A few problems with policy gradient algorithms (REINFORCE)

- The algorithm is not sample efficient, one trajectory is only used to update the policy once and it is then tossed away.
- the training can be unstable, because the gradient policy gradients are "offset reinforced" by the state action value

$$q_\pi(s,a) = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_T$$

So actions with outlier rewards can change the policy gradient substantiatlly and make the learning unstable.

A2C is another policy gradient based method that tackles the above problems. Main ideas of A2C areSample experiences from different randomizations of the environment in parallel, so that more experiences with more variety can be used to train the policy. Beside training a policy net (actor), we also train a critic (value net), to predict the value of the state $v(s)$. We can use it compute the "advantage" of the action