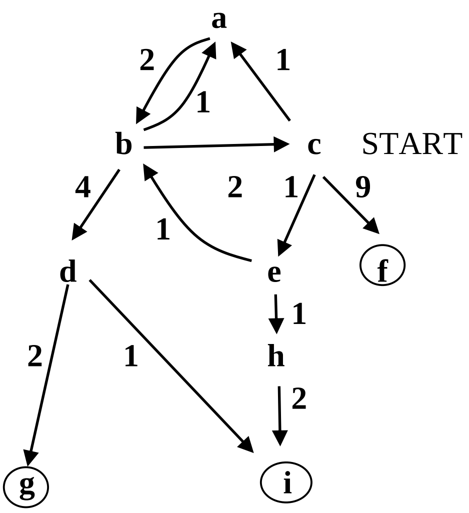
Artificial Intelligence, Master degree - Exam 5 September 2019

Instructions:
Time: 70 min. Use of literature, notes and electronic devices is not allowed. Please state your answers short and clear, answering the questions directly to the point.
Oral exam: Monday, 9 September at 12 pm

1. Consider the following state space:



Let **c** be the start state of search. **f**, **g** and **i** are goal states. Let search algorithms generate the successor nodes of a node in alphabetical order. For example, the order of successors of node **c** is: **a**, **e**, **f**.

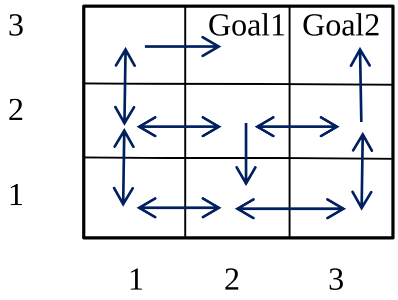
Assume that search algorithms A*, IDA* and RBFS detect cycles, and they immediately reject a node that completes a cycle. But they handle a graph as a tree. That is, if the algorithm reaches a node N by an alternative path then a copy N' of N is created and N' is treated as a new node. If two nodes have equal f-value then the node that was generated first is expanded first.

Heuristic values h of the nodes are given in the following table:

X	a	b	c	d	e	f	g	h	i
h(X)	1	2	1	1	5	0	0	8	0

(a) Which solution path is returned by algorithm A*?
(b) Which solution path is returned by algorithm IDA*? What is the value of f-limit in the last iteration of the execution of IDA* in this case?
(c) Which solution path is returned by algorithm RBFS?
(d) During the execution of RBFS in this case, what is the value of bound B for searching the subtree below node a? How is the backed-up value F(b) changing?
(e) Are there any nodes whose F-values are *inherited* from their parents during the execution of RBFS in this case? If yes, state these nodes.
(f) Which solution path is returned by the RTA* algorithm What are the stored h-values of nodes a and c?

2. The 2-dimensional grid 3x3 below specifies a reinforcement learning problem.



The arrows indicate possible actions and corresponding transitions between states. The actions are l, r, d, u (left, right, down, up). As shown in the figure above, not all actions are possible in every state. E.g. in state (2,2), actions r, l and d are possible. The system is deterministic except for the action »up« in state (3,2). Here action »up« causes transition upwards (into goal state (3,3)) with probability p=0.9, or transition to into state (2,2) with probability 0.1. In goal states Goal1 (2,3) and Goal2 (3,3) no transitions are possible. Rewards for all the transitions are equal 0, except for the following:

> Transition into any of the goal states is rewarded by 1
> Transition from (2,2) to (2,1) is rewarded by 2

The agent is rewarded by the discounted total cumulative reward; the discount factor gamma is 0.5. The agent is initially in the start state (1,2).

(a) What is the cumulative reward from the start state (1,2) of action sequence [u,r]?

(b) What is the cumulative reward from the start state (1,2) of infinite action sequence: [r,d,l,u,r,d,l,u,r,d,l,u….]. That is: repeat indefinitely the cycle of four actions [r,d,l,u].

(c) What is approximately the expected utility U_{32} of state (3,2) with the following policy: in state (3,2) do "up", and in state (2,2) do "right". (Note that doing "up" in (3,2) may result in transition to (2,2)).

(d) What is the optimal policy that gives maximum utility of state (1,2). Justify your answer, without necessarily calculating this utility.

3. (a) A technique called TD-learning is often used in reinforcement learning. What does TD stand for? State the TD update rule used in value-based learning (applied when a transition from state **s** to **s'** has been observed). What is the role of parameter **alpha** in this rule? What is an appropriate policy regarding the adjustment of **alpha** during learning?
(b) In reinforcement learning, what does GLIE refer to? What is the English phrase that is abbreviated as GLIE? Explain the basic idea of the GLIE scheme.
(c) In reinforcement learning, what is the difference between value-based learning and Q-learning: What is being learned in each of these two cases?

