n, V_{n+1} = x_{n+1}\}$

⑥ goal test: complete assignment

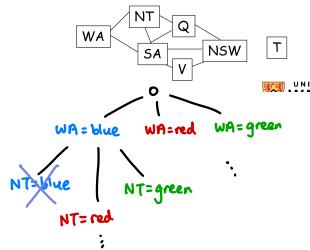
BACKTRACKING SEARCH

```

function BACKTRACKING-SEARCH(csp) returns a solution, or failure
    return RECURSIVE-BACKTRACKING({}, csp)
function RECURSIVE-BACKTRACKING(assignment, csp) returns a solution, or failure
    if assignment is complete then return assignment
    var  $\leftarrow$  SELECT-UNASSIGNED-VARIABLE(Variables[csp], assignment, csp)
    for each value in ORDER-DOMAIN-VALUES(var, assignment, csp) do
        if value is consistent with assignment according to Constraints[csp] then
            add { var = value } to assignment
            result  $\leftarrow$  RECURSIVE-BACKTRACKING(assignment, csp)
            if result  $\neq$  failure then return result
            remove { var = value } from assignment
    return failure
  
```

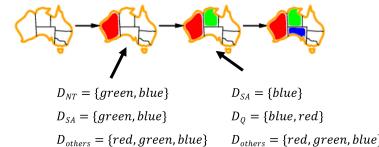
- this is DFS that choose values for one variable at a time
- we "backtrack" when a variable has no legal values to assign

EXAMPLE: MAP COLORING



MOST CONSTRAINED VARIABLE HEURISTIC

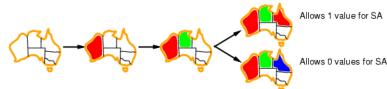
Q: Idea: Choose the variable which has the fewest "legal" moves.



Q: In a tie, choose the variable with the most constraints on the remaining variables.
(ie "most constraining variable").

LEAST CONSTRAINING VALUE HEURISTIC

Q: Idea: Given a variable, choose the "least constraining value", ie the one that rules out the fewest values in the remaining variables.

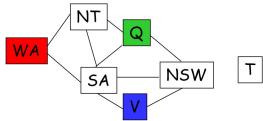


FORWARD CHECKING

Q₁: Idea: We keep track of remaining legal values for unassigned variables, & terminate search when any variable has no legal values.

Q₂: This helps us detect failure early.

EXAMPLE: MAP COLORING



WA	NT	Q	NSW	V	SA	T
RGB						
R	GB	RGB	RGB	RGB	GB	RGB
R	B	G	RB	RGB	B	RGB
R	B	G	RB	B	X	RGB

$$\begin{array}{l} \text{WA} = R \\ \text{Q} = G \\ \text{V} = B \end{array}$$

this is the empty set;
⇒ the current assignment does not lead to a solution.

Chapter 5: Uncertainty

Q₁: Refer to STAT 231/330 for more details.

Q₂: We use \sim to denote the complement of an event (ie $\sim A$).

BAYES RULE

Q₁: For 2 events A, B, note

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Proof: $P(A)P(B|A) = P(A \wedge B) = P(B)P(A|B)$.

$$\therefore P(B|A) = \frac{P(B)P(A|B)}{P(A)}.$$

Q₂: In particular, it allows us to compute a belief about hypothesis B given evidence A.

Q₃: More general forms:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

$$P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A|X)}{P(B|X)}$$

$$P(A=v_i|B) = \frac{P(B|A=v_i)P(A=v_i)}{\sum_{k=1}^n P(B|A=v_k)P(A=v_k)}$$

PROBABILISTIC INFERENCE

Q₁: Idea: Given a prior distribution $P(X)$ over variables X of interest & given new evidence $E=e$ for some variable E, revise our degrees of belief; ie the "posterior" $P(X|E=e)$.

ISSUES

Q₁: Specifying the full joint distribution for X_1, \dots, X_n requires an exponential number of possible "worlds".

Q₂: So, inference is also slow since we need to sum over these exponential number of worlds:

$$P(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n | X_i) = \frac{P(X_1, \dots, X_n)}{\sum_{x_1} \sum_{x_{i-1}} \sum_{x_{i+1}} \dots \sum_{x_n} P(X_1, \dots, X_n)}$$

CONDITIONAL INDEPENDENCE

Q₁: Two variables X, Y are "conditionally independent" given Z if

$$P(X=x | Z=z) = P(X=x | Y=y, Z=z)$$

$$\Leftrightarrow P(X=x, Y=y | Z=z) = P(X=x | Z=z)P(Y=y | Z=z)$$

$$\Leftrightarrow \forall x \in \text{dom}(X), y \in \text{dom}(Y), z \in \text{dom}(Z)$$

Q₂: If we know the value of Z , nothing we learn about Y will influence our beliefs about X .

VALUE OF INDEPENDENCE

Q₁: If X_1, \dots, X_n are mutually independent, then we can specify the full joint distribution using only n parameters (ie linear) instead of $2^n - 1$ (ie exponential).

Q₂: Although most domains do not exhibit complete mutual independence, they do instead exhibit a fair amount of conditional independence.

NOTATION: $P(X)$

Q₁: We define " $P(X)$ " as the marginal distribution over X .

- $P(X=x)$ is a number, $P(X)$ is a distribution.

NOTATION: $P(X|Y)$

Q₁: We define " $P(X|Y)$ " as the family of conditional distributions over X ; one for each $y \in \text{dom}(Y)$.

EXPLOITING CONDITIONAL INDEPENDENCE: CHAIN RULE

Consider a story:

- If Pascal woke up too early E , Pascal probably needs coffee C ; if Pascal needs coffee, he's likely grumpy G . If he is grumpy then it's possible that the lecture won't go smoothly L . If the lecture does not go smoothly then the students will likely be sad S .



E - Pascal woke up too early G - Pascal is grumpy S - Students are sad
 C - Pascal needs coffee L - The lecture did not go smoothly

S is independent of E, C, G given L
 L is independent of E, C given G & so on.

$$\Rightarrow P(S|L, G, C, E) = P(S|L)$$

$$P(L|G, C, E) = P(L|G)$$

$$P(G|C, E) = P(G|C)$$

Then

$$\begin{aligned} P(S, L, G, C, E) &= P(S|L, G, C, E) P(L|G, C, E) P(G|C, E) \cdot \\ &\quad P(C|E) P(E) \\ &= \underline{P(S|L) P(L|G) P(G|C) P(C|E) P(E)}. \end{aligned}$$

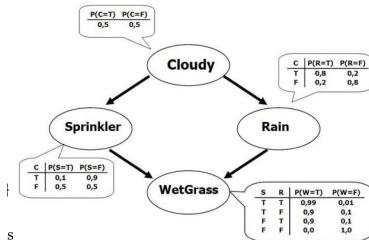
Q: In this, we can specify the full joint distribution by specifying the five local conditional distributions.

Chapter 6:

Bayesian Networks

BAYESIAN / BELIEF / PROBABILISTIC NETWORKS (BN)

B₁ "Bayesian networks" are graphical representations of the direct dependencies over a set of variables, alongside a set of conditional probability tables (CPT) quantifying the strength of the influences.



B₂ In particular, it has

- ① A DAG with nodes = variables X_i , &
- ② A set of CPTs $P(X_i | \text{Parents}(X_i))$ for each X_i .

B₃ Key notions:

- ① parents/children of a node;
- ② ancestors/descendants of a node; &
- ③ family: set of nodes consisting of X_i & its parents.

SEMANTICS OF A BAYES NET

B₁ Idea: Every X_i is conditionally independent of all its non-descendants given its parents; ie

$$P(X_i | S \cup \text{Par}(X_i)) = P(X_i | \text{Par}(X_i))$$

$\forall S \subseteq \text{NonDescendants}(X_i)$

B₂ Also, the joint distribution is recoverable using the parameters (CPTs) in the BN:

$$\begin{aligned} P(x_1, \dots, x_n) &= P(x_n | x_{n-1}, \dots, x_1) P(x_{n-1} | x_{n-2}, \dots, x_1) \dots P(x_1) \\ &= P(x_n | \text{Par}(x_n)) P(x_{n-1} | \text{Par}(x_{n-1})) \dots P(x_1). \end{aligned}$$

CONSTRUCTING A BN

B₁ Idea:

- ① Take any ordering of the variables, and then for X_n to X_1 :
 - let $\text{Par}(X_n)$ be any subset $S \subseteq \{X_1, \dots, X_n\}$ such that X_n is independent of $\{X_1, \dots, X_{n-1}\} - S$ given S .
 - Continue this for X_{n-1}, \dots, X_1 .

② In the end, we get a DAG, which is also a BN by construction.

B₂ Note the order in which we consider the variables changes the resultant BN!

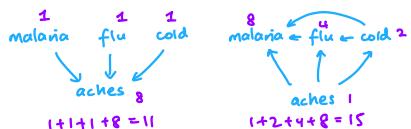
eg
order: mal, cold, flu, aches
malaria → flu → cold
aches ← aches ← aches

order: aches, cold, flu, malaria
malaria ← flu ← cold
aches ← aches ← aches

COMPACTNESS

B₁ In a BN, if each rv is directly influenced by at most k others, then each CPT will have at most 2^k entries.

B₂ So, the entire network of n variables is specified by $n \cdot 2^k$ parameters.



d-SEPARATION

- Q: First, we say a set of variables E "d-separates" X & Y if it "blocks" every undirected path in the BN between X & Y.

TESTING INDEPENDENCE

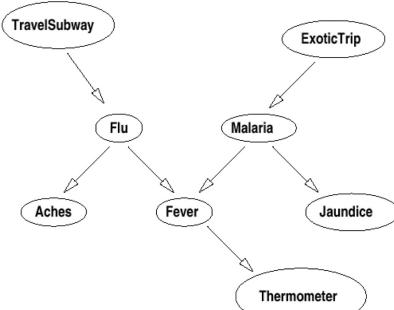
- Q: Then, X & Y are conditionally independent given evidence E if E d-separates X & Y.

BLOCKING IN d-SEPARATION

- Q: Let P be an undirected path from X → Y. Then the evidence set E "blocks" path P if
- ① one arc on P goes into Z & one goes out of Z, & Z ∉ E;
 - ② both arcs on P leave Z & Z ∉ E; or
 - ③ both arcs on P enter Z & neither Z nor any of its descendants are in E.
- $X \rightsquigarrow Z \rightsquigarrow Y$
- $X \leftarrowtail Z \rightarrowtail Y$
- $X \rightsquigarrow Z \rightsquigarrow Y$



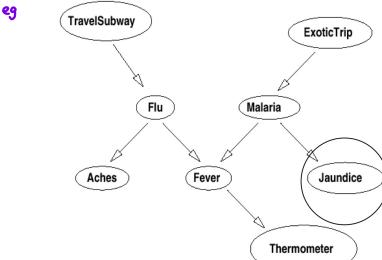
EXAMPLE



- ① subway & thermometer
 - dependent
 - but independent given the flu
 - ↳ since flu blocks the only path (rule 1)
- ② aches & fever
 - dependent
 - but independent given the flu
 - ↳ since flu blocks the only path (rule 2)
- ③ flu & malaria
 - independent
 - dependent given fever or thermometer
 - ↳ rule 3
- ④ subway & exotic trip
 - independent
 - dependent given thermometer
 - ↳ rule 3

SIMPLE FORWARD INFERENCE (CHAIN)

- Q: Idea: To compute the marginal distribution, we can use simple forward "propagation" of probabilities.



$$\begin{aligned}
 P(J) &= \sum_{M, ET} P(J, M, ET) \quad (\text{marginalization}) \\
 &= \sum_{M, ET} P(J|M, ET) P(M|ET) P(ET) \\
 &\quad (\text{chain rule}) \\
 &= \sum_{M, ET} P(J|M) P(M|ET) P(ET) \\
 &\quad (\text{conditional independence}) \\
 &= \sum_M P(J|M) \sum_{ET} P(M|ET) P(ET)
 \end{aligned}$$

all these terms can now be found in the CPTs.

- Q: We can do something similar if we have "upstream" evidence.

$$\begin{aligned}
 \text{eg } P(J|ET) &= \sum_m P(J, m|ET) \\
 &= \sum_m P(J|m, ET) P(m|ET) \\
 &= \sum_m P(J|m) P(m|ET)
 \end{aligned}$$

terms found in CPTs

SIMPLE BACKWARD INFERENCE

- Q: Idea: For "downstream" evidence, we must reason backwards, which we can use Bayes' rule:

$$\begin{aligned}
 \text{eg (using same BN as above)} \quad P(LET|J) &= \alpha P(J|ET) P(ET), \quad \alpha = \frac{1}{P(J)} \\
 &= \alpha \sum_m P(J, m|ET) P(ET) \\
 &= \alpha \sum_m P(J|m, ET) P(m|ET) P(ET) \\
 &= \alpha \sum_m P(J|m) P(m|ET) P(ET).
 \end{aligned}$$

We can then calculate $P(J) = \sum_{ET} P(J|ET) P(ET)$.

VARIABLE ELIMINATION

The "variable elimination" algorithm is a general inference tool for BNs.

FACTORS

A "factor" is a function $f(X_1, \dots, X_k)$.

We can represent factors as a table of numbers, one for each instantiation of the variables X_1, \dots, X_k .

We denote $f(X, Y)$ to be a factor over the variables $X \cup Y$, where X & Y are sets of variables.

