a)

$$V(x,\theta_v) = \theta_v^T B(x)$$

$$\nabla V(x,\theta_v) = \frac{\partial[V(x,\theta_v)]}{\partial\theta_v} = \mathrm{B(x)}$$

b)

$$\pi(u|x,\theta) = \frac{e^{h(x,u,\theta_\pi)}}{\sum_a e^{h(x,a,\theta)}} \qquad h(x,\ u,\ \theta_\pi) = \theta_\pi^T \Psi(x,\ u)$$

$$\nabla In\pi(u_t|x_t,\theta_\pi) = \frac{\partial[\pi(U_t|x_t,\theta_\pi)]}{\partial\theta_\pi} = \psi(x_t,u_t)\frac{\sum_a e^{\theta_\pi^T\psi(x_t,u_t)} - e^{\theta_\pi^T\psi(x_t,u_t)}}{\sum_a e^{\theta_\pi^T\psi(x_t,u_t)}}$$

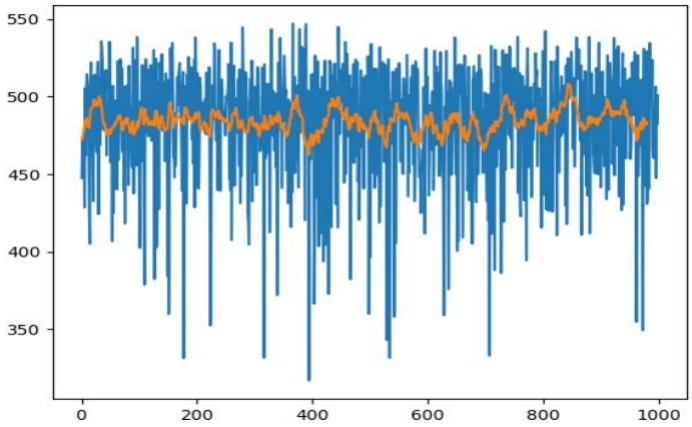
$$= \psi(x_t,u_t)[1-\pi(u_t|x_t,\theta_\pi)]$$

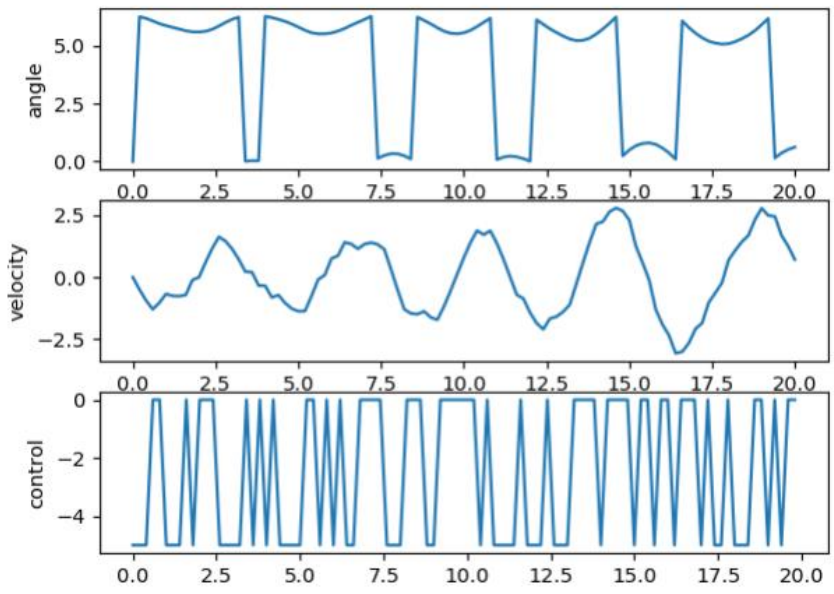
c)

policy_learning_rate = 0.00000001

Iteration times = 1000

It can't balance the pendulum appropriately



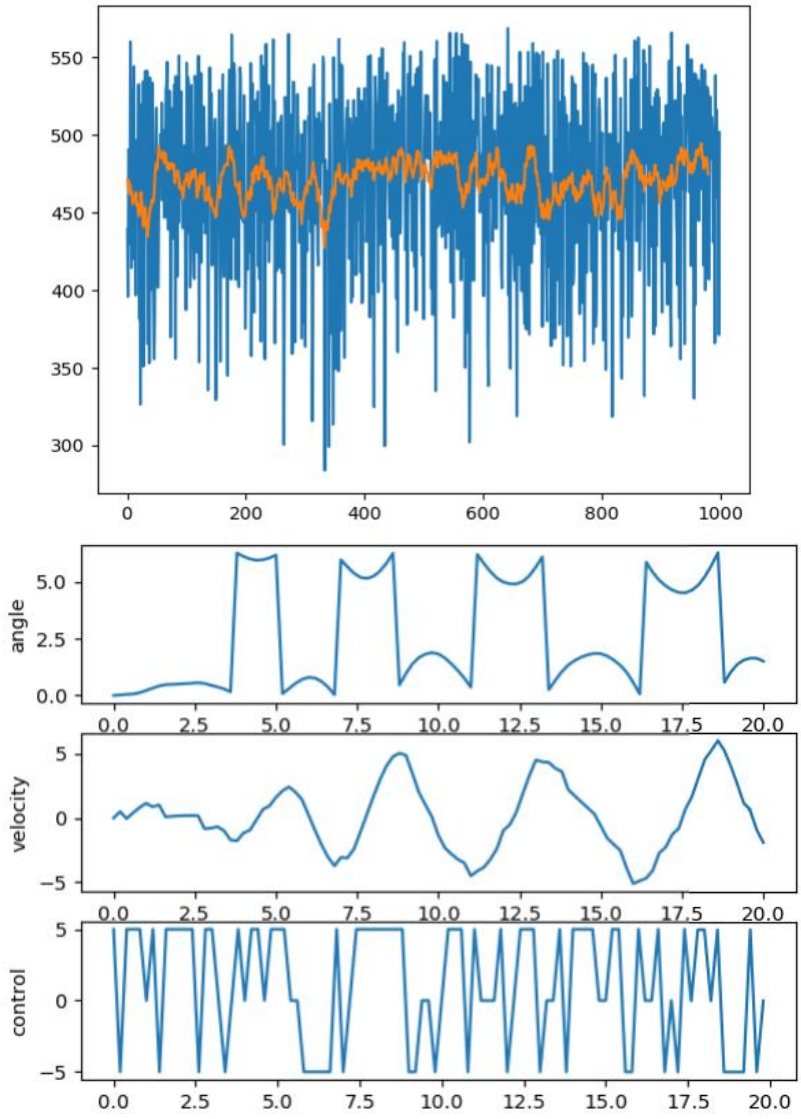


d)

policy_learning_rate = 0.00000001 value_learning_rate=0.01

Iteration times = 1000

It can balance the pendulum appropriately



e) The REINFORCE algorithm with baseline was easier to use.