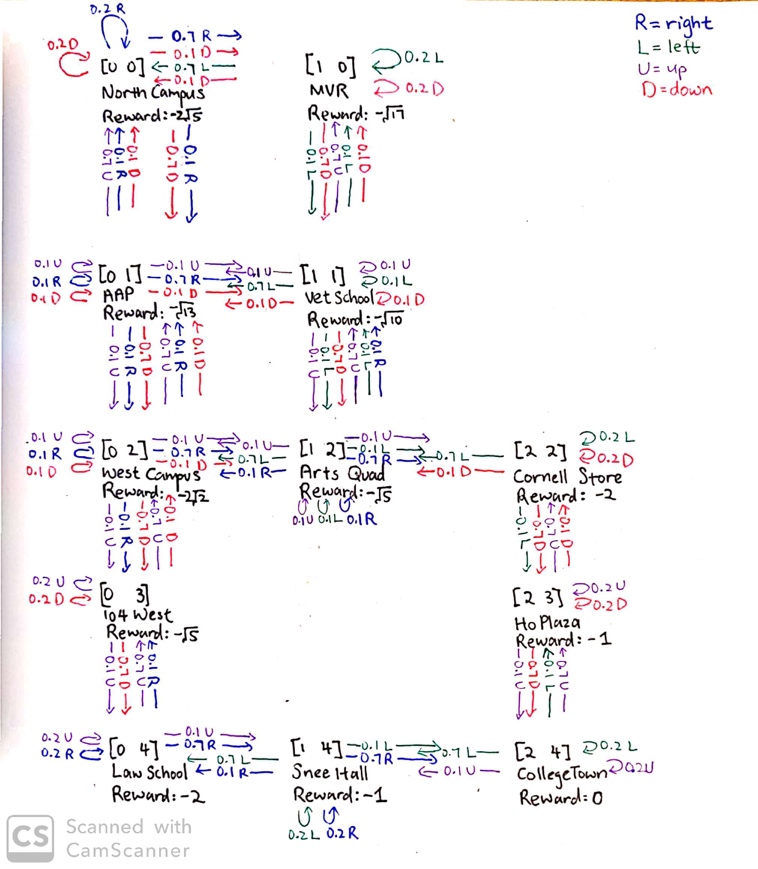
1. MPD and Utility Function
   1. 
   2. Let , , ,
2. Discount Rewards
   1. 1) If the environment does not contain a terminal state or if the agent never reaches one, then all environment histories will be infinitely long, and utilities will be infinite, so you can’t compare state sequences then.

2) Discounting favors near-term rewards, which is what we want to do, but you can’t do that if you’re just adding up all future rewards as is.

* 1. Lower bound is when

­

Upper bound is when

­

Anything between the two bounds is also finite.

* 1. You should **decrease** because now you have

1. Policy Iteration
   1. There are 2 states and 3 actions in this environment.
   2. :

Converge!

1. Q-learning
   1. The Q-value from the best action to take in each state should converge to the maximum expected utility of that state.
   2. If