Note each CPT in a Bayes net is a factor of its family.

eg $P(C|A, B) \rightarrow$ factor of A, B, C .

PRODUCT OF FACTORS: fg

Let $f(X, Y), g(Y, Z)$ be factors with variables Y in common.

Then the "product" of f & g , $h=fg$, is defined to be

$$h(X, Y, Z) = f(X, Y) \times g(Y, Z)$$

eg

$f(A, B)$	$g(B, C)$	$h(A, B, C)$
ab 0.9	bc 0.7	abc 0.63
a~b 0.1	b~c 0.3	a~bc 0.02
~ab 0.4	~bc 0.2	~a~bc 0.28
~a~b 0.6	~b~c 0.8	~a~b~c 0.12

SUM VARIABLE OUT OF A FACTOR: $\sum_x f$

Let $f(X, Y)$ be a factor, where X is a variable & Y is a variable set.

Then, we can "sum out" variable X from f to produce a new factor $h = \sum_x f$, where

$$h(Y) = \sum_{x \in \text{Dom}(X)} f(x, Y).$$

eg

	$f(A, B)$	$h(B)$
ab 0.9	b 1.3	
a~b 0.1	~b 0.7	
~ab 0.4		
~a~b 0.6		

RESTRICTING FACTORS: $f|_{X=x}$

Let $f(X, Y)$ be a factor with variable X & variable set Y .

Then, we "restrict" factor f to $X=x$, ie $h=f|_{X=x}$, by doing

$$h(Y) = f(x, Y).$$

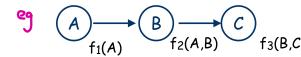
eg

	$f(A, B)$	$h(B) = f _{A=a}$
ab 0.9	b 0.9	
a~b 0.1	~b 0.1	
~ab 0.4		
~a~b 0.6		

NO EVIDENCE CASE

Idea: Computing prior probability of the query variable X can be seen as applying these operations on factors.

EXAMPLE 1

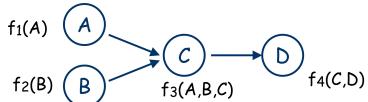


$$\begin{aligned} P(C) &= \sum_{A,B} P(C|B) P(B|A) P(A) \\ &= \sum_B P(C|B) \sum_A P(B|A) P(A) \\ &= \sum_B f_3(B, C) \sum_A f_2(A, B) f_1(A) \\ &\quad \text{---} \\ &= \sum_B f_3(B, C) f_4(B) \\ &\quad \text{---} \\ &= f_5(C). \end{aligned}$$

Numerical example:

	$f_1(A)$	$f_2(A, B)$	$f_3(B, C)$	$f_4(B)$	$f_5(C)$
a 0.9	ab 0.9	bc 0.7	b 0.85	c 0.625	
a~b 0.1	a~b 0.1	b~c 0.3	~b 0.15	~c 0.375	
~ab 0.4		~bc 0.2			
~a~b 0.6		~b~c 0.8			

EXAMPLE 2



eg $P(D) = \sum_{A,B,C} P(D|C) P(C|B,A) P(B) P(A)$

 $= \sum_C P(D|C) \sum_B P(B) \sum_A P(C|B,A) P(A)$
 $= \sum_C f_4(C,D) \sum_B f_2(B) \sum_A f_3(A,B,C) f_1(A)$
 $= \sum_C f_4(C,D) \sum_B f_2(B) f_5(B,C)$
 $= \sum_C f_4(C,D) f_6(C)$
 $= f_7(D)$

*define f_5, f_6, f_7
according to the
brackets

ALGORITHM (NO EVIDENCE)

Input: query var Q , remaining vars Z , & $F = \text{set of factors corresponding to CPTs}$ for $\{Q\} \cup Z$.

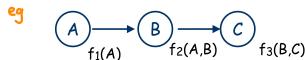
- Choose an elimination ordering Z_1, \dots, Z_n of variables in Z .
- For each Z_j -- in the order given -- eliminate $Z_j \in Z$ as follows:
 - Compute new factor $g_j = \sum_{Z_j} f_1 \times f_2 \times \dots \times f_k$, where the f_i are the factors in F that include Z_j
 - Remove the factors f_i (that mention Z_j) from F and add new factor g_j to F
- The remaining factors refer only to the query variable Q . Take their product and normalize to produce $P(Q)$

eg query: $P(D)$
elim. order: A, B, C

Steps:

- add $f_5(B,C) = \sum_A f_3(A,B,C) f_1(A)$;
remove $f_1(A), f_3(A,B,C)$
- add $f_6(C) = \sum_B f_2(B) f_5(B,C)$
- we don't need to sumout f_3 as we removed it in step ①
remove $f_2(B), f_5(B,C)$
- add $f_7(D) = \sum_C f_4(C,D) f_6(C)$
remove $f_4(C,D), f_6(C)$
- The remaining factor $f_7(D)$ is our (possibly unnormalized) probability $P(D)$

EVIDENCE CASE



eg $P(A|C=c) = \alpha P(A) P(C=c|B) P(B|A)$ (Bayes' thm)

 $= \alpha P(A) \sum_B P(C=c|B) P(B|A)$
 $= \alpha f_1(A) \sum_B f_3(B,c) f_2(A,B)$
 $= \alpha f_1(A) \sum_B f_4(B) f_2(A,B)$
 $= \alpha f_1(A) f_5(A)$
 $= f_6(A)$

ALGORITHM (WITH EVIDENCE)

Input: Given query var Q , evidence vars E (observed to be e), remaining vars Z & set of factors involving CPTs for $\{Q\} \cup Z$, F :

- Replace each factor $f \in F$ that mentions a variable(s) in E with its restriction $f|_{E=e}$ (somewhat abusing notation)
- Choose an elimination ordering Z_1, \dots, Z_n of variables in Z .
- For each Z_j -- in the order given -- eliminate $Z_j \in Z$ as follows:
 - Compute new factor $g_j = \sum_{Z_j} f_1 \times f_2 \times \dots \times f_k$, where the f_i are the factors in F that include Z_j
 - Remove the factors f_i (that mention Z_j) from F and add new factor g_j to F
- The remaining factors refer only to the query variable Q . Take their product and normalize to produce $P(Q)$

eg same example
Query: $P(A|D=d)$

- replace $f_4(C,D)$ with $f_5(C) = f_4(C,d)$
- proceed similar to before

ANALYSIS

\bullet_1 After eliminating z_j , the factors remaining in set F refer only to x_{j+1}, \dots, z_n & Q .

\bullet_2 Also, no factor mentions any evidence variable E after the initial restriction.

\bullet_3 Note

- ① The number of iterations is linear in the # of variables;
- ② The complexity is exponential in the # of variables.

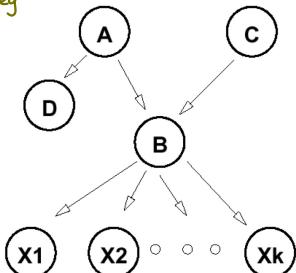
POLYTREES

\bullet_1 Polytrees are basically "trees" (ie no undirected cycles) that can have multiple start nodes.

\bullet_2 Idea: In these, the inference is linear wrt the size of the network.

\bullet_3 To do this, we eliminate only "singly-connected nodes".

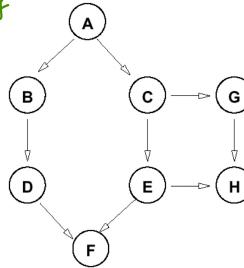
eg



- eliminate D, A, C, X₁, ..., X_k
- if we eliminate B before these, we get factors that include all of A, C, X₁, ..., X_k!

LEAST NEIGHBORS HEURISTIC

eg



- A, F, H, G, B, C, E is good
- ↳ we eliminate the nodes with 2 neighbors first
- ↳ leaving the nodes with 3 neighbors at the end.
- if we started with B, the ordering would be bad since the size of the factors is larger.

\bullet_1 Idea: When choosing an ordering, prioritize nodes with the least number of neighbors.

RELEVANCE

\bullet_1 Motivation: Certain variables have no impact on the query.

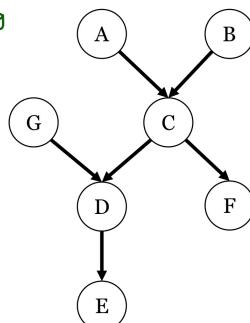
eg $A \rightarrow B \rightarrow C$

- To calculate $P(A)$, we only need to look at A's CPT!
- However, if we do var elimination, we get trivial factors (ie whose value is just 1)

\bullet_2 Thus, when considering variables, we can restrict our attention to only the "relevant" ones;

- ie given query Q & evidence E :
- ① Q is relevant;
- ② If Z is relevant, $\text{Parents}(Z)$ are relevant; &
- ③ If $E \in E$ is a descendant of a relevant node, then E is relevant.

eg



① $Q = P(F)$
relevant: F, C, B, A

② $Q = P(C|F, E)$
relevant: F, C, B, A, E, D, Q

③ $Q = P(C|F, E, C)$
relevant: whole graph, but really none except C, F since C cuts off all influence of others.

Chapter 7: Causal Inference

CAUSALITY

"Causality" is the study of how things influence each other & how causes lead to effects.

CAUSAL DEPENDENCE

We say " X causes Y " if changes to X induce changes in Y .

Note joint distributions captures correlations between X & Y , not causations.

- $P(Y|X) \neq X \text{ causes } Y$

CAUSAL BAYESIAN NETWORK

A "causal Bayesian network" is one where all edges indicate direct causal effects.



CAUSAL INFERENCE

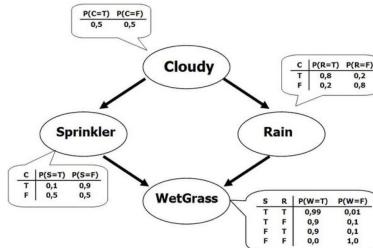
Causal networks can solve "intervention queries"; ie what the effect of an action is.
- but non-causal networks cannot.

OBSERVATION VS INTERVENTION

Observational queries are in the form "What is the likelihood of Y given X ?" ie $P(Y|X=x)$.

Interventional queries are in the form "How does doing X affect Y ?"
ie $P(Y|\text{do}(X=x))$.
- the "do" keyword specifies the query is an intervention.

EXAMPLE: CAUSAL GRAPH



observational: $P(WG|S=\text{true})$

- factors: $P(C), P(R|C), P(S|C), P(WG|S, R)$

- evidence: $S = \text{true}$

- eliminate: C, R

interventional: $P(WG|\text{do}(S=\text{true}))$

- we can remove the CPT from S since we "explicitly set" $S = \text{true}$.

- factors: $P(C), P(R|C), P(WG|S, R)$

- evidence: $S = \text{true}$

- eliminate: C, R

INFERENCE WITH THE DO OPERATOR

To do inference for $P(X|\text{do}(Y=y), Z=z)$:

① Remove edges pointing to Y &

$P(Y|\text{Parents}(Y))$

② Perform variable elimination as usual
(evidence is $Y=y, Z=z$).

COUNTER-FACTUAL ANALYSIS

- "Counter-factual analysis" explores outcomes that did not occur, but could have occurred under different conditions.
 - basically a "what-if?" analysis
 This can help test causal relationships.
 eg "would the patient have died if he was not treated"

STRUCTURAL CAUSAL MODEL / SCM

- Idea: We want to separate causal relations from "noise".

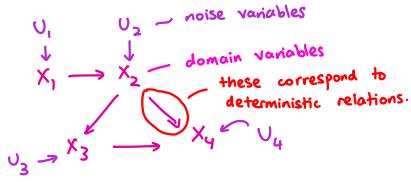
A "structural causal model" consists of

- ① X : endogenous/domain variables
- ② U : exogenous variables/noise
- ③ Only deterministic relations given by equations in the form

$$X_i = f(\text{parents}(X_i), U_i)$$

where U_i corresponds to the noise variable associated with X_i .

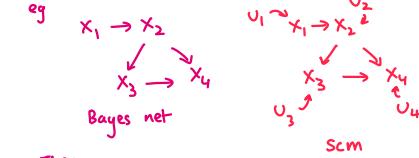
eg



$$X_1 = f_1(U_1), \quad X_2 = f_2(X_1, U_2).$$

$$X_3 = f_3(X_2, U_3), \quad X_4 = f_4(X_3, X_2, U_4)$$

- We can convert SCMs to causal Bayesian networks, but not v.v.



Then

$$P(X_1) = \sum_{U_1} P(U_1) f(X_1 | U_1)$$

$$P(X_2 | X_1) = \sum_{U_2} P(U_2) f(X_2 | X_1, U_2)$$

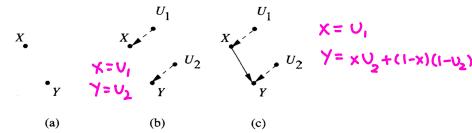
- SCMs are more descriptive since they separate causal relations from noise.

METHOD

For a causal model M , to find $P(Y=y | e, \text{do}(X=x))$:

- ① Update $P(u)$ to find $P(u|e)$ (abduction);
 - u = noise variables
- ② Replace the equations corresponding to variables in set X by the equations $X=x$ (action); &
- ③ Use the modified model to calculate $P(Y=y)$.