4. Consider a qualitative model of QSIM type. There are 5 variables in the system: X, DX,Y, VX, VY. The landmarks for these variables are:

X: minf, x0, 0, x1, inf
DX: minf, x0, 0, x1, inf
Y: minf, 0, y0, inf
VX, VY: minf, 0, inf

The constraints in the model are:

deriv(X, VX)
deriv(Y, VY)
plus(X, DX, x1/std) % X + DX = x1/std (DX is difference between landmark x1 and X)
VX = M_0^+(DX) % Monotonically inc. function with corresponding values (zero,zero)
VY = M_0^+(DX) % Monotonically inc. function with corresponding values (zero,zero)

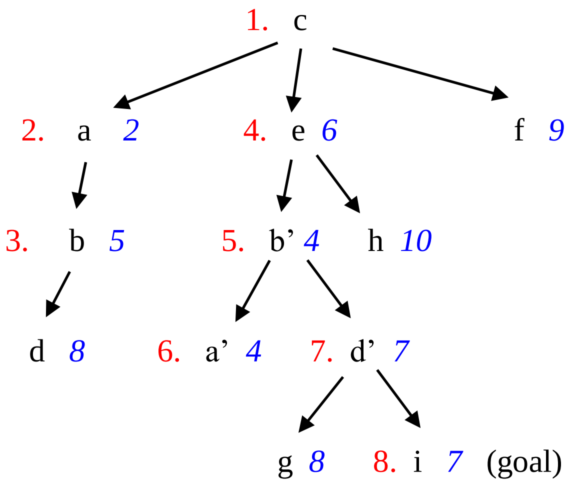
Initial values of X and Y in the start state at time t0 are: X(t0) = x0, Y(t0) = y0

(a) Determine the qualitative state of the system at time t0; that is qualitative magnitudes and directions of change of all the variables: X, Y, DX, VX, VY.
(b) Determine the qualitative state of the system at time interval t0..t1.
(c) Determine the qualitative state of the system at time t1.
(d) If we continue the simulation, the system eventually reaches a steady state. What is this state?

ANSWERS (WITH DETAILED SOLUTION OF PROBLEM 1)

1. (a) Solution path returned by A* is: c, e, b, d, i

Trace of A*: (Note: Not required as answer to this question)
Numbers in blue italics: f-values; Numbers in red Roman: order of expaned nodes



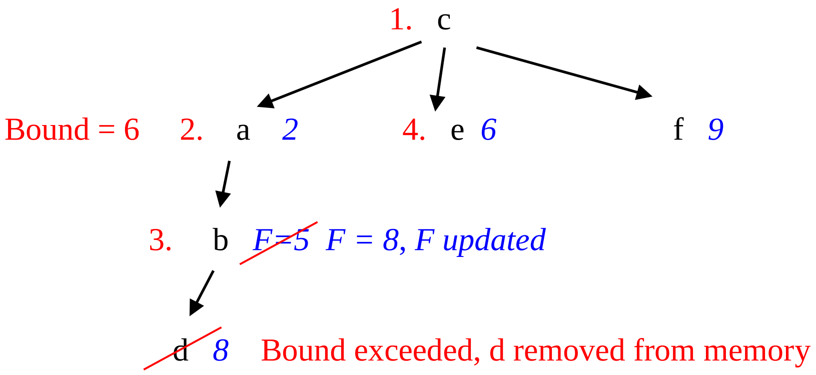
(b) Solution path returned by IDA*: c, e, b, d, i

Trace of IDA*:

f-limit	Visited nodes
1	c, a, e, f
2	c, a, b, e, f
5	c, a, b, d, e, f
6	c, a, b, d, e, b, a, d, h, f
7	c, a, b, d, e, b, a, d, g, i (goal)

(c) Solution path returned by RBFS: c, e, b, d, i
Note: No need to trace RBFS because RBFS always returns a path of equal cost as A*

(d) Answer: Bound for searching below a is 6; F(b) = 5, then F(b) = 8
This answer can be obtained by tracing part of executin of RBFS:


1. c
Bound = 6
2. a *2*
4. e *6*
f *9*
3. b ~~*F=5*~~ *F = 8, F updated*
~~d~~ *8*
Bound exceeded, d removed from memory

(e) Inheritance never occurs

(f) RTA* returns path: c, a, b, d, i
stored h(c) = 6, stored h(a) = infinity

Trace of RTA*:

Current node	Stored h of current node (i.e. 2nd best f)
c	6
a	infinity
b	8
d	2
i (goal)	

2.

(a) $U^{(u,r)}(s12) = 0.5$

(b) $U^{cycle}(s12) = 1.067$

(c) $U^{policy\ c}(s32) = 12/13 = 0.923$

(d) Optimal policy is cycle (r,d,l,u)

3.

(a) TD-rule for value learning: U(s) ← U(s) + alpha * (r(s,s’) + gamma * U(s’) – U(s))
Parameter alpha determines the degree of adjustment: greater alpha means more vigorous adjustment. Appropriate policy regarding alpha is: alpha decreases with number of visits of state s.

(b) GLIE = Greedy in the LImit of Exploration
This refers to the idea of appropriate tradeoff between exploration and exploitation: initially tend to explore, later tend to exploit (become greedy)

(c) In value-based learning, utilities U(s) are learned. In Q-learning, Q(s,a) values are learned (utility of performing action a in state s).

4.

	Time	X	Y	DX	VX	VY
(a)	t0	x0/inc	y0/inc	x1..inf/dec	0..inf/dec	0..inf/dec
(b)	t0..t1	x0..0/inc	y0..inf/inc	-“-	-“-	-“-
(c)	t1	0/inc	-“-	-“-	-“-	-“-
(d)	steady state	x1/std	y0..inf/std	0/std	0/std	0/std