HW4

① So that y^π is an equilibrium point, we must prove that:

$$\dot{y}^\pi = 0$$

$$\dot{y}^\pi = \mathbb{E}_\pi\left[c_t + (\gamma - 1)\, y^\pi\right] =$$

$$= \mathbb{E}_\pi\left[c_t + (\gamma - 1)\frac{c_\pi}{1-\gamma}\right] =$$

$$= \mathbb{E}_\pi\left[c_t - c_\pi\right] = c_\pi - c_\pi = 0$$

↓

In average, policy π will lead to cost c_π!
(definition)

② $$\dot{E} = \underbrace{(y^t - y^\pi)}_{>0}(\dot{y}^t - \dot{y}^\pi)$$

Since the cost to go for an arbitrary policy must necessarily be higher than for the optimal policy.

On the other hand,

$$\dot{y}^{t} - \dot{y}^{\pi} = \mathbb{E}_{\pi}\left[c_t + (\gamma-1)y^{t}\right]$$

$$- c_{\pi} - (\gamma-1)y^{\pi} =$$

$$= \cancel{c_{\pi}} + (\gamma-1)y^{t} - \cancel{c_{\pi}} - (\gamma-1)y^{\pi} =$$

↳ for the same reason as before

$$= \underbrace{(\gamma-1)}_{<0}\underbrace{(y^{t} - y^{\pi})}_{>0}$$

$$\Rightarrow \dot{y}^{t} - \dot{y}^{\pi} < 0 \Rightarrow$$

$$\Rightarrow \dot{E} < 0$$

Along all trajectories of the ode described above.

③ If the trajectories of the TD algorithm follows indeed to ode described in this exercise, we know that the energy will decrease in each iteration, therefore our cost to go will monotonically converge to the optimal one.