Setting: Markov Decision Process (MDP)

consists of:

S	possible states
A	possible actions
$P(s'\|s,a)$	transition probabilities
$R(s,a)$	(expected) reward (actual reward might be random)
$\gamma \in [0,1)$	discount factor
$\pi_\theta(a\|s)$	policy: a distribution over actions, parametrized by network weights θ

$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$ trajectory

Note: This is "Markov" because decision need only depend on the current state s, rather than on previous states.

Objective: Maximize the expected, discounted return

$$J(\theta) = E_{\tau \sim \pi_\theta}\left\{\sum_{t=0}^{\infty} \gamma^t r_t\right\}$$

(the above is the objective for the decision at $t=0$).

State-value function: $V^\pi(s) = E\left\{\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right\}$

i.e., the expected return when starting at state s and applying policy π.

Action-value function: $Q^\pi(s,a) = E\left\{\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a\right\}$

i.e., the expected return when starting at state s, taking action a, and then proceeding according to policy π.

Bellman equations:

$$V^\pi(s) = \sum_{a \in A} \pi(a|s) Q^\pi(s,a)$$

$$Q^\pi(s,a) = R(s,a) + \gamma E_{s' \sim P(\cdot|s,a)}\{V^\pi(s')\}$$

Advantage function: $A^\pi(s,a) \triangleq Q^\pi(s,a) - V^\pi(s)$

The advantage function measures the improvement in value (or deterioration, if negative) that you get from taking action a at state s, instead of following the policy π.

Policy gradient theorem:

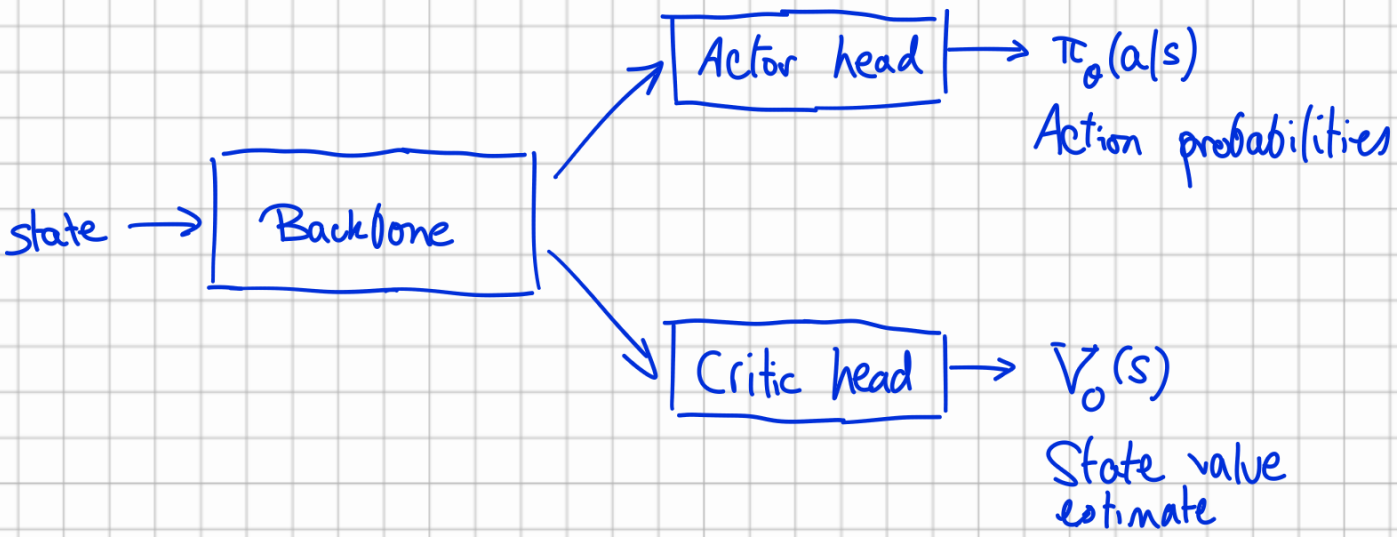
$$\nabla_\theta J(\theta) = E_{s,a\sim\pi_\theta}\left\{\left(\nabla_\theta \log \pi_\theta(a|s)\right) \cdot A^{\pi_\theta}(s,a)\right\}$$

Intuition:

Consider some state s, and suppose there is some action a for which $A^{\pi_\theta}(s,a) > 0$.
Then we ought to prefer a over the current policy's decision at s.
We do this by making the gradient of $\log \pi_\theta(a|s)$ positive at those a's, which will push $\pi_\theta(a|s)$ upwards at those points.