EXAMPLE



Model 1

	$u_2 = 0$	$u_2 = 1$	Marginal
$x = 1$	$x = 0$	$x = 1$	$x = 0$
$y = 1$ (death)	0	0.25	0.25
$y = 0$ (recovery)	0.25	0	0.25

Model 2

	$u_2 = 0$	$u_2 = 1$	Marginal
$x = 1$	$x = 0$	$x = 1$	$x = 0$
$y = 1$ (death)	0	0.25	0.25
$y = 0$ (recovery)	0.25	0	0.25

model B:

$$\begin{array}{ccc} X & \leftarrow & U_1 \\ & \downarrow & \\ Y & \leftarrow & U_2 \end{array}$$

evidence: $X=\text{true}, Y=\text{true}$
 $\Rightarrow P(U_2=1 | \text{evidence}) = 1.$
 Then
 $Y = U_2 = 1.$

model C:

$$\begin{array}{ccc} X & \leftarrow & U_1 \\ & \downarrow & \\ Y & \leftarrow & U_2 \end{array}$$

evidence: $X=\text{true}, Y=\text{true}$
 $\Rightarrow P(U_2=1 | \text{evidence}) = 1.$
 Then

$$Y = XU_2 + (1-X)(1-U_2)$$

$$= 0(1) + (1-0)(1-1)$$

$$= 0.$$

Chapter 8: Reasoning Over Time

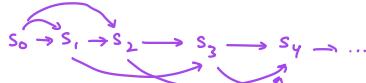
STATIC VS DYNAMIC INFERENCE

- Q₁ So far, we have assumed "static inference"; ie the world does not change.
- Q₂ However, we need to perform "dynamic inference" in the real world since the world evolves over time.
- Q₃ In particular, we need
- ① A set of all possible states/worlds;
 - ② A set of time-slices/snapshots;
 - ③ Different probability distributions for each state at each time-slice; &
 - ④ Dynamics encoding how distributions change over time.

STOCHASTIC PROCESS

- Q₁ A "stochastic process" is defined by
- ① a set of states S ;
 - ② some stochastic dynamics $P(s_t | s_{t-1}, \dots, s_0)$.

eg



- Q₂ This is a Bayes net with 1 r.v. per time slice.

Q₃ Problems:

- ① We may have infinite variables; and so
 - ② We may have infinitely large conditional probability tables.
- Q₄ To solve this, we will assume
- ① "Stationary process": dynamics do not change over time; ie the CPT is the same regardless of the time step.
 - ② "Markov assumption": current state depends only on a finite history of past states.

K-ORDER MARKOV PROCESS

- Q₁ Idea: The last k states are sufficient for inference.

eg - first-order: $P(s_t | s_{t-1}, \dots, s_0) = P(s_t | s_{t-1})$

$$s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow \dots$$

- second-order: $P(s_t | s_{t-1}, \dots, s_0) = P(s_t | s_{t-1}, s_{t-2})$

$$\begin{matrix} s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow \dots \\ \searrow \quad \nearrow \\ \end{matrix}$$

- Q₂ Advantage: we can specify the entire process with finitely many time slices.

eg for 1st order: $s_{t-1} \rightarrow s_t$

- dynamics: $P(s_t | s_{t-1})$

- prior: $P(s_0)$

HIDDEN MARKOV MODELS

- Q₁ Motivation: In general,

- ① States are not directly observable;
- ② Uncertain dynamics increase state uncertainty; but
- ③ Observations made from sensors reduce state uncertainty.

- Q₂ A "Hidden Markov model" encapsulates this and includes

- ① a set of states S ;
- ② a set of observations O ;
- ③ a transition model $P(s_t | s_{t-1}, \dots, s_0)$;
- ④ an observation model $P(o_t | s_{t-1}, \dots, s_0)$;
- ⑤ a prior $P(s_0)$.

eg 1st order HMM:

$$\begin{matrix} s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4 \\ \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ o_1 \quad o_2 \quad o_3 \quad o_4 \end{matrix}$$

- $P(s_t | s_{t-1})$: State transition with uncertainty

- $P(o_t | s_t)$: uncertainty in measurements from sensors

INFERENCE IN TEMPORAL MODELS

Q₁: We have 4 common tasks:

- ① Monitoring: $P(s_t | o_t, \dots, o_1)$
- ② Prediction: $P(s_{t+k} | o_t, \dots, o_1)$
- ③ Hindsight: $P(s_k | o_t, \dots, o_1)$, $k < t$
- ④ Most likely explanation: $\underset{s_0, \dots, s_t}{\operatorname{argmax}} P(s_0, \dots, s_t | o_t, \dots, o_1)$

Q₂: We can solve ①-③ using variable elimination & ④ with a variant.

MONITORING

Q₁: Idea: We want to compute

$$P(s_t | o_t, \dots, o_1).$$

i.e. the distribution of the current state given observations.

Q₂: We can solve this using the "forward algorithm", which corresponds to variable elimination:

1. Factors: $P(s_0)$, $P(s_i | s_{i-1})$, $P(o_i | s_i)$, $1 \leq i \leq t$
2. Restrict o_1, \dots, o_t to observations made
3. Sumout s_0, \dots, s_{t-1} ; i.e.

PREDICTION

Q₁: Goal: we want to compute

$$P(s_{t+k} | o_t, \dots, o_1);$$

i.e. the distribution over future state given observations.

Q₂: We can also use the forward algorithm:

1. Factors: $P(s_0)$, $P(s_i | s_{i-1})$, $P(o_i | s_i)$, $1 \leq i \leq t+k$
2. Restrict o_1, \dots, o_t to observations made
3. Sumout s_0, \dots, s_{t+k-1} , o_{t+1}, \dots, o_{t+k}

HINDSIGHT

Q₁: Goal: we want to compute

$$P(s_k | o_t, \dots, o_1)$$

Q₂: We can use "forward-backward algorithm" to solve this:

1. Factors: $P(s_0)$, $P(s_i | s_{i-1})$, $P(o_i | s_i)$, $1 \leq i \leq t+k$
2. Restrict o_1, \dots, o_t
3. Sumout s_0, \dots, s_{k-1} , s_{k+1}, \dots, s_t

MOST LIKELY EXPLANATION

Q₁: Goal: We want to compute

$$\underset{s_0, \dots, s_t}{\operatorname{argmax}} P(s_0, \dots, s_t | o_t, \dots, o_1).$$

Q₂: We can use the "Viterbi algorithm" to solve this:

1. Factors: $P(s_0)$, $P(s_i | s_{i-1})$, $P(o_i | s_i)$, $1 \leq i \leq t$
2. Restrict o_1, \dots, o_t
3. "Maxout" s_0, \dots, s_t

COMPLEXITY OF TEMPORAL INFERENCE

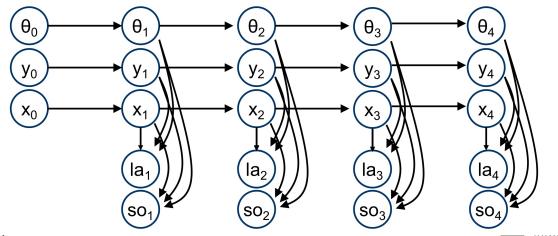
Q₁: HMMs are Bayes nets with a polytree structure.

Q₂: Thus, variable elimination is

- ① Linear wrt # of time slices;
- ② Linear wrt the largest CPT.

DYNAMIC BAYESIAN NETWORKS

- Idea: Encode states & observations with several random variables, and exploit conditional independence to save time & space.



- This allows us to write the transition and observation models very compactly.

NON-STATIONARY PROCESS

- If the process is not stationary, we can add new state components until dynamics are stationary.

NON-MARKOVIAN PROCESS

- If the process is not Markovian, we can add new state components until dynamics are Markovian.

- However, note this may significantly increase computational complexity.

- so we should find the smallest state description that is Markovian & stationary.

Chapter 9: Decision Tree Learning

INDUCTIVE LEARNING

Q₁: Idea: Given a training set of examples of the form $(x, f(x))$, return a "hypothesis" function h that approximates f .

Q₂: Types:
① Classification; &
② Regression.

HYPOTHESIS SPACE

Q₃: The "hypothesis space" is the set of all hypotheses h that the learner may consider.

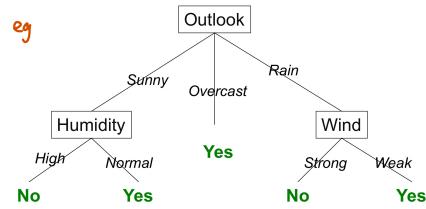
REALIZABLE

Q₁: we say a learning problem is "realizable" if the hypothesis space contains the true function.
Q₂: We can use a large hypothesis space, but there is a tradeoff between the expressiveness of a hypothesis class & the complexity of finding a simple, consistent hypothesis within the space.

DECISION TREES

Q₁: A decision tree contains
① Nodes, labelled with attributes;
② Edges, labelled with attribute values;
&
③ Leaves, labelled with classes.

Q₂: Idea: Classify an instance by starting at the root, testing the attribute specified by the root, then moving down the branch corresponding to the value of the attribute; we continue this until we reach a leaf, then which we return the class.



Q₃: We can express any boolean function as a decision tree.
- but some functions require exponentially large trees

INDUCING A DECISION TREE

B1 Idea: We find a small tree consistent with the training examples by recursively choosing the most significant attribute as the root of the subtree.

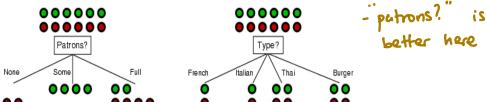
B2 Algorithm:

```
function DTL(examples, attributes, default) returns a decision tree
    if examples is empty then return default
    else if all examples have the same classification then return the classification
    else if attributes is empty then return MODE(examples)
    else
        best ← CHOOSE-ATTRIBUTE(attributes, examples)
        tree ← a new decision tree with root test best
        for each value  $v_i$  of best do
            examplesi ← {elements of examples with best =  $v_i$ }
            subtree ← DTL(examplesi, attributes - best, MODE(examplesi))
            add a branch to tree with label  $v_i$  and subtree subtree
        return tree
```

CHOOSING AN ATTRIBUTE

B1 In particular, at each iteration, we want to choose an attribute that is most useful for classifying examples.

B2 Ideally, a good attribute is one that splits the examples into either "all positive" or "all negative".



B3 For a training set with p positive examples & n negative examples, the "entropy" is

$$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right).$$

B4 Then, if an attribute A divides the training set E into subsets E_1, \dots, E_v according to their values for A , where A has v distinct values, then the "remainder" of A is

$$\text{remainder}(A) = \sum_{i=1}^v \frac{p_i + n_i}{p+n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

B5 Finally, the "information gain" (IG) of A is

$$\text{IG}(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - \text{remainder}(A)$$

B6 We choose the attribute with the largest IG.

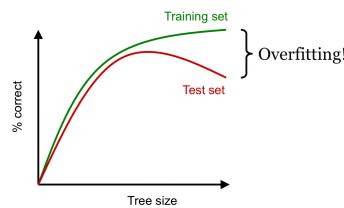
PERFORMANCE OF A LEARNING ALGORITHM

B we can verify the performance of a learning algorithm by using a test set, which are examples the algorithm did not see during training.

OVERRFITTING

B1 We say a hypothesis $h \in H$ "overfits" the training data if there exists some alternative hypothesis $h' \in H$ such that

- ① h' has smaller error than h over the training examples; but
- ② h' has smaller error than h over the entire distribution of instances.



B2 Overfitting can occur if

- ① the data is noisy; or
- ② the training set is too small to give a representative sample of the target function.

B3 To avoid overfitting, we can

- ① prune statistically irrelevant nodes; or
- ② stop growing tree when the test set performance decreases, using "cross-validation".

CHOOSING TREE SIZE

Q: However, since we are now choosing the tree size based on the test set, it becomes part of the training set when optimizing the tree size.

Q: So, we cannot trust the test set to be representative of future accuracy.

Q: Solution: we split the data into
① training set: compute the decision tree;
② validation set: optimize hyperparameters
eg tree size

③ test set: measure performance

Q: Choosing tree size based on the validation set:

```
Let TS be the Tree Size
For TS = 1 to max value
    decisionTreeTS ← train(TS, trainingData)
    accuracyTS ← eval(decisionTreeTS, validationData)
TS* ← argmaxTS accuracyTS
decisionTreeTS* ← train(TS*, trainingData ∪ validationData)
accuracy ← eval(decisionTreeTS*, testData)
Return k*, accuracy
```

```
eval(decisionTree, dataset)
    correct ← 0
    For each (x, y) ∈ dataset
        if y = decisionTree(x) then correct ← correct + 1
    accuracy ←  $\frac{\text{correct}}{|\text{dataset}|}$ 
    return accuracy
```

CROSS-VALIDATION

Q: Idea: Repeatedly split the training data into two parts; one for training and one for validation, and then report the average validation accuracy.

eg k=4



exp #

#1

#2

#3

#4

↳ then take the average of the validation accuracy.

Q: This ensures the validation accuracy is representative of future accuracy.

Q: In "k-fold cross validation", we split the training data into k equal size subsets, and run k experiments, each time validating on one subset & training on the remaining subsets.

Then, we report the average validation accuracy of the k experiments.

