Date: Sub:

سعیده شفاء منزوی ۹۹۴۱۱۴۲۸ تمرین کلاسی 2 :

مبنای این بحث تئوری ROBBINS-MONRO یا الگوریتم ROBBINS-MONRO است. یک تئوری بیان می‌کنیم و سپس آن را اثبات می‌کنیم.

فرض کنید که یک FMDP داریم با فضای state برابر با $\mathcal{X}$ و فضای اکشن A و تحریک احتمال حالت احتمال P و پاداش r.

نحوهٔ آپدیت Q به صورت زیر است:

$$Q_{t+1} = Q_t + \alpha_t [r_t - Q_t]$$

که برای $Q_{learning}$ به طور خاص داریم (با حالت خاص فرمول مشابه بالا می‌شود)

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t(x_t, a_t)[\ I\ ]$$

$$I = r_t + \gamma \max_{b \in A}(Q_t(x_{t+1}, b)) - Q_t(x_t, a_t)$$

برای همگرایی می‌دانیم که لازم است برای هر (x, a) عضو $X \times A$ (ضرب دکارتی این دو مجموعه) رابطه مقابل برقرار باشد:

① $\sum_t \alpha_t(x, a) = \infty$

② $\sum_t \alpha_t^2(x, a) < \infty$

و $0 \le \alpha_t(x, a) < 1$

تئوری بعدی نیز این است که یک فرآیند تصادفی Δ_t مقادیر حقیقی خواهد داشت

$$\Delta_{t+1}(x) = (1 - \alpha_t(x))\Delta_t(x) + \alpha_t(x) F_t(x)$$

و به صفر همگرا می‌شود اگر:

1) $0 \le \alpha_t \le 1$, $\sum_t \alpha_t(x) = \infty$ و $\sum_t \alpha_t^2(x) < \infty$

2) $\| \mathbb{E}[F_t(x) \mid \mathcal{F}_t] \|_W \le \gamma \|\Delta_t\|_W$, with $\gamma < 1$

3) $\mathrm{var}[F_t(x) \mid \mathcal{F}_t] \le C(1 + \|\Delta_t\|_W^2)$ for $C > 0$

Date: Sub:

سعید مسعودی فردی 99411427

اثبات قضیه ۱:

$$Q_{t+1}(x_t,a_t) = (1-\alpha_t(x_t,a_t))Q_t(x_t,a_t) + \alpha_t(x_t,a_t)[r_t + \gamma \max_{b\in A} Q_t(x_{t+1},b)]$$

ابتدا $Q^*(x_t,a_t)$ را کم می‌کنیم:

$$\Delta_t(x_t,a_t) = Q_t(x_t,a_t) - Q^*(x_t,a_t)$$

داریم:

$$\Delta_{t+1}(x_t,a_t) = (1-\alpha_t(x_t,a_t))\Delta_t(x_t,a_t) + \alpha_t(x_t,a_t)[F_t]$$

$$F_t = r_t + \gamma \max_{b\in A} Q_t(x_{t+1},b) - Q^*(x_t,a_t)$$

اگر مقدار زیر را بنامیم:

$$F_t(x,a) = r(x,a,X(x,a)) + \gamma \max_{b\in A} Q_t(y,b) - Q^*(x,a)$$

که $X(x,a)$ یک متغیر تصادفی مانند y از توزیع $P_a(x,y)$ است.

$$E[F_t(x,a)|F_t] = \sum_{y\in X} P_a(x,y)[r(x,a,y) + \gamma \max_{b\in A} Q_t(y,b)] - Q^*(x,a)$$

$$= (HQ_t)(x,a) - Q^*(x,a)$$

با این روش که $Q^* = HQ^*$

$$E[F_t(x,a)|F_t] = (HQ_t)(x,a) - HQ^*(x,a)$$

Date: Sub: ۹۹۴۱۱۴۰۵ ... فرزی

با توجه به این نکته که $(Hq)(x,a) = \sum_{y\in X} P_a(x,y)\left[r(x,a,y) + \gamma \max_{b\in A} q(y,b)\right]$

و $\|Hq_1 - Hq_2\|_\infty \le \gamma \|q_1 - q_2\|_\infty$

لذا:

$$\|E[F_t(x,a) \mid \mathcal{F}_t]\|_\infty \le \gamma \|Q_t - Q^*\|_\infty = \gamma \|\Delta_t\|_\infty$$

همچنین داریم:

$$\mathrm{var}[F_t(x) \mid \mathcal{F}_t] = E\left[\left(r(x,a,X(x,a)) + \gamma \max_{b\in A} Q_t(y,b) - Q^*(x,a) - (HQ_t)(x,a) + Q^*(x,a)\right)^2\right]$$

$$= E\left[\left(r(x,a,X(x,a)) + \gamma \max_{b\in A} Q_t(y,b) - (HQ_t)(x,a)\right)^2\right]$$

$$= \mathrm{var}\left[r(x,a,X(x,a)) + \gamma \max_{b\in A} Q_t(y,b) \mid \mathcal{F}_t\right]$$

که در اینجا داریم: (C = ثابت)

$$\mathrm{var}[F_t(x) \mid \mathcal{F}_t] \le C(1 + \|\Delta_t\|_w^2)$$

بنابراین Δ_t به صفر همگرا شده که Q_t به Q^* همگرا بود.
در نتیجه اثبات می‌گردد.

PITICO