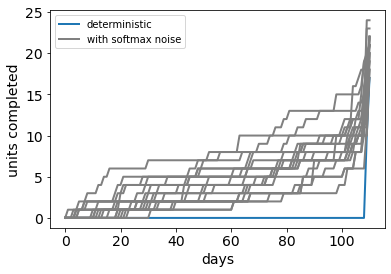
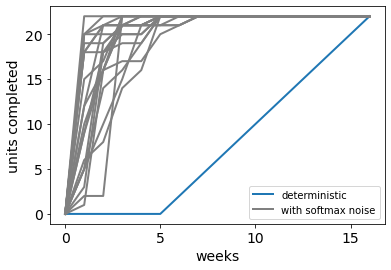
Task: complete at least 14 units of work in 110 days.   
Reward schedule: Rewards come at the deadline with a threshold of atleast 14 units and extra credit for upto 22 credits (but for a flatter rate).

Parameters: reward=20.0 for each unit upto 14, and then reward=5.0 for each unit until 22,  
 reward for each unit of shirk = 0.1,  
 effort work = -0.3  
States = No. of units completed (highest=30)  
Actions available in each state = no. of units to work (if action = do x units, 30-x units are ‘shirk’)  
Transition probabilities = binomial distribution given by the efficacy with which each unit can be completed

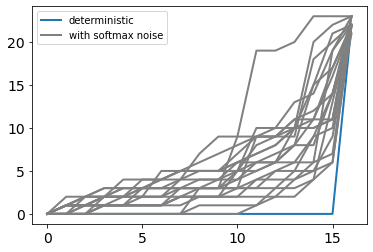
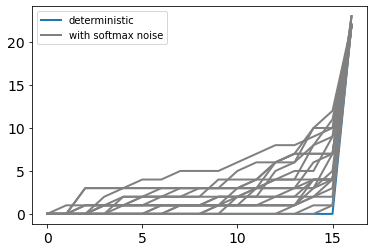
Example trajectory   
(deterministic policy and softmax policy with beta=5, discount = 0.9, efficacy = 0.5)  
  


* So the policy is to minimize effort by trying as many units as possible at the end (rather than spreading the effort out for example)
* Hence, the optimal policy is to do max no. of units even if there is no added benefit beyond 22 units, to maximize probability of finishing. This is a disadvantage of the structure, because people would do the units sequentially rather than trying 30 units parallelly and hope that some of them work out. For further simulations, I limit the maximum no. of units to 22 as a work-around.
* Even for a lower efficacy, the deterministic policy is to start a 3-4 days before deadline and that is enough. Students might actually be allocating work over weeks rather than days. Indeed, I can recover my model-agnostic clusters from data by aggregating days to weeks. So, now I switch to weeks.

Now, horizon = 16 weeks  
Reward = 4 for each unit upto 14 and then reward=1 for each unit until 22  
Max no. of units = 22  
softmax beta = 5  
Other parameters remain the same

When efficacy = 1.0 (equivalent to Zhang and Ma models with linear cost function)  
  
 discount = 1.0 

discount = 0.9 discount = 0.8

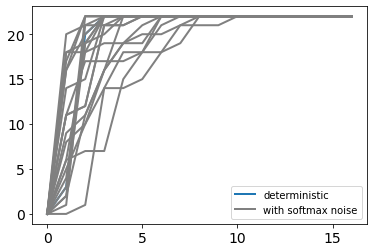
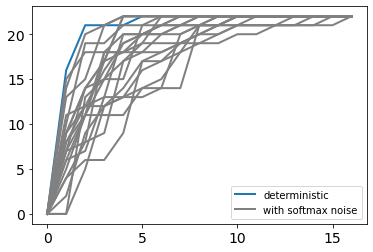
 