Q: Selecting tree size via cross-validation:

Let TS be the Tree Size

Let k be the number of trainData splits

For TS = 1 to max value

For i = 1 to k do (where i indexes trainData splits)
 decisionTree_{TS} ← train(TS, trainData_{1..i-1,i+1..k})
 accuracy_{TS,i} ← eval(decisionTree_{TS}, trainData_i)
 accuracy_{TS} ← average(accuracy_{TS,i})_{forall i}

```
TS* ← argmaxTS accuracyTS
decisionTreeTS* ← train(TS*, trainData1..i-1,i+1..k)
accuracy ← eval(decisionTreeTS*, testData)
Return TS*, accuracy
```

Chapter 10: Statistical Learning

Q₁ Idea: We have uncertain knowledge about the world, & learning reduces this uncertainty.

Q₂ In particular, we have our hypotheses H : our probabilistic theories of the world; &

① hypotheses H : our probabilistic theories of the world; &

② data D ; our evidence about the world.

BAYESIAN LEARNING

Q₁ "Bayesian learning" consists of

- ① the prior $P(H)$;
- ② the likelihood $P(d|H)$; &
- ③ our evidence $d = \{d_1, \dots, d_n\}$,

and we want to compute

$$P(H|d) = k P(d|H)P(H)$$

via Bayes' theorem.

Q₂ To predict an unknown quantity X , we can use

$$\begin{aligned} P(X|d) &= \sum_i P(X|d, h_i)P(h_i|d) \\ &= \sum_i P(X|h_i)P(h_i|d) \end{aligned}$$

EXAMPLE: CANDY

▪ Favorite candy sold in two flavors:

- Lime (hugh)
- Cherry (yum)

▪ Same wrapper for both flavors

▪ Sold in bags with different ratios:

- 100% cherry $\rightarrow h_1$
- 75% cherry + 25% lime $\rightarrow h_2$
- 50% cherry + 50% lime $\rightarrow h_3$
- 25% cherry + 75% lime $\rightarrow h_4$
- 100% lime $\rightarrow h_5$

} our hypotheses

Assume prior is

$$P(H) = \{0.1, 0.2, 0.4, 0.2, 0.1\}$$

If we assume candies are "iid":

$$P(d|h_i) = \prod_j P(d_j|h_i)$$

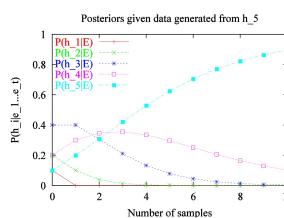
Suppose first 10 candies all taste lime:

$$\Rightarrow P(d|h_5) = 1^{10} = 1.$$

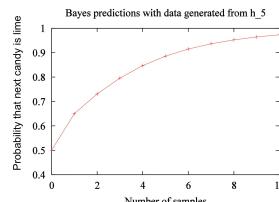
$$\Rightarrow P(d|h_3) = 0.5^{10} \approx 0.0001$$

$$\Rightarrow P(d|h_1) = 0^{10} = 0$$

Posterior:



Prediction:



Bayesian Learning Properties

Properties:

- ① Optimal: given prior, no other prediction is correct more often than the Bayesian one
- ② No overfitting: all hypotheses weighted & considered.

But when the hypothesis space is large, Bayesian learning may be intractable.

MAXIMUM A POSTERIORI / MAP

Idea: Make our prediction based on the most probable hypothesis h_{MAP} ; ie

$$h_{MAP} = \operatorname{argmax}_h P(h|d)$$

This "approximates" Bayesian learning.

eg candy example

1 lime: $h_{MAP} = h_3$, $P(\text{lime}|h_{MAP}) = 0.5$

2 limes: $h_{MAP} = h_4$, $P(\text{lime}|h_{MAP}) = 0.75$

3 limes: $h_{MAP} = h_5$, $P(\text{lime}|h_{MAP}) = 1$

etc.

However, the prediction from MAP is less accurate than the Bayesian prediction since it relies on only one hypothesis h_{MAP} .

It also has "controlled overfitting" (prior can be used to penalize complex hypotheses).

Also, finding h_{MAP} may be an intractable optimization problem!

$$\begin{aligned} h_{MAP} &= \operatorname{argmax}_h P(h|d) \\ &= \operatorname{argmax}_h P(h) P(d|h) \\ &= \operatorname{argmax}_h P(h) \prod_i P(d_i|h) \\ &= \operatorname{argmax}_h (\log P(h) + \sum_i \log P(d_i|h)) \end{aligned}$$

MAXIMUM LIKELIHOOD / ML

Idea: Simplify MAP by assuming the priors are uniform i.e. $P(h_i) = P(h_j) \forall i, j$, and let

$$h_{ML} = \operatorname{argmax}_h P(d|h)$$

and make our prediction based on h_{ML} only:

$$P(X|d) \approx P(X|h_{ML})$$

Properties:

① Less accurate than Bayesian & MAP; but ML, MAP & Bayesian predictions converge as data increases.

② Subject to overfitting.

Finding h_{ML} is easier than finding h_{MAP} :

$$h_{ML} = \operatorname{argmax}_h \sum_i \log P(d_i|h)$$

STATISTICAL LEARNING

Note,

- ① if the data is known, ie all attributes are known, then learning is easy.
- ② if the data is unknown, then learning is harder.

EXAMPLE 1: CANDY 1

- hypothesis h_0 : $P(\text{cherry})=0$, $P(\text{lime})=1-\theta$.

- data d: c cherries, l limes

ML hypothesis: θ is relative freq. of observed data

↳ $\theta = \frac{c}{c+l}$, $P(\text{cherry}) = \frac{c}{c+l}$, $P(\text{lime}) = \frac{l}{c+l}$
Then

$$\hookrightarrow P(d|h_0) = \theta^c (1-\theta)^l$$

$$\Rightarrow \log P(d|h_0) = c \log \theta + l \log (1-\theta)$$

$$\Rightarrow \frac{d \log P(d|h_0)}{d\theta} = \frac{c}{\theta} - \frac{l}{1-\theta}$$

Set this to 0 to find optimal θ : $\Rightarrow \theta = \frac{c}{c+l}$

EXAMPLE 2: CANDY 2

BN:	flavor	$P(F=\text{cherry}) = \theta$
	wrapper	$P(W=\text{red} F)$
	F	θ_1
	C	θ_2
	R	θ_3

Hypothesis: $h_{\theta_1, \theta_2, \theta_3}$

Data: - c cherries; r_c green wrappers, r_r red wrappers

- R times; r_g green wrappers, r_r red wrappers.

Then

$$L(\theta, \theta_1, \theta_2) = P(d|h_{\theta, \theta_1, \theta_2}) = \theta^c (1-\theta)^R \theta_1^{r_g} (1-\theta_1)^{R-r_g} \theta_2^{r_r} (1-\theta_2)^{R-r_r}$$

Getting $L(\theta, \theta_1, \theta_2)$, and setting $\frac{\partial L}{\partial (\theta, \theta_1, \theta_2)} = 0$, we get

$$\theta = \frac{c}{c+r_r}, \quad \theta_1 = \frac{r_g}{r_g+r_r}, \quad \theta_2 = \frac{r_r}{r_g+r_r}.$$

LAPLACE SMOOTHING

Idea: If there is no sample for a certain outcome, we may get overfitting.

e.g. no cherries eaten so far

$$\hookrightarrow P(\text{cherry}) = \theta = \frac{c}{c+0} = 0$$

\hookrightarrow this is dangerous since it rules out outcomes.

To solve this, we employ "Laplace (add-one) smoothing", where we add one to all counts.

$$\text{e.g. } P(\text{cherry}) = \theta = \frac{c+1}{c+r_r+2} (> 0).$$

NAIVE BAYES MODEL

Idea: we want to predict a class based on attributes A_i .



Parameters:

$$\textcircled{1} \quad \theta = P(C=\text{true})$$

$$\textcircled{2} \quad \theta_{i1} = P(A_i=\text{true} | C=\text{true})$$

$$\textcircled{3} \quad \theta_{i2} = P(A_i=\text{true} | C=\text{false})$$

Assumption: A_i 's are independent given C .

Note Naive Bayes models usually don't perform as well as decision tree models since the latter does not assume conditional independence of the attributes.

Parameter learning:

$$\textcircled{1} \quad \text{Parameters: } \theta_{V, \text{pac}(V)=v}$$

$$- \theta_{V, \text{pac}(V)=v} = P(V|\text{pac}(V)=v)$$

- we can get this from the CPFs

\textcircled{2} Data d:

$$d_i = \langle V_1=v_{1,i}, \dots, V_n=v_{n,i} \rangle$$

\textcircled{3} Max likelihood:

$$\widehat{\theta}_{V, \text{pac}(V)=v} = \frac{\#\{V, \text{pac}(V)=v\}}{\#\{\text{pac}(V)=v\}}$$

- $\text{pac}(V)$ = parents of V

Chapter III: Neural Networks

ARTIFICIAL NEURAL NETWORKS

- Q: Idea: Mimic the brain to do computation;
in particular:
① Nodes correspond to neurons;
② Links correspond to synapses (links).

UNIT

- Q: For each unit i , it has
① Weights, w — refers to the strength of the link from unit i to unit j :

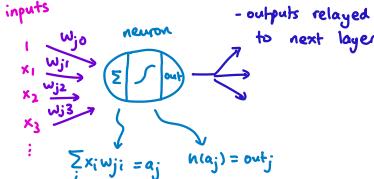
$$a_j = \sum_i w_{ji} x_i + w_{j0} = w_j \bar{x}$$

- ② Activation function, h — corresponds to the numerical signal produced:

$$y_j = h(a_j).$$

- h should be non-linear

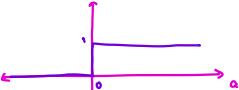
Picture:



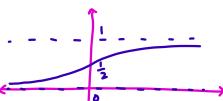
- Q: Note the unit should be "active" (ie output near 1) when fed with the "right" inputs, and "inactive" (output near 0) when fed with the "wrong" inputs.

COMMON ACTIVATION FUNCTIONS

- Q: "Threshold" function:



- Q: "Sigmoid" function:



LOGIC GATES

- Q: Idea: We want to design ANNs to represent boolean functions:

- ① AND:

$$\begin{array}{l} w_0 \\ x_1 w_1 \\ x_2 w_2 \\ \hline \text{threshold} \end{array}$$

$$a = w_0(1) + w_1(x_1) + w_2(x_2).$$

$$h(a) = \begin{cases} 0, & a \leq 0 \\ 1, & \text{otherwise.} \end{cases}$$

We can use $w_0 = -1.5$
 $w_1 = w_2 = 1$.

- ② OR:

↳ we can use $w_0 = -0.5$, $w_1 = w_2 = 1$.

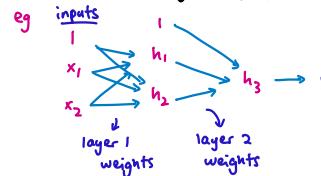
- ③ NOT:

↳ we can use $w_0 = 0$, $w_1 = -1$.

NETWORK STRUCTURES

- Q: Types:

- ① Feed-forward network: consists of a directed acyclic graph.



- ② Recurrent network: consists of a directed cyclic graph.

- can memorize information

PERCEPTRON

- Q: A "perceptron" is a single layer feed-forward network.

- Q: Note a perceptron is a linear separator.

MULTILAYER NETWORKS

- Q: Idea: Neural networks with ≥ 1 hidden layer of sufficiently many sigmoid units can approximate any function arbitrarily closely.

(see slides for idea)

WEIGHT TRAINING

- Q₁: Our parameters are the weights in the layers $\langle W^{(1)}, W^{(2)}, \dots \rangle$.
- Q₂: Idea: We want to minimize the errors.
- Q₃: To do this, we can use backpropagation.

LEAST SQUARED ERROR

- Q₁: Our loss/error function is

$$E(W) = \frac{1}{2} \sum_n E_n(W)^2 = \frac{1}{2} \sum_i \|f(x_i; W) - y_i\|_2^2$$

We want to minimize this.

- Q₂: To do this, we can use sequential gradient descent:

$$w_{ji} \leftarrow w_{ji} - \eta \cdot \frac{\partial E_n}{\partial w_{ji}}$$

- Q₃: To compute the gradient efficiently, we can use backpropagation, or in reality, automatic differentiation.

BACKPROPAGATION ALGORITHM

- Q₁: First phase: forward phase - compute output z_j for each unit j .



$$- z_j = h(a_j), \quad a_j = \sum_i w_{ji} z_i$$

- Q₂: Second phase: "backward phase" - compute δ_j at each unit j .

- For each w_{ji} : $\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{ji}} = \delta_j z_i$,
 $\delta_j = \frac{\partial E_n}{\partial a_j}$.

- Then

$$\delta_j = \begin{cases} h'(a_j)(z_j - y_j) & \text{(base case: } j \text{ is output unit)} \\ h'(a_j) \sum_k w_{jk} \delta_k & \text{(recursive case: } j \text{ is hidden)} \end{cases}$$

- Since $a_j = \sum_i w_{ji} z_i$, thus $\frac{\partial a_j}{\partial w_{ji}} = z_i$.

EXAMPLE

<see annotated slides>

Chapter 12: Deep Neural Networks

DEEP NEURAL NETWORKS

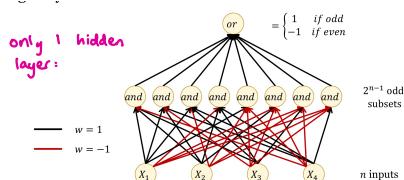
- Θ_1 A "deep neural network" is a NN with many hidden layers.

Θ_2 Advantage: high expressivity.

