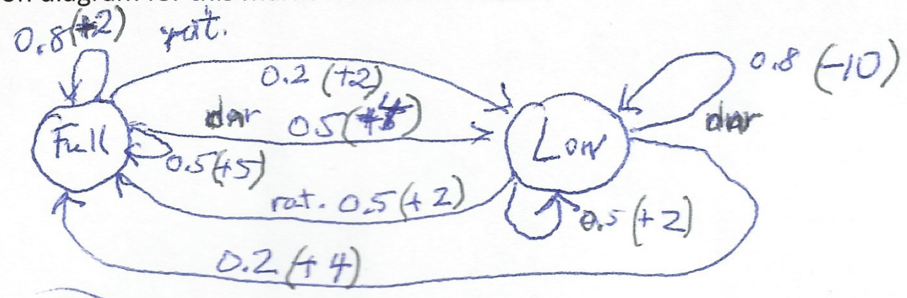
Without rationing, there is still some slight possibility (0.2) of the reservoirs filling up, and people are happy (but not as happy as when the reservoirs start out full, due to some folks having really low water).

But the worst happens when the reservoirs are low, there is no rationing, then the reservoirs stay low, and during that low-to-low period, many customers have outages. Here the reward is −10.

Draw a state-transition diagram for this Markov Decision Process:



The discount rate gamma is 0.5, reflecting the idea that making people happy today is more important than in the future, especially given the fact that the company doesn't yet have a contract with the state for the next biennium. Note that this is technically, however, an infinite-horizon MDP.

The company's agent is implementing rationing ALL THE TIME. What is it doing wrong?

Use the Value Iterations method to determine a simple approximation of the expected values of each state under the optimal policy that suggests what the problem is.

In particular, assume $V_0(s) = 0$ for each s in {F, L}, and then compute $V_1(s)$ using one step of the Bellman update.

Here is the formula for the Bellman Update operation:

$$V_{k+1}(s) = \max_a \Sigma_{s'} \; T(s, a, s') \, [\, R(s, a, s') + \gamma V_k(s') \,]$$

V0(F) = 0; V0(L) = 0;

V1(F) = max(T(F, rat, F) [R(F, rat, F) + gamma V0(F)] + T(F, rat, L) [R(F, rat, L) + gamma V0(L)],
T(F, dnr, F) [R(F, dnr, F) + gamma V0(F)] + T(F, dnr, L) [R(F, dnr, L) + gamma V0(L)])
= max(0.8 * [2 + 0.5 * 0] + 0.2 * [2 + 0.5 * 0], 0.5 * [5 + 0.5 * 0] + 0.5 * [4 + 0.5 * 0])
= max(2.0, 4.5)
= 4.5
(the better action here seems to be Do Not Ration)

V1(L) = max(T(L, rat, L) [R(L, rat, L) + gamma V0(L)] + T(L, rat, F) [R(L, rat, F) + gamma V0(F)],
T(L, dnr, L) [R(L, dnr, L) + gamma V0(L)] + T(L, dnr, F) [R(L, dnr, F) + gamma V0(F)])
= max(0.5 * [2 + 0.5 * 0] + 0.5 * [2 + 0.5 * 0], 0.2 * [4 + 0.5 * 0] + 0.8 * [-10 + 0.5 * 0])
= max(2.0, -7.2)
= 2.0
(the better action here seems to be Ration)

So on the basis of this 1-step approximation of the values iteration result, we see a disagreement between the Wa-Wa-Wa agent's policy and the policy we have determined.

Another step of value iteration would confirm that the above policy (Do not ration when reservoirs are mostly full; ration when reservoirs are low) is optimal.