CME 242 : Bulk

16 mars 2020

Reinforcement Learning

Monte-Carlo learning for model-free prediction

In Monte-Carlo learning for model-free prediciton, the incremental step in the every-visit method can be written as :

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t))$$

Exercise : Show that for a fixed learning-rate, the MC method is equivalent to an exponentially decaying average of the episodes returns.

Answer :

Let $s \in \mathcal{S}$. Using the iteration formula, we can write :

$$\begin{aligned} V_n(s) &\leftarrow V_{n-1}(s) + \alpha(G_t - V_{n-1}(s) \\ &= \alpha G_n + (1-\alpha)V_{n-1}(s) \\ &= \alpha G_n + (1-\alpha)(\alpha G_{n-1} + (1-\alpha)V_{n-2}(s)) \\ &= \alpha G_n + \alpha(1-\alpha)G_{n-1} + (1-\alpha)^2 V_{n-2}(s) \\ &= \ldots \\ &= \sum_{i=0}^{n-1} \alpha(1-\alpha)^i G_{n-i} \end{aligned}$$

Then, we remark that $(\alpha(1-\alpha)^{n-i})_{1\leq i\leq n}$ are weights when $n \to \infty$ that exponentially decay. Indeed, we can write by assuming that $|1-\alpha| < 1$:

$$\begin{aligned} \sum_{i=1}^{+\infty} \alpha(1-\alpha)^i &= \alpha \frac{1}{1-(1-\alpha)} \\ &= 1. \end{aligned}$$

Greedy-ϵ theorem

For any ϵ-greedy policy π, the ϵ-greedy policy π' with respect to q_π is an improvement $v_{\pi'}(s) \geq v_\pi(s)$

Proof :

Let s a state.

$$
\begin{aligned}
q_\pi(s,\pi'(s)) &= \sum_{a\in\mathcal{A}} \pi'(a|s) q_\pi(s|a) \\
&= \epsilon/m \sum_{a\in\mathcal{A}} q_\pi(s|a) + (1-\epsilon) max_{a\in\mathcal{A}} q_\pi(s,a) \\
&\geq \epsilon/m \sum_{a\in\mathcal{A}} q_\pi(s|a) + (1-\epsilon) \sum_{a\in\mathcal{A}} \frac{\pi(a|s)-\epsilon/m}{1-\epsilon} q_\pi(s,a) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) q_\pi(s|a) \\
&= v_\pi(s).
\end{aligned}
$$

Indeed, since $\sum_{a\in\mathcal{A}} \pi(a|s) = 1$ and there is m actions in $\mathcal{A}$, we have :

$$
\begin{aligned}
(1-\epsilon) \sum_{a\in\mathcal{A}} \frac{\pi(a|s)-\epsilon/m}{1-\epsilon} q_\pi(s,a) &\leq (1-\epsilon) max_{a\in\mathcal{A}} q_\pi(s,a) \sum_{a\in\mathcal{A}} \frac{\pi(a|s)-\epsilon/m}{1-\epsilon} \\
&= (1-\epsilon) max_{a\in\mathcal{A}} q_\pi(s,a) \frac{1-m\epsilon/m}{1-\epsilon} \\
&= (1-\epsilon) max_{a\in\mathcal{A}} q_\pi(s,a)
\end{aligned}
$$

Finally, by using the Policy improvement theorem, we have the result.

Greedy in the Limit with Infinite Exploration (GLIE)

— All state-action pairs are explored infintely many times

$$\boxed{\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \quad lim_{k\to+\infty} N_k(s,a) = \infty}$$

— The policy converges on a greedy policy :

$$\boxed{lim_{k\to+\infty} \pi_k(a|s) = \mathbf{1}(a = argmax_{a'\in\mathcal{A}}(q_k(s,a')))}$$