Actor-critic setup:

In practice, $V(s)$ and $Q(s,a)$ are unknown.
One approach is to train a network $V_\phi(s)$ to estimate $V(s)$, usually as a separate head ("critic") in addition to the head that makes the actual action decision.
This is called an actor-critic setup:



Unbiased Advantage Estimate

Define the one-step time-difference error:

$$\delta_t \triangleq \underbrace{r_t + \gamma V_\phi(s_{t+1})}_{\text{An estimate of } Q^\pi(s,a)} - V_\phi(s_t)$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\text{An estimate of } A^\pi(s,a)}$$

We can combine multiple time-difference errors to obtain:

$$A_1(s_t, a_t) \triangleq \sum_{l=0}^{\infty} \gamma^l \delta_{t+l}$$

$$= \sum_{l=0}^{\infty} \gamma^l \left(r_{t+l} + \gamma V_\phi(s_{t+l+1}) - V_\phi(s_{t+l}) \right)$$

$$= \underbrace{\sum_{l=0}^{\infty} \gamma^l r_{t+l}}_{\text{(discounted) return}} + \underbrace{\sum_{l=0}^{\infty} \gamma^{l+1} V_\phi(s_{t+l+1}) - \sum_{l=0}^{\infty} \gamma^l V_\phi(s_{t+l})}_{\text{all terms cancel out, except for } -V_\phi(s_t)}$$

Therefore, by the definition of Q^π,

$$E(\hat{A}_1(s_t, a_t)) = Q^\pi(s_t, a_t) - V_\phi(s_t) = A(s_t, a_t)$$

(by def. of $A(s,t)$, assuming that V_ϕ is an unbiased est of V.)

So, (assuming V_ϕ is unbiased), $\hat{A}_1(s,a)$ is an unbiased estimator of the advantage.

Generalized Advantage Estimate

The above estimator can be generalized to

$$\hat{A}_\lambda^{GAE}(s,a) \triangleq \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+1} \quad \text{for } \lambda \in [0,1]$$

Where the previous estimator was obtained with $\lambda = 1$.

This gives a bias-variance tradeoff:

- $\lambda = 0$: $\hat{A}_0 = \delta_{t+1}$

  This yields low variance since it depends only on the current timestep,

  but high bias because (I think) it depends on the specific choice of action at time t (and the corresponding reward r_t).

- $\lambda = 1$: This gives the unbiased estimate we saw before, but it has high variance because it is the sum of many instantiations of the policy r.v.

Typical values are $\lambda \sim 0.95$.

Proximal Policy Objective (PPO)

Recall that the gradient of the objective is given by (policy gradient theorem):

$$\nabla_\theta J(\theta) = E_{s,a\sim\pi_\theta}\left\{\left(\nabla_\theta \log \pi_\theta(a|s)\right)\cdot A_t^{\pi_\theta}(s,a)\right\}.$$

One problem with this approach is that the gradient can push the policy very far from its current value, causing inconsistency.

A common mitigation is to look at surrogate objective functions. In PPO these are based on the probability ratio $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$

which checks if the new policy (post-update) is more or less likely to take action a_t in state s_t.

Now, maximizing $J(\theta)$ (as shown above) is (almost? not sure) the same as maximizing

$$E\{r_t(\theta)\,\hat{A}_t(\theta)\}$$

But, we can now tweak this to limit the size of the change:

$$L^{clip}(\theta) = \hat{E}\left\{\min\left[r_t(\theta)\hat{A}(t),\ clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\cdot\hat{A}(t)\right]\right\}$$

for some $\varepsilon \sim 0.1$.

Intuition:

- When $\hat{A}>0$, if $r_t(\theta) \geq 1+\varepsilon$, this is $(1+\varepsilon)\hat{A}$;
  if $r_t(\theta) < 1+\varepsilon$, this is $r_t(\theta)\hat{A}$
  so the expression becomes $\min(r_t(\theta), 1+\varepsilon)\cdot\hat{A}$.
  Thus we clip the ratio at $1+\varepsilon$, limiting the incentive to make large changes to π_θ.
- When $\hat{A}<0$, if $r_t(\theta) \leq 1-\varepsilon$, this is $(1-\varepsilon)\hat{A}$;
  if $r_t(\theta) > 1-\varepsilon$, this is $r_t(\theta)\hat{A}$.
  So the expression becomes $\max(1-\varepsilon, r_t(\theta))\cdot\hat{A}$,
  which prevents us from too much penalty on bad actions.

The PPO objective actually contains two additional components:

$$L^{PPO}(\theta) = \hat{E}\left\{ L^{clip}(\theta) - c_1 L_t^{VF}(\theta) + c_2 L^{ent}(\theta) \right\}$$

clipped probability ratio obj (see above)

value function loss

entropy bonus

Valve function loss: MSE estimate for the critic head:

$$L_t^{VF}(\theta) = \left(V_\phi(s_t) - V_t^{obs}\right)^2$$

where V_t^{obs} is computed as the actual discounted return obtained from time t until the end of the episode.

Entropy bonus: Gives a small bonus to policies with distributions closer to uniform, since such policies encourage exploration. The bonus is computed as the entropy of the policy at state s_t:

$$L_t^{ent}(\theta) = -\sum_{a \in A} \pi_\theta(a|s_t) \log \pi_\theta(a|s_t)$$

Typical values: $c_1 \in [0.5, 1]$

$c_2 \sim 0.01$