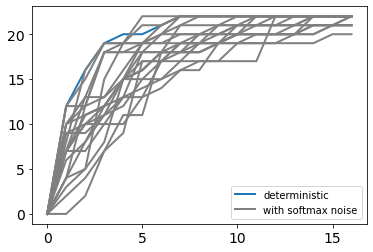
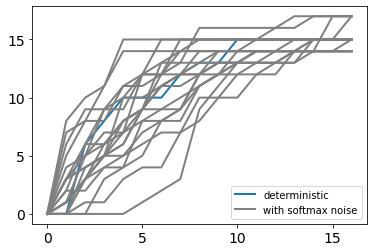
Sufficient to work in the last timestep because efficacy = 1 (tendency to put off is greater with larger discount rates)

Discount = 1.0, different efficacies

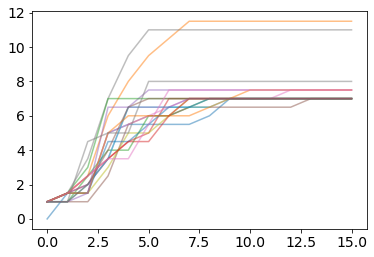
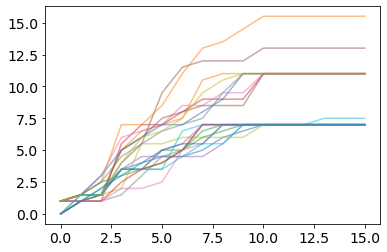
no preference for working now or later. So randomly choose at each timestep how many units to do. Average time of completion depends on efficacy.

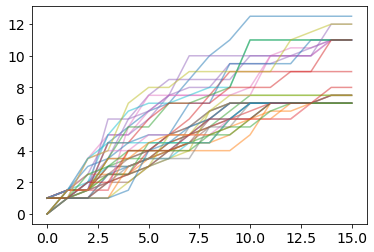
Efficacy = 0.9 efficacy = 0.7

  Efficacy = 0.5 efficacy = 0.2

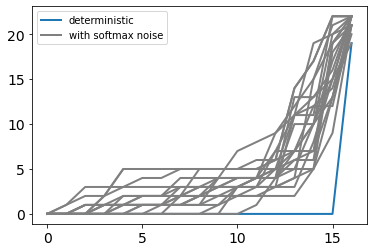
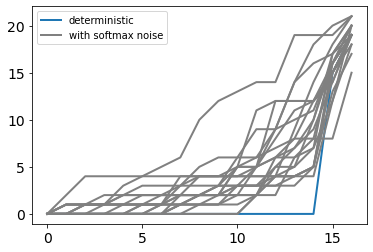
Matches some patterns from the data where most work is done early:



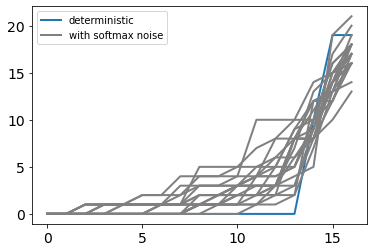
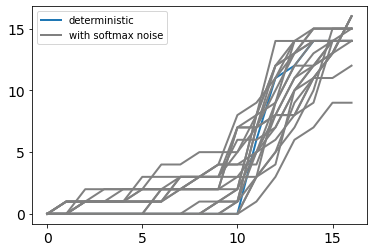
this works better with discount factor, limits on actions, and gaps between real and assumed efficacy:   


With discount = 0.9, work later. Starting time depends on efficacy

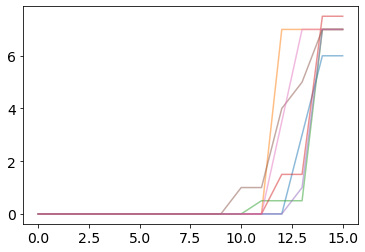
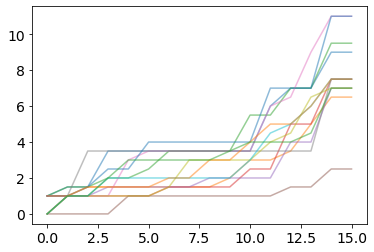
Efficacy = 0.9 efficacy = 0.7