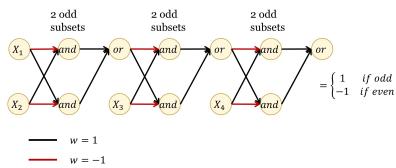
EXPRESSIVENESS

- Θ_1 Idea: Although NNs with 1 layer of sigmoid/hyperbolic units can approximate arbitrarily closely NNs with several layers, the number of units may decrease exponentially as the number of layers increases.

- Θ_2 Example: parity function



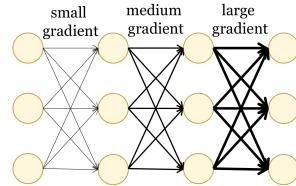
$2n-2$ hidden layers:



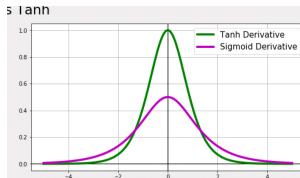
- with more hidden layers, we need less hidden nodes.

VANISHING GRADIENTS

- Θ_1 Idea: Deep NNs using sigmoid/hyperbolic units often suffer from vanishing gradients.



Θ_2 This is because the derivatives of the Sigmoid & tanh functions are ≤ 1 .



Θ_3 eg $y = \sigma(w_3 \sigma(w_2 \sigma(w_1 x)))$

$$w_1 \rightarrow h_1 \rightarrow h_2 \rightarrow h_3$$

Then

$$\frac{\partial y}{\partial w_3} = \sigma'(a_3) \sigma(a_2)$$

$$\frac{\partial y}{\partial w_2} = \sigma'(a_3) w_3 \sigma'(a_2) \sigma(a_1)$$

$$\frac{\partial y}{\partial w_1} = \sigma'(a_3) w_3 \sigma'(a_2) w_2 \sigma'(a_1) \times$$

as products of factors ≤ 1 gets "longer", the gradient vanishes

- Θ_3 Solution: we use the "rectified linear unit" activation function:

$$h(a) = \max(0, a).$$

- gradient is 0 or 1
- sparse computation

- Θ_4 "Soft" version / "softplus":

$$h(a) = \log(1 + e^a)$$

- note this does not prevent gradient vanishing ($\text{gradient} < 1$)

OVERFITTING

Θ_1 : Idea: As the number of parameters is often larger than the amount of data, it increases the risk of overfitting.

DROPOUT

Θ_1 : This helps solve overfitting.

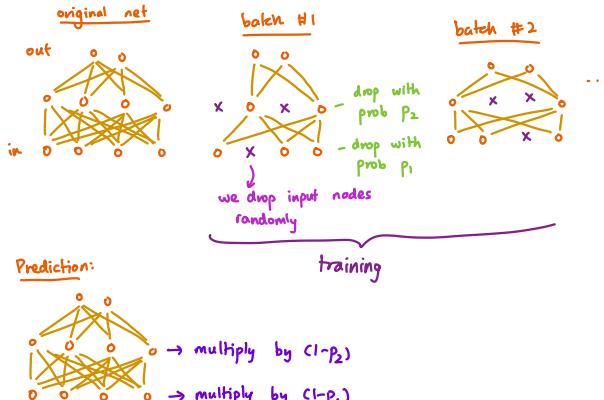
Θ_2 : Idea: Randomly "drop" some units from the network when training.

Θ_3 : Training: at each iteration of gradient descent:

- ① each input unit is dropped with probability p_1 ; &
- ② each hidden unit is dropped with probability p_2 .

Θ_4 : Prediction:

- ① Multiply each input unit by $1-p_1$; &
- ② Multiply each hidden unit by $1-p_2$.



Θ_3 : Algorithm:

Training: let \odot denote elementwise multiplication

• Repeat

- For each training example (x_n, y_n) do
 - Sample $z_n^{(l)}$ from $Bernoulli(1 - p_l)^{k_l}$ for $1 \leq l \leq L$
 - Neural network with dropout applied:

$$f_n(x_n, z_n; W) = h_L(W^{(L)} \left[\dots h_2 \left(W^{(2)} \left[h_1 \left(W^{(1)} \left[\bar{x}_n \odot z_n^{(1)} \right] \right) \odot z_n^{(2)} \right] \dots \odot z_n^{(L)} \right] \right])$$
 - Loss: $Err(y_n, f_n(x_n, z_n; W))$
 - Update: $w_{kj} \leftarrow w_{kj} - \eta \frac{\partial Err}{\partial w_{kj}}$
- End for

• Until convergence

Prediction: $f(x_n; W) = h_L(W^{(L)} \left[\dots h_2 \left(W^{(2)} \left[h_1 \left(W^{(1)} [\bar{x}_n (1 - p_1)] (1 - p_2) \right] \dots (1 - p_L) \right] \right] \right])$

Θ_4 : Intuitively, each dropout iteration trains a different sub-network, and we merge these during training.

Chapter 13: Decision Networks

MOTIVATION

Q: Sometimes, we need to make decisions under uncertainty.

PREFERENCE ORDERING: \geq

Q: A "preference ordering" \geq is a ranking of all possible states of affairs/worlds S .

- these could be outcomes of actions, states in a search problem, etc

Q: In particular, we use the notation

① $s \geq t \Rightarrow s$ is at least as good as t

② $s > t \Rightarrow s$ is strictly preferred to t

③ $s \sim t \Rightarrow$ agent is indifferent between s & t

where s & t are states.

Q: If an agent's actions are deterministic, then we know what states will occur.

Q: Otherwise, we can represent this using lotteries:

$$L = (p_1, s_1; \dots; p_n, s_n)$$

where state s_i occurs with probability p_i .

AXIOMS

Q: Given 3 states A, B & C:

① either $A > B$, $A \times B$ or $A \sim B$

(orderability);

② $A > B$, $B > C \Rightarrow A > C$

(transitivity);

③ $A \succcurlyeq B \succcurlyeq C \Rightarrow \exists p$ s.t. $[p, A; 1-p, C] \sim B$

(continuity);

④ $A \sim B \Rightarrow [p, A; 1-p, C] \sim [p, B; 1-p, C]$

(substitutability);

⑤ $A > B \Rightarrow (p \geq q \Leftrightarrow [p, A; 1-p, B] \succ [q, A; 1-q, B])$

(monotonicity)

⑥ $[p, A; 1-p, [q, B; 1-q, C]] \sim [p, A; (1-p)q, B; (1-p)(1-q), C]$

(decomposability)

UTILITY FUNCTION

Q: A "utility function" $U: S \rightarrow \mathbb{R}$ associates a "utility" with each outcome.

Q: In particular, $U(s)$ measures our degree of preference for s .

Q: Note U induces a "preference ordering" \geq_U over S by $s \leq_U t \Leftrightarrow U(s) \leq U(t)$.

EXPECTED UTILITY: $EU(d)$

Q: Idea: Under uncertainty, each decision d induces a distribution P_d over possible outcomes, where $P_d(s)$ is the probability of outcome s under decision d .

Q: The "expected utility" of decision d is

$$EU(d) = \sum_{s \in S} P_d(s) U(s)$$

PRINCIPLE OF MAXIMUM EXPECTED UTILITY (MEU)

Q: MEU states the optimal decision under conditions of uncertainty is the one with the highest expected utility.

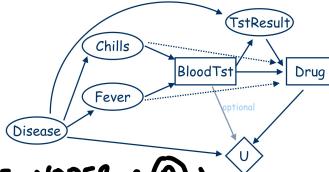
DECISION NETWORKS / INFLUENCE DIAGRAMS

Q₁: "Decision networks" provide a way of representing sequential decision problems.

Idea:

- ① Represent the variables like in a BN;
- ② Add decision/controllable variables; &
- ③ Add utility variables that describe how good different states are.

eg



CHANCE NODES (A)

Q₁: "Chance nodes" are random variables.

- denoted by circles

Q₂: Like a BN, they contain CPTs with probabilistic inference on their parents.

DECISION NODES (A)

Q₁: "Decision nodes" are variables set by the decision maker.

- denoted by squares

Q₂: In particular, the parents reflect information available at the time the decision is to be made.

eg



- the values of chills & fever need to be observed before the decision to take the test must be made

VALUE NODES (A)

Q₁: "Value nodes" specify utility of a state.

- denoted by a diamond

Q₂: In particular, the utility depends only on the states of the parents of the value node.

ASSUMPTIONS

Q₁: We assume

① decision variables are totally ordered: D_1, \dots, D_n

- ie decisions are made in sequence

D_1, \dots, D_n

② "no-forgetting" property: any information that is available when decision D_i is made is available when D_j is made, i.e. $j > i$.

- thus all parents of D_i are parents of D_j
- we use dashed lines to indicate this



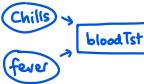
POLICIES: 8

Q₁: A policy δ is a set of mappings δ_i , one for each decision node D_i , where

$$\delta_i : \text{Dom}(\text{Par}(D_i)) \rightarrow \text{Dom}(D_i).$$

Q₂: In particular, δ_i associates a decision with each parent assignment for D_i .

eg



$$\begin{aligned}\delta_{\text{BT}}(c, f) &= \text{bt} \\ \delta_{\text{BT}}(c, \neg f) &= \neg \text{bt} \\ \delta_{\text{BT}}(\neg c, f) &= \text{bt} \\ \delta_{\text{BT}}(\neg c, \neg f) &= \neg \text{bt}\end{aligned}$$

VALUE OF A POLICY: EU(δ)

Q₁: The "value" of policy δ is the expected utility given that decisions are executed according to δ .

Q₂: Essentially,

$$\text{EU}(\delta) = \sum_X P(X, \delta(X)) U(X, \delta(X))$$

where $\delta(X)$ denotes the assignment to decision variables dictated by δ given the assignment X .

OPTIMAL POLICIES

\bullet we say a policy δ^* is "optimal" if

$$EU(\delta^*) \geq EU(\delta)$$

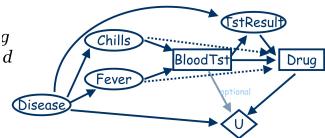
for all policies δ .

\bullet_2 To compute the best policy:

- ① Start with the last decision;
- ② For each assignment to parents & for each decision value, compute the expected value of choosing that value of D;
- ③ Set the policy choice for each value of parents to be the value of D that has max value;
- ④ Repeat these steps for each decision in "reverse" order.

\bullet_3 To compute the expected values, we can use variable elimination.

eg



- eg suppose we have asst cc,f,bt,pos to $P(\text{Drug})$
- we want $EU(\text{Drug} = \text{md} | c, f, bt, pos)$
- in variable elimination, we can treat C, F, BT, DR as evidence
- then eliminate remaining variables - in this case, only Disease is left
- we are left with the factor

$$EU(\text{md} | c, f, bt, pos) = \sum P(\text{Dis} | c, f, bt, pos, \text{md}) U(\text{Dis}, \text{bt}, \text{md})$$

\bullet_4 Finally, we find which D maximises EUD (evidence), which will be in the optimal policy.

OPTIMAL POLICIES FOR BNs

\bullet_1 In BNs, utility nodes are just factors that can be dealt with using variable elimination.

\bullet_2 Thus, for this case, we can just use variable elimination.

OPTIMIZING POLICIES: NOTES

\bullet_1 Idea: If a decision node D has no decisions that follow it, we can find its policy by instantiating each of its parents and computing the expected utility of each decision for each parent instantiation.

- no-forgetting \Rightarrow all other decisions are already instantiated.

\bullet_2 When a decision D is optimized, we can treat it as a random variable.

- just treat the policy as a new CPT
- given parent instantiation x, D gets $\delta(x)$ with probability 1

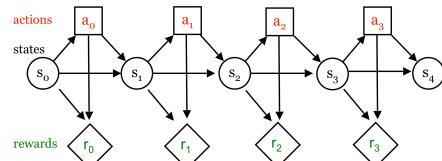
\bullet_3 At each iteration of the decision optimization process, we can optimize D; by using simple variable elimination calculations.

Chapter 14: Markov Decision Processes

SEQUENTIAL DECISION MAKING

- "Sequential decision making" combines static decision making (eg in decision networks) & sequential inference (eg HMMs, dynamic BNs)

MARKOV DECISION PROCESSES



Idea: These are indefinite/infinite/large finite decision networks.

Formal definition: a Markov decision process has

- states $s \in S$;
- actions $a \in A$;
- rewards $r \in R$;
- transition model $P(s_t | s_{t-1}, a_{t-1})$;
- reward model $R(s_t | a_t)$;
- discount factor $0 \leq \gamma \leq 1$;
- horizon (# of time steps) h .

Our goal is to find the optimal policy; ie an optimal way to act at every state to maximize the utility/reward.

CURRENT ASSUMPTIONS

- Assumptions:
- Process is stochastic;
- Process is sequential;
- States are fully observable;
- Model is complete;
- no learning is required
- States & actions are discrete.
- note that we can cycle between states.

TRANSITION MODEL: $P(s_t | s_{t-1}, a_{t-1})$

Assumptions:

- Markov: $P(s_t | s_{t-1}, a_{t-1}, s_{t-2}, a_{t-2}, \dots) = P(s_t | s_{t-1}, a_{t-1})$
- Stationary: $P(s_t | s_{t-1}, a_{t-1})$ is same given $(s_t, a_{t-1}, s_{t-1}) \forall t$.

REWARD MODEL

- Reward function: $R(s_t, a_t) = r_t$

Assumption: the reward function is stationary;
ie $R(s_t, a_t)$ is the same for a given (s, a) .

