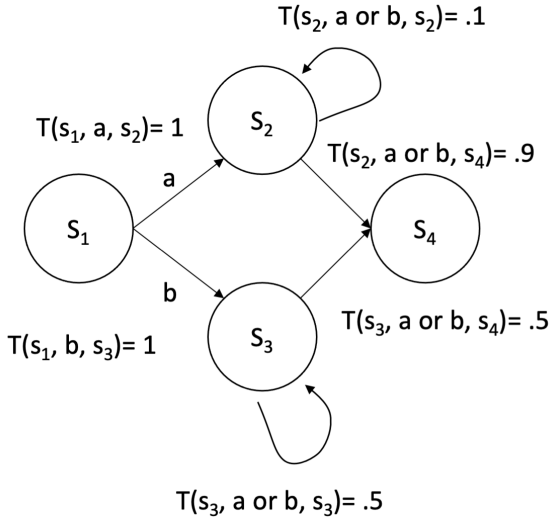
Practice Exam E

1 The best laid plans

1. (10 points) A *stochastic shortest paths* problem is a specific type of MDP in which

   - $\mathcal{S}$ and $\mathcal{A}$ are discrete sets of states and actions, as in a standard MDP.
   - There is a *goal set* $G \subset S$.
   - The transition function $T(s, a, s') = P(S_{t+1} = s' \mid S_t = s, A_t = a)$, is almost as usual, except that all states in G are absorbing; that is, for all $s \in G$, and all $a \in A$, $T(s, a, s) = 1$.
   - The reward function is $R(s, a, s') = 0$ for all $s \in G$ and $R(s, a, s') = -1$ otherwise (it can really be any negative value, but we will restrict our attention to this case).
   - The discount factor $\gamma = 1$.

   (a) (2 points) Here is a simple SSP.



   - $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$
   - $\mathcal{A} = \{a, b\}$
   - $\mathcal{G} = \{s_4\}$
   - $T(s_1, a, s_2) = 1$, $T(s_1, b, s_3) = 1$,
     $T(s_2, a \text{ or } b, s_2) = .1$, $T(s_2, a \text{ or } b, s_4) = .9$, $T(s_3, a \text{ or } b, s_3) = .5$, $T(s_3, a \text{ or } b, s_4) = .5$

   What is the optimal action to take in s_1?

   Solution: a

   (b) (2 points) What is the optimal Q function for the following state-action pairs? You can write an unevaluated numerical expression, but it may help to be reminded that $\sum_{i=1}^{\infty} ar^i = (ar)/(1-r)$.

   $Q(s_1, a)$ **-2.11**

   $Q(s_1, b)$ **-3**

$Q(s_2, a)$ **-1.11**

$Q(s_3, a)$ **-2**

(c) (2 points) Now consider the same SSP, but where we change just the transition function for s_3:

$$T(s_3, a \text{ or } b, s_3) = 1.0, \quad T(s_3, a \text{ or } b, s_4) = 0.0$$

What is the optimal Q function for the following state-action pairs?

$Q(s_1, b)$ $-\infty$

$Q(s_3, a)$ $-\infty$

(d) (2 points) Will value iteration converge on either or both of these SSPs? Explain.

Solution: It will converge on the first but not the second, for which, on every iteration, the value function will change by 1.

(e) (2 points) For each of the SSPs, is there a finite number of iterations after which you could terminate value iteration and extract an optimal policy? Explain your answer. (You don't have to provide a precise number).

Solution: There is a point at which the estimated value of action b is worse than the value of action a and after that, if we stop, the greedy policy will in fact be optimal.

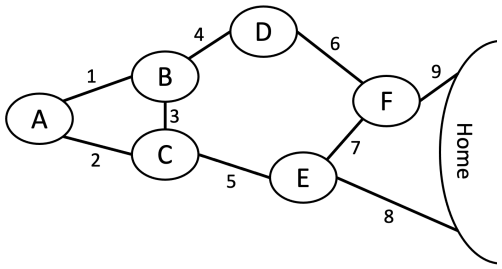
2 Random Bridges 1

2. You are on an archipelago (collection of islands) connected by drawbridges that go up and down. You always know what island you are on, but you don't know when the bridges will be up or down.