Efficacy = 0.5 efficacy = 0.2

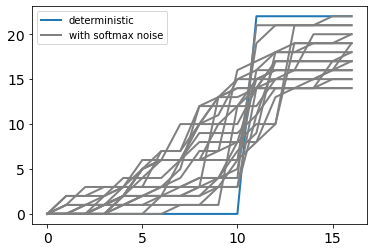
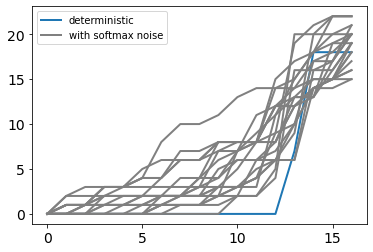
Matches clusters where work is put off:

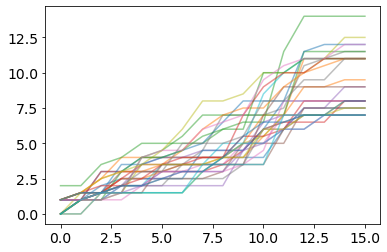
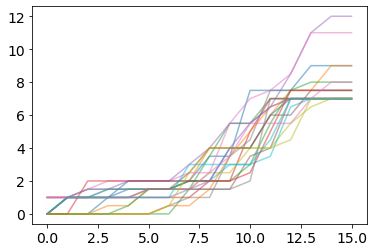
What if there is a discrepancy between real and assumed efficacies?

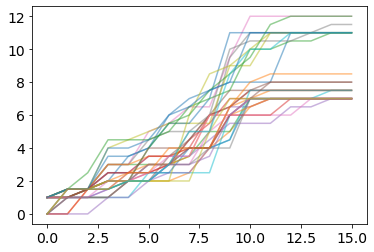
When real > assumed, agent completes earlier systematically

Real efficacy = 0.9, assumed efficacy = 0.2 real = 0.9, assumed = 0.3

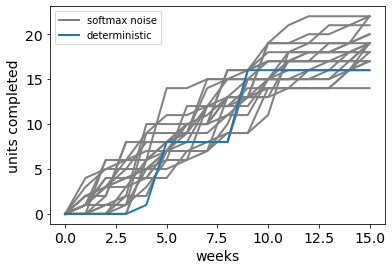
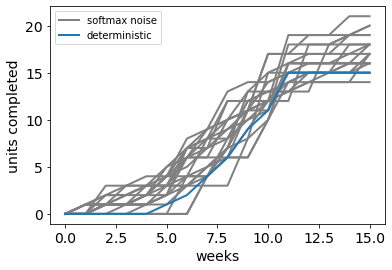
 

Matches clusters where work is put off but completed systematically before deadline (ramping up but not towards deadline but an earlier timestep)

  
  
  
The following is not explained by any of these, start promptly (no delays) but spread out and completed early. Works with limits on number of actions. Or with small convex exponents



real eff = 0.9, assumed = 0.2, reward = 6, r\_extra = 6/4   
 action\_lim = 8 exponent = 1.2

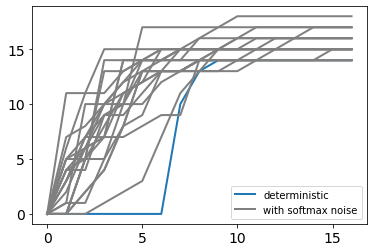
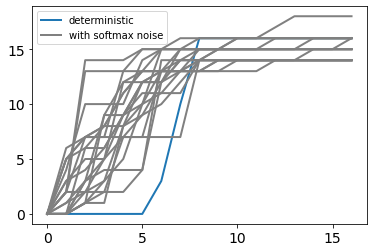
 

What if efficacy linearly reduces in time?

Efficacy = max\_efficacy(1 – t/horizon)

There are opposite tendencies from discounting and the decreasing efficacies: so deterministic policy is to start (and stop) sometime in between. Doesn’t reach full 22 credits most times due to low efficacy later on

Max\_efficacy = 0.9 Max\_efficacy = 0.6

Matches the clusters where work is finished earlier but not all of the credits are completed (because low efficacy later on).

Finally, the assumption here was that rewards are delayed. This might not be the case because once the threshold of 14 units is reached, students get the credits. So, these are actually immediate rewards! We have to turn to hyperbolic discounting/ differential discounts, and other mechanisms that don’t depend on a delay.