However, the terminal reward does not have to be stationary.
eg +1/-1 for winning/losing

Goal: maximize sum of expected rewards $\sum_t R(s_t, a_t)$.

DISCOUNTED REWARDS

Idea: If h is infinite, then $\sum_t R(s_t, a_t) = \infty$, which is not ideal.

Solution: use "discounted rewards"

$$\sum_t \gamma^t R(s_t, a_t)$$

where $\gamma \in [0, 1]$ is the discount factor.

Intuition: we prefer utility sooner than later.

POLICY

The "policy" is the choice of action at each time step.

Formally, this maps states to actions

$$\pi(s_t) = a_t$$

Goal: Find the optimal policy

$$\pi^* = \arg\max_{\pi} \sum_{t=0}^h \gamma^t E_{\pi}[r_t]$$

POLICY OPTIMIZATION

To evaluate a policy, we can compute the value of following it:

$$V^\pi(s_0) = \sum_{t=0}^h \gamma^t \sum_{s_t} P(s_t | s_0, \pi) R(s_t, \pi(s_t))$$

The optimal policy is such that

$$V^\pi(s_0) \geq V^\pi(s_0) \quad \forall \pi.$$

Algorithms:

- ① Value iteration
- ② Policy iteration
- Computation can be done
- ① "offline": before the process starts
- ② "online": as the process evolves.

VALUE ITERATION

Idea: we find the max values iteratively at the t^{th} time step:

$$V_0(s) = \max_a R(s, a) \quad \forall s$$

$$V_t(s) = \max_a R(s, a) + \gamma \sum_{s'} P(s' | s, a) V_{t-1}(s') \quad \forall s$$

In particular,

$$a_t = \operatorname{argmax}_a R(s, a) + \gamma \sum_{s'} P(s' | s, a) V_{t-1}(s') \quad \forall s$$

Algorithm:

valueIteration(MDP)

$$V_0(s) \leftarrow \max_a R(s, a) \quad \forall s$$

For $n = 1$ to h do

$$V_n^*(s) \leftarrow \max_a R(s, a) + \gamma \sum_{s'} \Pr(s' | s, a) V_{n-1}^*(s') \quad \forall s$$

Return V^*

We can represent value iteration in a matrix form:

$$\begin{aligned} R^a &\in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} \\ V_n^* &\in \mathbb{R}^{|\mathcal{S}|} \\ T^a &\in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|} \end{aligned}$$

HORIZON EFFECT

If h is finite, the policy is non-stationary, and there is no guarantee to converge.

- best action different at each time step

If h is infinite, the policy is stationary, and there is a guarantee for the value iteration to converge.

- same best action at each time step

INFINITE HORIZON

To deal with an infinite horizon, we can use

- ① a large enough n and execute the policy at the n^{th} iteration; or
- ② continue iterating until $|V_n - V_{n-1}| < \epsilon$.
- ϵ is the "threshold".

POLICY ITERATION

Idea: we alternate between 2 steps:

- ① Policy evaluation: given π ,

$$V^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s' | s, \pi(s)) V^\pi(s') \quad \forall s$$

- ② Policy improvement:

$$\pi(s) \leftarrow \operatorname{argmax}_a R(s, a) + \gamma \sum_{s'} P(s' | s, a) V^\pi(s') \quad \forall s$$

Algorithm:

policyIteration(MDP)

Initialize π_0 to any policy

$$n \leftarrow 0$$

Repeat

$$\text{Eval: } V_n = R^{\pi_n} + \gamma T^{\pi_n} V_n$$

$$\text{Improve: } \pi_{n+1} \leftarrow \operatorname{argmax}_a R^a + \gamma T^a V_n$$

$$n \leftarrow n + 1$$

$$\text{Until } \pi_{n+1} = \pi_n$$

Return π_n

COMPLEXITY

Value iteration:

- ① Each iteration: $O(|\mathcal{S}|^2 |\mathcal{A}|)$

- ② Many iterations: linear convergence

Policy iteration:

- ① Each iteration: $O(|\mathcal{S}|^3 + |\mathcal{S}|^2 |\mathcal{A}|)$

- ② Few iterations: linear-quadratic convergence

Chapter 15:

Reinforcement Learning

PROBLEM



Q₁: We want to learn to choose actions that maximize rewards.

Q₂: We have states, actions & rewards, but do not know the transition or reward models.

Q₃: Goal: We want to find

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \sum_{t=0}^n \gamma^t E_{\pi}[r_t]$$

Q₄: Idea: We want to learn the model.

COMPONENTS

- Q: RL agents may include
- ① the model $P(s'|s, a)$, $R(s, a)$;
 - ② the policy $\pi(s)$;
 - ③ the value function $V(s)$.

CATEGORIES

Value based

- No policy (implicit)
- Value function

Policy based

- Policy

- No value function

Actor critic

- Policy
- Value function

Model based

- Transition and reward model

Model free

- No transition and no reward model (implicit)

Online RL

- Learn by interacting with environment

Offline RL

- No environment
- Learn only from saved data

MODEL FREE EVALUATION

Q₁: Idea: Given a policy π , estimate $V^{\pi}(s)$ without any transition or reward model.

Q₂: Strategies:

① Monte-Carlo evaluation:

$$V_{\pi}(s) = E_{\pi} \left[\sum_t \gamma^t r_t | s, \pi \right]$$

$$\approx \frac{1}{n(s)} \sum_{k=1}^{n(s)} \sum_t [\gamma^t r_t^{(k)} | s, \pi]$$

→ several sample approximation

② Temporal difference (TD) evaluation:

$$V_{\pi}(s) = E[r | s, \pi(s)] + \gamma \sum_{s'} P(s' | s, \pi(s)) V^{\pi}(s')$$

$$\approx r + \gamma V^{\pi}(s')$$

→ one sample approximation

MONTE CARLO EVALUATION

Q₁: Let

$$G_n = \sum_t \gamma^t r_t^{(n)}$$

Q₂: Then

$$V_n^{\pi}(s) \approx V_{n-1}^{\pi}(s) + \frac{1}{n(s)} (G_{n(s)} - V_{n-1}^{\pi}(s))$$

Q₃: Incremental update:

$$V_n^{\pi}(s) \leftarrow V_{n-1}^{\pi}(s) + \alpha_n (G_n - V_{n-1}^{\pi}(s)), \quad \alpha_n = \frac{1}{n(s)}$$

TEMPORAL DIFFERENCE EVALUATION

\exists_1 Incremental update:

$$V_n^{\pi}(s) \leftarrow V_{n-1}^{\pi}(s) + \alpha_n(r + \gamma V_{n-1}^{\pi}(s') - V_{n-1}^{\pi}(s))$$

\exists_2 If α_n is decreased appropriately with the # of times a state is visited, then $V_n^{\pi}(s)$ converges to the correct value.

\exists_3 Sufficient conditions for $\alpha_n(s)$:

$$\begin{aligned}\sum_n \alpha_n &= \infty \\ \sum_n \alpha_n &< \infty\end{aligned}$$

\exists_4 We often choose $\alpha_n(s) = \frac{1}{n(s)}$.

\exists_5 Algorithm:

TDevaluation(π, V^{π})

Repeat

 Execute $\pi(s)$

 Observe s' and r

 Update counts: $n(s) \leftarrow n(s) + 1$

 Learning rate: $\alpha \leftarrow \frac{1}{n(s)}$

 Update value: $V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha(r + \gamma V^{\pi}(s') - V^{\pi}(s))$
 $s \leftarrow s'$

Until convergence of V^{π}

Return V^{π}

COMPARISON

Monte Carlo	TD
- unbiased estimate	- biased estimate
- high variance	- low variance
- needs many trajectories	- needs less trajectories

MODEL-FREE CONTROL

\exists_1 Idea: Instead of evaluating the state value function $V^{\pi}(s)$, evaluate the "state-action value function" $Q^{\pi}(s, a)$

$$Q^{\pi}(s, a) = E(r|s, a) + \gamma \sum_{s'} P(s'|s, a) V^{\pi}(s')$$

- value of executing a followed by π

\exists_2 Then, we use the policy

$$\pi^*(s) = \operatorname{argmax}_a Q^{\pi}(s, a)$$

BELLMAN'S EQUATION

\exists_1 Optimal state-action value function:

$$Q^*(s, a) = E[r|s, a] + \gamma \sum_{s'} P(s'|s, a) \max_a Q^*(s', a')$$

$$\text{where } V^*(s) = \max_a Q^*(s, a), \\ \pi^*(s) = \operatorname{argmax}_a Q^*(s, a).$$

Q-LEARNING

Qlearning(s, Q^*)

Repeat

 Select and execute a ↗ how do we choose the action?

 Observe s' and r

 Update counts: $n(s, a) \leftarrow n(s, a) + 1$

 Learning rate: $\alpha \leftarrow \frac{1}{n(s, a)}$

 Update Q-value:

$$Q^*(s, a) \leftarrow Q^*(s, a) + \alpha \left(r + \gamma \max_{a'} Q^*(s', a') - Q^*(s, a) \right)$$

↳ vs, a

Until convergence of Q^*

Return Q^*

EXPLORATION VS EXPLOITATION

\exists_1 Idea: If the agent always chooses the action with the highest value, it is "exploiting", and the learned model is not accurate.

\exists_2 By taking random actions ("exploration"), the agent may learn the model, but parts of it will never be used.

\exists_3 Thus, we need a balance.

COMMON EXPLORATION METHODS

\exists_1 Methods:

① ϵ -greedy: with prob ϵ , execute random action; otherwise, execute the best action $a^* = \operatorname{argmax}_a Q(s, a)$

② Boltzmann exploration: increasing temp T increases stochasticity

$$P(a) = \frac{e^{Q(s, a)/T}}{\sum_a e^{Q(s, a)/T}}$$

EXPLORATION & Q-LEARNING

\exists_1 Q-learning converges to optimal Q-values if

- ① every state is visited infinitely often;
- ② the action selection becomes greedy as $t \rightarrow \infty$;
- ③ the learning rate is decreased fast enough, but not too fast;

$$\sum_n \alpha_n \rightarrow \infty, \quad \sum_n (\alpha_n)^2 < \infty.$$

e.g. do ϵ -greedy, but decrease ϵ over time

Chapter 16: Deep Reinforcement Learning

LARGE STATE SPACES

Q₁: Idea: For large state spaces, Q-learning is impractical since the update function has complexity proportional to the state space size.

Q₂: We need to approximate

- ① the policy $\pi(s) \rightarrow a$;
- ② the Q-function $Q(s, a) \rightarrow \mathbb{R}$; &
- ③ the value function $V(s) \rightarrow \mathbb{R}$.

Q-FUNCTION APPROXIMATION

Q₁: Let $s = (x_1, \dots, x_n)^T$.

- ① Linear:

$$Q(s, a) \approx \sum_i w_{ai} x_i$$

- ② Non-linear (e.g. neural network):

$$Q(s, a) \approx g(x; w)$$

GRADIENT Q-LEARNING

Q₁: Idea: We want to minimize the squared error between

- ① the Q-value estimate: $Q_w(s, a)$
- ② the target: $r + \gamma \max_{a'} Q_{\bar{w}}(s', a')$.

Q₂: Squared error:

$$\text{Err}(w) = \frac{1}{2} [Q_w(s, a) - r - \gamma \max_{a'} Q_{\bar{w}}(s', a')]^2$$

Q₃: Gradient:

$$\frac{\partial \text{Err}}{\partial w} = [Q_w(s, a) - r - \gamma \max_{a'} Q_{\bar{w}}(s', a')] \frac{\partial Q_w}{\partial w}$$

Q₄: We can then use gradient descent.

Initialize weights w at random in $[-1, 1]$

Observe current state s

Loop

Select action a and execute it

Receive immediate reward r

Observe new state s'

Gradient: $\frac{\partial \text{Err}}{\partial w} = [Q_w(s, a) - r - \gamma \max_a Q_w(s', a')] \frac{\partial Q_w(s, a)}{\partial w}$

Update weights: $w \leftarrow w - \alpha \frac{\partial \text{Err}}{\partial w}$

Update state: $s \leftarrow s'$

CONVERGENCE OF APPROXIMATION

Q-LEARNING

Q₁: Given $\sum a_t = \infty$, $\sum a_t^2 < \infty$:

- ① Linear approximation Q-learning converges; but
- ② Non-linear approximation Q-learning may diverge.
 - adjusting w to increase $\&$ at (s, a) may introduce errors at nearby state-action pairs.

MITIGATING DIVERGENCE

To mitigate divergence, we can use

① Experience replay;

② Using 2 networks:

- Q-network;

- target network.

EXPERIENCE REPLAY

Idea: store previous experiences $\langle s, a, s', r \rangle$ into a buffer & sample a mini-batch of previous experiences at each step to learn by Q-learning.

Advantages:

① break correlations between successive updates (more stable learning)

② less interactions with environment needed (better data efficiency)

TARGET NETWORK

Idea: use a separate target network which is only updated periodically.

repeat for each (s, a, s', r) in mini-batch:

$$w \leftarrow w - \alpha_i \left[Q_w(s, a) - r - \gamma \max_{a'} Q_{\bar{w}}(s', a') \right] \frac{\partial Q_w(s, a)}{\partial w}$$

$$\bar{w} \leftarrow w$$

- similar to value iteration.

Advantage: mitigate divergence.

DEEP Q-NETWORK / DQN

A "deep Q-network" uses gradient

Q-learning with

① deep neural networks;

② experience replay;

③ target network.