   Every hour, each bridge is randomly set to be up or down for that hour. Each reset is made independently (that is, whether the bridge was up on the previous time step doesn't affect whether it will be up on the next time step), but the probability of resetting to "up" versus "down" varies per bridge. Let p_i be the probability that bridge i is reset to up.

   On any island, your action choices are to walk over any bridge that's connected to it, or to wait. Bridges that are up cannot be crossed; if you try to go over a bridge that's up, you will stay on the island you are currently on. If you go over a bridge that's down, you will end up on the island at its other end.

   Crossing a bridge takes one hour. So, every hour, you attempt to cross a bridge, and then all bridges are immediately randomly reset. You want to minimize the time required to reach your **home** island. You therefore reward yourself -1 for every action taken and 0 for reaching the eternal resting (absorbing) home state. Here's a map of the archipelago:



(a) (5 points) Consider the policy

$$\pi = \{A : 1, B : 4, C : 5, D : 6, E : 8, F : 9\}$$

Write down an expression for $V_\pi(A)$, the value of being on island A and executing policy π, using only bridge-up probabilities $p_1, \ldots, p_9$.
Recall that, for $r < 1$, $\sum_{i=0}^{\infty} r^i = 1/(1-r)$.

Solution: Consider repeatedly trying to get over bridge 1. The value of that process is V, where

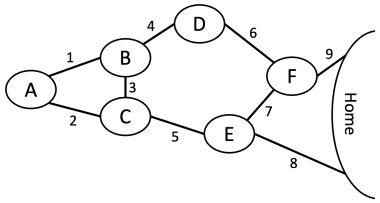
$$\begin{aligned} V &= -1 + (1 - p_1) \cdot 0 + p_1 * V \\ V(1 - p_1) &= -1 \\ V &= -1/(1 - p_1) \end{aligned}$$

Another way to see this is that the cost is

$$-1 - p_1 - p_1^2 - p_1^3 - \ldots$$

and use the result about the sum of a geometric series.
So:

$$-\left(\frac{1}{1-p_1} + \frac{1}{1-p_4} + \frac{1}{1-p_6} + \frac{1}{1-p_9}\right)$$



(b) (5 points) Write down an expression for $V^*(D)$ in terms of $p_1, \ldots, p_9$. You can also use $V^*(A)$, $V^*(B)$, $V^*(C)$, $V^*(D)$, $V^*(E)$, and/or $V^*(F)$ in your expression.

Solution:

$$V^*(D) = -1 + \max(((1 - p_4)V^*(B) + p_4 V^*(D)), ((1 - p_6)V^*(F) + p_6 V^*(D)))$$

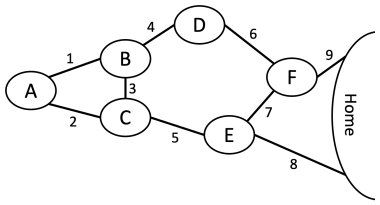
Which is also

$$V^*(D) = \max\left(\frac{-1}{1-p_4} + V^*(B), \frac{-1}{1-p_6} + V^*(F)\right)$$

(c) (3 points) Provide a set of values for $p_1, \ldots, p_9$ that would make the following policy optimal:

$$\pi = \{A : 1, B : 4, C : 3, D : 6, E : 8, F : 7\}$$

Solution: Let $p_5 = 0$, $p_9 = 0$ and everything else be 1.



(d) (3 points) Prove that this policy is optimal under the values you chose for $p_1, \ldots, p_9$.
Hint: you can prove that a policy π is optimal by first computing its value function V_π, and then prove that this value function V_π is a fixed point under value iteration.

