Note

Wednesday, August 12, 2020 8:48 PM

Define:

$r(s,a)[i]$ as the reward obtained by performing action a at state s for the ith time

$n(s,a)$ is the number of times performing action a at state s

$R^{\wedge}(s,a) = \frac{1}{n(s,a)} \sum_{t=0} r(s,a)[i]$, is the avearage reward obtained by performing action a at state s

In many reinforcement learning problems, taking the same action a at state s can result in different reward r(s, a). Objective of a reinforcement learning problem is to maximize the discounted returns of average rewards

$$\mathrm{E}_{s_0,a_0,\ldots.\sim\boldsymbol{\pi}}\left(\sum_{\mathrm{t=0}} \gamma^t R^{\wedge}(s_t,a_t)\right)$$

$R^{\wedge}(s_t,a_t)$ here has a confidence interval

$\mathrm{CI}(R^{\wedge}(s_t,a_t)) := [\, R^{\wedge}(s_t,a_t) - \boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{n(s,a)}},\ R^{\wedge}(s_t,a_t) + \boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{n(s,a)}}]$ that we are $1 - \delta^{R}_{n(s,a)}$ confident with.

By Hoeffding's Inequality:

$$\boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{n(s,a)}} = \sqrt{\frac{\ln\frac{2}{\delta^{R}_{n(s,a)}}}{n(s,a)}}$$

The count-based exploration bonus method proposed to maximize the discounted returns of upper tail of the confidence interval average rewards

$$\mathrm{E}_{s_0,a_0,\ldots.\sim\boldsymbol{\pi}}\left(\sum_{\mathrm{t}=0}\gamma^t\left(\hat{R}(s_t,a_t)\ +\ \boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{n}(\boldsymbol{s},\boldsymbol{a})}\right)\right)$$

which transforms the estimate of of $\tilde{Q}(s,a)$ from

$$\tilde{Q}(s,a) = \hat{R}(s,a) + \gamma \sum_{s'} T(s'|s,a) \max_{a'} \tilde{Q}(s',a')$$

to be

$$\tilde{Q}(s,a) = \left(\hat{R}(s,a) + \boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{n}(\boldsymbol{s},\boldsymbol{a})}\right) + \gamma \sum_{s'} T(s'|s,a) \max_{a'} \tilde{Q}(s',a')$$

And we use $\frac{\beta}{\sqrt{n(s,a)}}$ to represent $\boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{n}(\boldsymbol{s},\boldsymbol{a})} := \sqrt{\frac{\ln\frac{2}{\delta^R_{n(s,a)}}}{n(s,a)}}$ in computation