Initialize weights w and \bar{w} at random in $[-1, 1]$

Observe current state s

Loop

Select action a and execute it

Receive immediate reward r

Observe new state s'

Add $\langle s, a, s', r \rangle$ to experience buffer

Sample mini-batch of experiences from buffer

For each experience $\langle \hat{s}, \hat{a}, \hat{s}', \hat{r} \rangle$ in mini-batch

$$\text{Gradient: } \frac{\partial Err}{\partial w} = [Q_w(\hat{s}, \hat{a}) - \hat{r} - \gamma \max_{\hat{a}'} Q_{\bar{w}}(\hat{s}', \hat{a}')] \frac{\partial Q_w(\hat{s}, \hat{a})}{\partial w}$$

$$\text{Update weights: } w \leftarrow w - \alpha \frac{\partial Err}{\partial w}$$

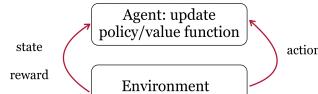
Update state: $s \leftarrow s'$

Every c steps, update target: $\bar{w} \leftarrow w$

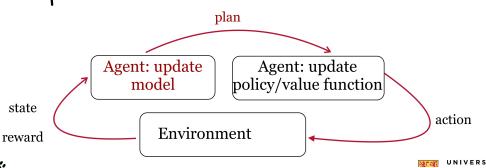
Chapter 16: Model-Based Reinforcement Learning

MODEL-FREE VS MODEL-BASED RL

Q1 In model-free online RL, there are no explicit transition or reward models.



Q2 In model-based online RL, we learn an explicit transition and/or reward model.



Q3 Benefit: Increased sample efficiency

Q4 Drawback: Increased complexity.

MODEL-BASED RL METHOD

Q1 Idea: At each step:

- ① Execute action;
 - ② Observe resultant state & reward;
 - ③ Update transition/reward models;
 - ④ Update policy/value function.
- Q2 Algorithm with value iteration:

ModelBasedRL(s)

Repeat

Select and execute a // similar to Q-learning

Observe s' and r

Update counts: $n(s, a) \leftarrow n(s, a) + 1$,
 $n(s, a, s') \leftarrow n(s, a, s') + 1$

Update transition: $\Pr(s'|s, a) \leftarrow \frac{n(s, a, s')}{n(s, a)} \forall s'$

Update reward: $R(s, a) \leftarrow \frac{r + (n(s, a) - 1)R(s, a)}{n(s, a)}$

Solve: $V^*(s) = \max_a R(s, a) + \gamma \sum_{s'} \Pr(s'|s, a) V^*(s') \forall s$

$s \leftarrow s'$

Until convergence of V^*

Return V^*

COMPLEX MODELS

Q1 Idea: Use function approximations for the transition & reward models:

- ① Linear model:

$$\text{pdf}(s'|s, a) = N(s'|w^T x, \sigma^2 I)$$

- ② Non-linear model:

- stochastic: Gaussian process;

$$\text{pdf}(s'|s, a) = GP(s|m(\cdot), k(\cdot, \cdot))$$

- deterministic: neural networks;

$$s' = T(s, a) = NN(s, a)$$

PARTIAL PLANNING

Q1 Idea: In complex models, fully optimizing the policy/value function at each time step is intractable.

Q2 To mitigate this, we can do partial planning that involves

- ① a few steps of Q-learning; &
 ② learning from simulated experience.

MODEL-BASED RL WITH Q-LEARNING

ModelBasedRL(s)

Repeat

Select and execute a , observe s' and r

Update transition: $w_T \leftarrow w_T - \alpha_T(T_{w_T}(s, a) - s') \nabla_{w_T} T_{w_T}(s, a)$

Update reward: $w_R \leftarrow w_R - \alpha_R(R_{w_R}(s, a) - r) \nabla_{w_R} R(s, a)$

Repeat a few times:
 sample \hat{s}, \hat{a} arbitrarily

$$\delta \leftarrow R_{w_R}(\hat{s}, \hat{a}) + \gamma \max_{\hat{a}'} Q_{w_Q}(T_{w_T}(\hat{s}, \hat{a}), \hat{a}') - Q_{w_Q}(\hat{s}, \hat{a})$$

Update Q: $w_Q \leftarrow w_Q - \alpha_Q \delta \nabla_{w_Q} Q_{w_Q}(\hat{s}, \hat{a})$

$s \leftarrow s'$
 Until convergence of Q

Return Q

PARTIAL PLANNING VS REPLAY BUFFER

① Idea: In model-free Q-learning with a replay buffer, we update the Q-function based on samples from the replay buffer; in the previous algorithm, we update Q by generating samples from the model.

② Replay buffer:

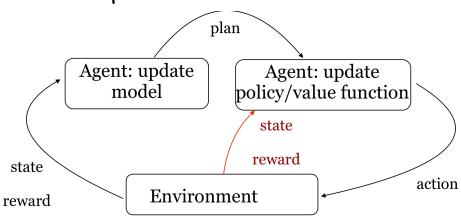
- ① simple;
- ② real samples; but
- ③ no generalization to other state-action pairs.

③ Partial planning with model:

- ① complex;
- ② simulated samples; but
- ③ generalization to other state-action pairs.

DYNA - Q

④ Idea: We learn an explicit transition & reward model & learn directly from real experience.



- outer loop: similar to model-based
- inner loop: similar to model-free

⑤ Algorithm:

Dyna-Q(s)

Repeat

Select and execute a , observe s' and r
Update transition: $w_T \leftarrow w_T - \alpha_T(T_{w_T}(s, a) - s') \nabla_{w_T} T_{w_T}(s, a)$
Update reward: $w_R \leftarrow w_R - \alpha_R(R_{w_R}(s, a) - r) \nabla_{w_R} R(s, a)$

$$\delta \leftarrow r + \gamma \max_{a'} Q_{w_Q}(s', a') - Q_{w_Q}(s, a)$$

Update Q : $w_Q \leftarrow w_Q - \alpha_Q \delta \nabla_{w_Q} Q_{w_Q}(s, a)$

Repeat a few times:

sample \hat{s}, \hat{a} arbitrarily

$$\delta \leftarrow R_{w_R}(\hat{s}, \hat{a}) + \gamma \max_{\hat{a}'} Q_{w_Q}(T_{w_T}(\hat{s}, \hat{a}), \hat{a}') - Q_{w_Q}(\hat{s}, \hat{a})$$

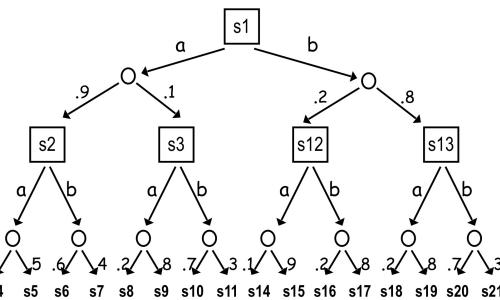
Update Q : $w_Q \leftarrow w_Q - \alpha_Q \delta \nabla_{w_Q} Q_{w_Q}(\hat{s}, \hat{a})$

$$s \leftarrow s'$$

Return Q

PLANNING FROM CURRENT STATE: MONTE CARLO TREE SEARCH

Idea: instead of planning at arbitrary states, plan from the current state, which helps improve the next action.



To make this tractable:

- Approximate leaf values with value of default policy;

$$Q^*(s, a) \approx Q^T(s, a) \approx \frac{1}{n(s, a)} \sum_{k=1}^n G_{ik}$$

- Approximate chance nodes' expectation by sampling from transition model;

$$Q^*(s, a) \approx R(s, a) + \gamma \sum_{s'} V(s')$$

- For decision nodes, only expand the most promising actions.

$$a^* = \underset{a}{\operatorname{argmax}} \quad Q(s, a) + c \sqrt{\frac{2 \ln(n(s))}{n(s, a)}}$$

$$V^*(s) = \max_{a^*} Q^*(s, a^*)$$

Algorithm:

UCT(s_0)

```
create root node0 with state state(node0) ←  $s_0$ 
while within computational budget do
    nodei ← TreePolicy(nodei)
    value ← DefaultPolicy(state(nodei))
    Backup(nodei, value)
    return action(BestChild(node0, 0))
```

TreePolicy($node$)

```
while node is nonterminal do
    if node is not fully expanded do
        return Expand(node)
    else
        node ← BestChild(node, C)
return node
```

Expand($node$)

```
choose  $a \in$  untried actions of  $A(state(node))$ 
add a new child  $node'$  to node
with state( $node'$ ) ←  $T(state(node), a)$ 
return  $node'$ 
```

deterministic transition

BestChild($node, c$)

```
return arg max $node' \in children(node)$   $V(node') + c \sqrt{\frac{2 \ln(n(node))}{n(node')}}$ 
```

DefaultPolicy($node$)

```
while node is not terminal do
    sample  $a \sim \pi(a | state(node))$ 
     $s' \leftarrow T(state(node), a)$ 
return  $R(s, a)$ 
```

Single Player

Backup($node, value$)

```
while node is not null do
     $n(node)V(node) + value$ 
     $V(node) \leftarrow \frac{n(node)V(node) + value}{n(node) + 1}$ 
     $n(node) \leftarrow n(node) + 1$ 
    node ← parent(node)
```

Two Players (adversarial)

BackupMinMax($node, value$)

```
while node is not null do
     $n(node)V(node) + value$ 
     $V(node) \leftarrow \frac{n(node)V(node) + value}{n(node) + 1}$ 
     $n(node) \leftarrow n(node) + 1$ 
    value ← - value
    node ← parent(node)
```

Chapter 17: Multi-Armed Bandits

STOCHASTIC BANDITS

- Q₁: A "bandit" has
 - ① a single state $\{s\}$;
 - ② a set of actions/arms A ;
 - ③ space of rewards (often rescaled to $[0, 1]$);
 - ④ finite/infinite horizons;
 - ⑤ average reward setting ($\gamma = 1$)
- Q₂: There is no transition function to be learned since there is a single state.
- Q₃: We only need to learn the stochastic reward function.

EXAMPLE: AD PLACEMENT

- Q₁: Idea:
 - ① Arms: set of possible ads
 - ② Rewards: 0 (no click), 1 (click)
- Q₂: What order should ads be presented to maximize revenue?
 - exploration vs exploitation problem

2-GREEDY

- Q₁: Idea: Select an arm at random with prob ϵ , and otherwise do a "greedy" selection (ie select arm with the highest average so far).

REGRET

- Q₁: Let $R(a)$ be the true (unknown) expected reward of a , and let

$$\begin{aligned} r^* &= \max_a R(a) \\ a^* &= \operatorname{argmax}_a R(a). \end{aligned}$$

- Q₂: The "expected regret" of a is

$$\text{loss}(a) = r^* - R(a).$$

- Q₃: The "expected cumulative regret" for n time steps is

$$\text{Loss}_n = \sum_{t=1}^n \text{loss}(a_t).$$

THEORETICAL GUARANTEES

- Q₁: If ϵ is constant, then for large enough t :

$$\begin{aligned} P(a_t \neq a^*) &\approx \epsilon \\ \text{Loss}_n &\approx \sum_{t=1}^n \epsilon \in O(n). \end{aligned}$$

- Q₂: If $\epsilon_t \approx \frac{1}{t}$, then for large enough t :

$$\begin{aligned} P(a_t \neq a^*) &\approx \epsilon_t \in O\left(\frac{1}{t}\right) \\ \text{Loss}_n &\approx \sum_{t=1}^n \frac{1}{t} \in O(\log n) \end{aligned}$$

EMPIRICAL MEAN

- Q₁: Idea: We want to quantify the empirical mean $\tilde{R}(a)$ from the true mean $R(a)$.

- Q₂: If we can write

$$|R(a) - \tilde{R}(a)| \leq \text{bound}$$

then we can select the arm with the best $\tilde{R}(a) + \text{bound}$, since $R(a) \leq \tilde{R}(a) + \text{bound}$.

POSITIVISM IN THE FACE OF UNCERTAINTY

- Q₁: Suppose there exists an oracle that returns an upper bound $UB_n(a)$ on $R(a)$ for each arm a based on n trials.

- Q₂: Suppose further

$$\lim_{n \rightarrow \infty} UB_n(a) = R(a).$$

- Q₃: Optimistic algorithm: at each step, select

$$\operatorname{argmax}_a UB_n(a)$$

- Q₄: This algorithm will converge to a^* .

Proof: Suppose we converge to suboptimal arm a' after infinitely many trials.
Then

$$R(a) = UB_\infty(a) \geq UB_\infty(a') = R(a') \quad \forall a'$$

But $R(a) \geq R(a')$ $\forall a'$ contradicts our assumption that a is suboptimal.

PROBABILISTIC UPPER BOUND

Θ_1 Idea: We cannot compute an upper bound with certainty as we are sampling.

Θ_2 But, we can obtain measures of that are upper bounds most of the time; ie

$$P(R(a) \leq f(a)) \geq 1 - \delta.$$

Θ_3 Hoeffding's inequality:

$$P(R(a) \leq \tilde{R}(a) + \sqrt{\frac{\log(\frac{1}{\delta})}{2n_a}}) \geq 1 - \delta$$

where $n_a = \#$ of trials for arm a .

