The equivalence of options and subgoals. Both specify likely future events and allow us to construct temporally abstract actions. Can we show an equvalence between these approaches?

In one case we condition the policy on a subgoal, and in the other case we condition with an option.

Based on Feudal nets (Fun) and The option-critic architecture (OpC).

OpC

$$Q_\Omega(s,w) = E_{a\sim\pi_\omega(s)}[Q_\omega(s,w,a)] \tag{1}$$
$$Q_\omega(s,w,a) = r(s_t) + \gamma E_{s'\sim\tau(s,\pi(s))}[U(s',w)] \tag{2}$$
$$U(s,w) = (1-\beta(s,w))Q_\Omega(s,w) + \beta(s,w)V_\Omega(s) \tag{3}$$

(these equations are copied from the OpC paper.)

FuN (with additions)

- $Q_\Omega(s,g)$ is the expected discounted reward of using π_ω to reach subgoal g and then following π_Ω afterwards.
- $Q_\omega(s,g,a)$ is the expected discounted reward of choosing an action given that we are attempting to achieve some goal, and the future ability to achieve subgoals.

$$Q_\Omega(s_t,g_t) = r(s_t) + \gamma \underset{s_{t+1}\sim\tau(s_t,\pi_\omega(s_t,g))}{E}[Q(s_{t+1},\pi_\Omega(s_{t+1}))] \quad \text{(manager)}$$
$$r_\omega(s_t,g_t) = Q_\Omega(s_{t-1},g_t) - \gamma Q_\Omega(s_t,g_t) \quad \text{(manager rewards the worker)}$$
$$Q_\omega(s_t,g_t,a_t) = r_\omega(s_t,g_t) + \gamma E_{a\sim\pi(s')}[Q_\omega(s_t,g_t,a_t)] \quad \text{(worker)}$$
(4)

(1) This is a non-standard definition. $r_\omega(s_t,g) = E_{s'\sim\tau(s,\pi_\omega(s,g))}[Q_\Omega(s',g)] - Q_\Omega(s_t,g)$ But can we justify it? (*intuitively it makes sense that we would reward the worker if it increases the expected rewards!? NO! Other way around. Any decrease in expected value means the reward must have been received at the current time step.*) To be able to caluclate this we will approximate it with $r_\omega(s_t,g) = Q_\Omega(s_{t+1},g) - Q_\Omega(s_t,g)$.

(1.1) As long as s_{t+1} is sampled IID from $\tau(s_t,\pi(s_t,g_t))$ then this estimator should have zero bias (but we have introduced more variance).

(1.2) Other approaches to subgoals "cheat" in the sense that they use the euclidean distance between the current state and goal state (see TDMs).

(1.3) This assumes the estimated values are perfectly fit to the current policy.

can come up with a counter example!? - in state A, and pick B/C with 50/50 chance. B is rewarded with 1 and C -1. $V(A) = 0, V(B) = 1, V(A) - \gamma V(B) = -1\gamma$ - in state A, policy always picks B from B/C. A is rewarded with 0.5 and B 0.5 and C -1. $V(A) = 0.5 + 0.5\gamma, V(B) = 0.5, V(A) - \gamma V(B) = 0.5 + 0.5\gamma - 0.5\gamma = 0.5$

(2) The introduction of the β is also non-standard for the Feudal net framework. But can we justify it? Feudal networks are implemented by running the manager at a lower temporal resolution than the worker. Thus the worker may recieve the same goal for k steps. We can use β to index the correct discounted reward. Maybe the worker policy is simply following the current subgoal, and thus the expected discounted reward is the value of the workers policy for the next k steps, in which case β should be zero for these k steps. Else, for β equals one, the manager picks a new subgoal, in which case we can recursively define it value as the value of the rolledout workers actions.

Equivalence

$$\text{Let } \; s' = \tau(s, w) = g \tag{5}$$

$$Q_\omega(s, w, a) = \sum_{t=0}^{T-1} \gamma^t r(s_t) + \gamma^T V(s') \tag{OpC}$$

$$= Q_\omega(s, g, a) \tag{6}$$

(7)

Generalisation/calc interpretation

> Intutition: Option-critics take the integral of Q_ω to construct Q_Ω while feudal networks take the derivative of Q_Ω to construct Q_ω *(assuming the introduction of (1), (2))*.

$$\text{definition of option manager} \tag{8}$$

$$Q_\Omega(s, w) = \int p(a|s, \pi_\omega) Q_\omega(s, w, a) da \tag{9}$$

So what is this reward, r_w? A way to think about it could be as the ...!?

(3) a_t is via greedy policy
(4) at convergence of Q to the true value of the current policy

$$
\begin{aligned}
r_\omega(s_t, g) &= Q_\Omega(s_{t-1}, g) - \gamma Q_\Omega(s_t, g) && \text{(by defn)} \\
&= r(s_t) && \text{(assuming (3), (4))} \\
Q_\Omega(s_{t+1}, g) - Q_\Omega(s_t, g) &\approx \frac{\partial Q_\Omega}{\partial t} && (10) \\
Q_\omega(s, g, a) &= E[\gamma^t r(s_t)] && (11) \\
&= \int \gamma^t r(s_t) dt && (12) \\
\int Q_\omega(s, g, a) da &= && (13)
\end{aligned}
$$

Comparison

pros/cons

- (con - options) calculating an integral for every choice…
- (con - subgoals) manager must be accurate, else can introduce bias/variance
- ?

TODOs

- Is there a nice way to visualise this!? Pictures!?
- ?