Some Derivations

MinKu Kang

July 11, 2018

$$\max\ [-H(x_{t+1}) + H(x_t)] \approx \max\ |\dot{x}_t|\Delta t = v_{max}\Delta t \tag{1}$$

$$\tag{2}$$

$$\begin{aligned} R^{max} =& r_0^{max} &+ r_1^{max} &+ ... + r_{T^*}^{max} &+ c_{T^*+1} + ... + c_\infty \\ R' =& r_0 &+ r_1 &+ ... + r_{T^*} &+ r_{T^*+1} + r_{T^*+2} + ... + r_{T^*+k} + c_{T^*+k+1} + ... + c_\infty \end{aligned}$$

$$\begin{aligned} R^{max} - R' = \underbrace{(r_0^{max} - r_0) + (r_1^{max} - r_1) + ... + (r_{T^*}^{max} - r_{T^*})}_{\geq 0} + \\ \underbrace{(c_{T^*+1} - r_{T^*+1}) + (c_{T^*+2} - r_{T^*+2}) + ... + (c_{T^*+k} - r_{T^*+k})}_{\geq 0?} \end{aligned}$$

One simple solution to make the second part always greater than zero is:

$$c_t > r_t \ \ \forall t.$$
$$c_t := r^{max}$$

Rather, one would be tempted to set c arbitrary large, but this can make the return arbitrary large, adversely affecting learning stability. Thus, we want to find an accurate estimate of the max reward, $\hat{r}^{max}$ to set c appropriately.

$$\begin{aligned} R_0^{max} &= r_0^* + r_1^* + ... + r_{T^*}^* \\ R_0 &= r_0 + r_1 + ... + r_{T^*} \end{aligned}$$

$$\begin{aligned} R_0^{max} - R_0 = r_0^* - r_0 + R_1^{max} - R_1 > 0 \\ \text{by the Priciple of Optimality,} \\ R_1^{max} - R_1 > 0. \\ \text{And note that,} \\ r_0^* - r_0 > -(R_1^{max} - R_1) < 0 \end{aligned}$$

This implies that $r_0^* - r_0$ can be negative, meaning that r_0 can be larger than r_0^*

$$R_0^{max} - R_0 = (r_0^* - r_0) + (r_1^* - r_1) + ... + (r_{T^*}^* - r_{T^*}) +$$
$$\underbrace{(c_{T^*+1} - r_{T^*+1}) + (c_{T^*+2} - r_{T^*+2}) + ... + (c_{T^*+k} - r_{T^*+k})}_{D_{1:k}}$$

By the Principle of Optimality, the following conditions should be satisfied:

$$\begin{aligned} D_{k:k} &> 0 \\ D_{k-1:k} &> 0 \\ &\cdots \\ D_{1:k} &> 0 \end{aligned}$$

A simple setting that can satisfy the above conditions is to set c_t as:

$$c_t := r^{max} \quad \forall t.$$

However, it is sometimes not trivial to find r^{max}, which might require the knowledge of dynamics model of the environment. For now, let us assume that we know r^{max}.

Now, let us consider the Reward Shaping term:

$$F(s_{t+1}, s_t) = \gamma\phi(s_{t+1}) - \phi(s_t).$$

This can be further expanded such that:

$$\begin{aligned} F(s_{t+1}, s_t) &= \phi(s_{t+1}) - \phi(s_t) + \gamma\phi(s_{t+1}) - \phi(s_{t+1}) \\ &= \phi(s_{t+1}) - \phi(s_t) + (\gamma - 1)\phi(s_{t+1}) \end{aligned}$$

Let us design a reward function such that:

$$\begin{aligned} r(s_t, a_t, s_{t+1}) &= r_{original}(s_t, a_t, s_{t+1}) + F_t \\ r_{original}(s_t, a_t, s_{t+1}) &:= -(\gamma - 1)\phi(s_{t+1}) = \alpha\phi(s_{t+1}),\ \alpha > 0 \\ \implies r(s_t, a_t, s_{t+1}) &= \phi(s_{t+1}) - \phi(s_t) \end{aligned}$$

Since the shaping term F_t does not affect the optimal policy, the optimal policy will maximize the discounted sum of $r_{original}$:

$$\sum_{i=0}^{\infty} \gamma^i \, r_{original}(s_i, a_i, s_{i+1}) = \sum_{i=0}^{\infty} \gamma^i \, \alpha\phi(s_{t+1}) = \alpha \sum_{i=0}^{\infty} \gamma^i \, \phi(s_{t+1})$$