UPPER CONFIDENCE BOUND / UCB

UCB(h)

$$V \leftarrow 0, n \leftarrow 0, n_a \leftarrow 0 \quad \forall a$$

Repeat until $n = h$

$$\text{Execute } \underset{a}{\operatorname{argmax}} \tilde{R}(a) + \sqrt{\frac{2 \log n}{n_a}}$$

Receive r

$$V \leftarrow V + r$$

$$\tilde{R}(a) \leftarrow \frac{n_a \tilde{R}(a) + r}{n_a + 1}$$

$$n \leftarrow n + 1, n_a \leftarrow n_a + 1$$

Return V

- we choose $\delta_n = \frac{1}{n^4}$ in Hoeffding's bound

UCB CONVERGENCE

Θ_1 UCB converges as $n \rightarrow \infty$.

Why? - as n increases, $\sqrt{\frac{2 \log n}{n_a}}$ increases
- so all arms are tried infinitely often

Θ_2 In particular,

$$\text{Loss}_n \in O(\log n)$$

BAYESIAN LEARNING

Θ_1 Let r^a be a random variable for a 's rewards.

Θ_2 Idea:

- ① Express uncertainty about θ by a prior $P(\theta)$;
- ② Compute posterior $P(\theta | r_1^a, \dots, r_n^a)$ based on samples r_1^a, \dots, r_n^a observed so far.

Θ_3 By Bayes' Theorem, we have

$$P(\theta | r_1^a, \dots, r_n^a) \propto P(\theta) P(r_1^a, \dots, r_n^a | \theta)$$

DISTRIBUTIONAL INFO

Θ_1 we can estimate

- ① the distribution over the next reward:

$$P(r_{n+1}^a | r_1^a, \dots, r_n^a) = \int_{\theta} P(r_{n+1}^a | \theta) P(\theta | r_1^a, \dots, r_n^a) d\theta$$

- ② the distribution over $R(a)$ when θ includes the mean:

$$P(R(a) | r_1^a, \dots, r_n^a) = P(\theta | r_1^a, \dots, r_n^a)$$

if $\theta = R(a)$.

BETA DISTRIBUTION: Beta(α, β)

Θ_1 The "Beta distribution" has the property that if $P(\theta) \sim \text{Beta}(\alpha, \beta)$, then

$$P(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

THOMPSON SAMPLING

Θ_1 Idea:

- ① Sample several potential average rewards

$$R_1(a), \dots, R_k(a) \sim P(R(a) | r_1^a, \dots, r_n^a) \quad \forall a$$

- ② Estimate the empirical average

$$\hat{R}(a) = \frac{1}{k} \sum_{i=1}^k R_i(a)$$

- ③ Then we can find

$$a^* = \underset{a}{\operatorname{argmax}} \hat{R}(a).$$

Θ_2 Algorithm:

ThompsonSampling(h)

$$V \leftarrow 0$$

For $n = 1$ to h

Sample $R_1(a), \dots, R_k(a) \sim P(R(a)) \quad \forall a$

$$\hat{R}(a) \leftarrow \frac{1}{k} \sum_{i=1}^k R_i(a) \quad \forall a$$

$$a^* \leftarrow \underset{a}{\operatorname{argmax}} \hat{R}(a)$$

Execute a^* and receive r

$$V \leftarrow V + r$$

Update $P(R(a^*))$ based on r

Return V

SAMPLE SIZE

- Q₁ Idea: In Thompson sampling, the amount of data n & sample size k regulate the amount of exploration.
- Q₂ In particular, as n & k increase, $\hat{R}(a)$ becomes less stochastic.
- Q₃ This ensures all actions are chosen with some probability.

ANALYSIS

- Q₁ Thompson sampling converges to the best arm.
- Q₂ Theoretical Loss $\in O(\log n)$.

Chapter 18: Game Theory

MULTI-AGENT DECISION MAKING

💡 Idea: In practice, there is usually more than one agent.

💡 Thus, each agent needs to account for other agents' actions/behaviors.

GAME

💡 A "game" is any set of circumstances whose outcomes depend on actions of two or more rational self-interested players.

PLAYERS

💡 "Players" are agents within the game that observe state & take actions.

RATIONAL

💡 We say an agent is "rational" if they choose their best actions, unless they are exploring.

SELF-INTERESTED

💡 We say an agent is "self-interested" if they only care about their own benefits.

GAME THEORY

💡 "Game theory" is a mathematical model of strategic interactions amongst ≥ 1 agents in a game.

INTERACTIONS

💡 An "interaction" occurs when one agent directly affects other agent(s).

💡 Thus, the utility for one agent depends on other agents.

STRATEGIC

💡 We say agents are "strategic" if they maximize their utility by taking into account their influence on the game via their actions.

LEARNING

💡 Idea: Each agent decides to act based on

- ① the world;
- ② other agents; &
- ③ their utility function.

TYPES OF GAMES

💡 Types:

- ① Cooperative - agents have a common goal
- ② Competitive - agents have a conflicting goal
- ③ Mixed - mix of both

NORMAL FORM GAMES

💡 A "normal form game" consists of

- ① a set of agents $I = 1, \dots, N, N \geq 2$;
- ② a set of actions for each agent $A_i = \{a_1^i, \dots, a_m^i\}$;
- ③ outcome of game is defined by a profile $a = (a_1, \dots, a_N)$;
- ④ total space of joint actions is $a \in A_1 \times \dots \times A_N$;
- ⑤ the utility functions are $u_i: A \rightarrow \mathbb{R}$.

PAYOUT MATRICES

💡 Idea: We can represent normal form games using "payoff matrices".

		Agent 2	
		One	Two
Agent 1	One	2, -2	-3, 3
	Two	-3, 3	4, -4

Diagram annotations:

- Labels: "actions agent 1 takes" points to the first column, "actions agent 2 takes" points to the second row.
- Utility labels: "utility of agent 1 from joint actions" points to the value 2, "utility of agent 2 from joint actions" points to the value 3.

PLAYING A NORMAL-FORM GAME

Idea: Players choose their actions at the same time.

- no communication with other agents
- no observation of other players' actions

Each player chooses a strategy σ_i which can be either:

- ① "Mixed" - probabilistic distribution over actions
- ② "Pure" - one action is always chosen

STRATEGY PROFILE

The "strategy profile" is the solution to a normal form game which outlines the strategy each agent plays.

We use σ_i to denote the strategy of player i .

We use σ_{-i} to denote the strategy of all players except i .

We use $u_i(\sigma)$ to denote the utility of agent i under strategy profile σ .

DOMINANT (STRATEGY)

We say for player i , a strategy σ_i "dominates" strategy σ'_i if

$$u_i(\sigma_i, \sigma_{-i}) > u_i(\sigma'_i, \sigma_{-i}) \quad \forall \sigma_{-i} \text{ &}$$

$$\exists \sigma_{-i} \text{ s.t. } u_i(\sigma_i, \sigma_{-i}) > u_i(\sigma'_i, \sigma_{-i})$$

A strategy is "dominant" if it dominates all other strategies.

DOMINANT STRATEGY EQUILIBRIUM / DSE

We say the strategy profile σ is a "DSE" if each player has a dominant strategy.

If a game has at least one DSE, then we say it is "dominance solvable".

BEST RESPONSE

Given a strategy profile σ_i, σ_{-i} , σ_i is a best response to the other agents' strategies σ_{-i} iff

$$u_i(\sigma_i, \sigma_{-i}) > u_i(\sigma'_i, \sigma_{-i}) \quad \forall \sigma'_i \neq \sigma_i$$

Note a rational agent will always play a best response.

NASH EQUILIBRIUM / NE

We say σ is a "Nash equilibrium" iff each agent i 's strategy σ_i is a best response to the other agent strategies σ_{-i} .

Alternatively, σ is a NE if no agent has any incentive to deviate from their current strategy σ_i .

SOLVING FOR NASH EQUILIBRIA

Method 1: Follow the chain of best responses until we reach a stable point; ie

- ① If some player is not playing a best response, switch to another strategy that is the best response.
- ② Repeat this until all players are playing the best response.

Method 2: Fix a strategy for one player & find the best response for the other.

PARETO DOMINANCE

We say an outcome σ "Pareto dominates" another outcome σ' iff

$$u_i(\sigma) \geq u_i(\sigma') \quad \forall i \quad \& \\ \exists i \text{ s.t. } u_i(\sigma) > u_i(\sigma')$$

PARETO OPTIMALITY

An outcome σ is "Pareto optimal" iff no other outcome σ' Pareto dominates σ .

MIXED STRATEGY NE

We say a mixed strategy σ is a NE if

$$E[u_i(\sigma_i, \sigma_{-i})] \geq E[u_i(\sigma'_i, \sigma_{-i})] \quad \forall \sigma'_i \neq \sigma_i$$

for each agent i .

NASH THEOREM

Every finite game has at least one (mixed) strategy that is a Nash equilibrium.

Chapter 19: Multi-Agent Reinforcement Learning

FRAMEWORK

Multi-agent Games + Sequential decision making



STOCHASTIC GAMES

Q₁: "Stochastic games" are N -agent MDPs.

Q₂: Components:

① N : # of agents

② S : shared state space

③ A^j : action space of agent j , $j=1,\dots,N$

④ r^j : reward function for agent j , $P(r^j|s, a^1, \dots, a^N)$

⑤ T : transition function, $P(s'|s, a^1, \dots, a^N)$

⑥ γ : discount factor, $0 \leq \gamma \leq 1$

⑦ h : horizon (# of time steps)

Q₃: Goal: find an optimal policy $\pi^* = \{\pi_1^*, \dots, \pi_N^*\}$ where

$$\pi_i^* = \underset{\pi_i}{\operatorname{argmax}} \sum_{t=0}^h \gamma^t E_{\pi} [r_i^t(s, a)],$$
$$a := \{a_1, \dots, a^N\}, \quad \pi := \{\pi_1^*, \dots, \pi_N^*\}$$

and $\pi_i : S \rightarrow \Sigma(A^i)$ (ie probability distribution over A^i).

Q₄: To play a stochastic game, players choose their actions at the same time, without communication or observation of other player's actions.

Q₅: At each state, all agents face a "stage game" (normal form game) with the values of the current state & joint action of each player being their utility.

OPTIMAL POLICY

Q₁: In MARL, the optimal policy should correspond to some equilibrium of the stochastic game.

Q₂: The most common solution concept is the "Nash equilibrium".

Q₃: We can define a "value function"

$$v_{\pi}^j(s) := \sum_{t=0}^{\infty} \gamma^t E_{\pi} [r_j^t | s_0 = s, \pi]$$

Q₄: Then, we say the stochastic game has a "Nash equilibrium" π^* iff

$$v_{(\pi_k^*, \pi_{-k}^*)}^j(s) \geq v_{(\pi_j^*, \pi_{-j}^*)}^j \quad \forall s \in S; \forall j; \forall \pi_j^*$$

INDEPENDENT LEARNING: NAIVE APPROACH

Q₁: Idea: Apply the single agent Q-learning directly to each agent.

Q₂: Limitations:

① Might not work well against opponents playing complex strategies;

② No guarantee of convergence;

③ Non-stationary transition & reward models.

COOPERATIVE SGs

Q: "Cooperative SGs" are those where the same reward function is shared across all agents.

OPTIMAL POLICY

Q: In this case, the equilibrium in the case of cooperative stochastic games is the Pareto dominating equilibrium.

OPPONENT MODELLING

Q₁: Idea: Each agent should maintain a belief over other agents' actions at the current state, as this is required to formulate its response.

Q₂: The method in which an agent accomplishes this is called "opponent modelling".

FICTITIOUS PLAY

Q₁: Idea: Each agent assumes that all opponents are playing a stationary mixed strategy.

Method:

- ① Agents maintain a count of the # of times another agent performs an action; ie

$$n_j(s, a_j) \leftarrow 1 + n_j(s, a_j) \quad \forall j, i$$

- ② Then, they update their "belief" about this strategy at each state according to

$$\mu_t^i(s, a_j) = \frac{n_t^i(s, a_j)}{\sum_{a_j'} n_t^i(s, a_j')} \quad \forall i, j$$

- ③ The agents can then calculate the best responses according to this belief.

LEARNING IN COOPERATIVE STOCHASTIC GAMES:

JOINT Q LEARNING / JQL

JointQLearning(s, Q)

Repeat

Repeat for each agent i

Select and execute a^i

Observe s' , r^i and a^{-i} , where $a^{-i} = \{a^1, \dots, a^{i-1}, a^{i+1}, \dots, a^N\}$

Update counts: $n(s, a) \leftarrow n(s, a) + 1$

Update counts: $n_t^i(s, a_j) \leftarrow 1 + n_{t-1}^i(s, a_j), \forall j$

Learning rate: $\alpha \leftarrow \frac{1}{n(s, a)}$

Update Q-value:

$$Q^i(s, a^i, a^{-i}) \leftarrow Q^i(s, a^i, a^{-i}) + \alpha(r^i + \gamma \max_{a^i} Q^i(s', a^i, \mu^i(s', a_1), \dots, \mu^i(s', a_N)) - Q^i(s, a^i, a^{-i}))$$

$s \leftarrow s'$

Until convergence of Q

Q₁: Idea: Modify & learning to include the opponents' action in the Q-updates.

Q₂: In particular, we want to find the "Nash Q function" for the game:

$$Q_*^i(s, a) = r^i(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V^i(s', T_1, \dots, T_N)$$

Q₃: This conveys the agent's immediate reward & discounted future rewards when all agents follow the Nash equilibrium policy.