Solution: Let $p_5 = 0$, $p_9 = 0$ and everything else be 1. The value function is

$$V = \{A : -5, B : -4, C : -5, D : -3, E : -1, F : -2\}$$

Given this value function, we can't improve our value at any of the islands:

- From A, bridge 1 has value $-1-4$ while bridge 2 has value $-1-5$, so we prefer 1.
- From B, bridge 1 has value $-1-5$ while bridge 3 has value $-1-5$ and bridge 4 has value $-1-3$, so we prefer 4.
- From C, bridge 2 has value $-1-5$, bridge 3 has value $-1-4$, bridge 5 has value $-1-5$, so we prefer 3
- From D, bridge 4 has value $-1-4$ while bridge 6 has value $-1-2$, so we prefer 6.
- From E, bridge 5 has value $-1-1$ while bridge 8 has value -1, so we prefer 9.
- From F, bridge 6 has value $-1-3$, bridge 7 has value $-1-1$ and bridge 9 has value $-1-2$, so we prefer 7.

3 Rusty Bridges

3. Now we are going to consider the same archipelago as in [delete this], but with a different model for how the bridges work. 200 years have passed, and the bridges have all rusted into some position, either up or down. However, communication infrastructure has also declined, and it is very foggy, and so you don't know any of the bridge positions with certainty.

   Let X_i be a random variable representing the position (up or down) of bridge i.

   (a) (6 points) For each of the following relationships among the bridges, draw a factor graph that has the fewest edges necessary to represent the distribution over the state of the bridges.

   i. Whether bridge i is down is independent of whether the bridge j is down, for all $i \neq j$.

   Solution: One factor for each bridge, connected just to that bridge.

   ii. You know that there is one and only one bridge that is not down.

   Solution: One factor connected to all the bridges.

iii. Bridges 4, 6, and 9 have been exposed to similar corrosive waves over the years.

> **Solution:** One factor connected to 4, 6, and 9, then three individual ones for the others.

For the rest of this problem, assume that whether bridge i is down is independent of whether bridge j is down, for all $i \neq j$. Let p_i denote the probability that bridge i is up.

As before, you want to make decisions about which bridges to cross. You decide to formulate your decision problem as a POMDP. You always know which island you are on. The initial belief state can be represented as $b_0 = (A, (p_1, p_2, \ldots, p_9))$. As before, if the bridge is up, you never be able to cross it, and if it's down, you will always be able to cross it. So, after you attempt to cross, you observe, with certainty, whether the bridge is up or down.

(b) (4 points) Consider belief state $b = (B, (1, 0, p_3, p_4, \ldots, p_9))$. What are the successor belief states under action 3, and their probabilities?

> **Solution:**
>
> - $(B, (1, 0, 1, p_4, \ldots, p_9))$ with probability p_3
> - $(C, (1, 0, 0, p_4, \ldots, p_9))$ with probability $1 - p_3$

(c) (1 point) Given fixed initial values $p_1, \ldots, p_9$, and starting from any island, how many possible belief states are there in this problem? Is it

- √ **less than** 6×3^9
- ◯ equal to 6×3^9
- ◯ greater than 6×3^9

(d) (3 points) Is belief state $(A, (p_1, p_2, 1, p_4, p_5, p_6, 1, p_8, p_9))$ reachable from $(A, (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9))$?

◯ Yes √ **No**

Explain your answer.

> **Solution:** We would have had to cross 1 or 2 in order to observe 3.

(e) (3 points) Now let's consider a strategy where, at each step, we use a most-likely-observation algorithm to find the best action. When there are multiple actions with the same return value, we pick the one with the smallest index (the bridge number). We execute the action and then replan based on the new observation. For what values of the p_i would the most-likely-observation strategy with re-planning fail to find a path even when there is indeed a policy with finite cost?

> **Solution:** When all the $p_i > 0.5$.

4 Simply unobservable

4. (4 points) (a) (2 points) Consider a POMDP in which the same observation is received, with probability 1, for every action in every state. In a problem with m states, n actions, and horizon h, characterize the size of an individual policy tree.

> **Solution:** It's a sequence of actions of length h.

(b) (2 points) In the class of POMDPs described above, the finite-horizon value function, as a function of the belief b, is still piecewise-linear and convex. In a problem with m states, n actions, and horizon h, what is the maximum number of pieces it could have?

Solution: n^h