RL 基础

集中不等式：

① Markov Inequality

$X \in R^+$，有 $Pr(X \geq s) \leq \frac{E(X)}{s}$

Proof: $E(X) = \int_0^s x p(x) dx + \int_s^\infty x p(x) dx$

$\geq \int_s^\infty x p(x) dx$

$\geq s \int_s^\infty p(x) dx = s Pr(X \geq s)$

$\Rightarrow Pr(X \geq s) \leq \frac{E(X)}{s}$

② Chebyshev's Inequality $Pr(|X - E(X)| \geq \epsilon) = Pr((X - E(X))^2 \geq \epsilon^2)$

$\leq \frac{E[(X - E(X))^2]}{\epsilon^2} = \frac{VAR(X)}{\epsilon^2}$

③ Chernoff Lemma → Chernoff Bound

mma: $Pr(X - E(X) \geq \epsilon) \leq \min_{\lambda \geq 0} \frac{E[\exp(\lambda(X - E[X]))]}{\exp(\lambda \epsilon)}$ and $Pr(X \geq \epsilon) \leq \min_{\lambda \geq 0} \frac{E[\exp(\lambda X)]}{\exp(\lambda \epsilon)}$

oof: $Pr(X - E(X) \geq \epsilon) \overset{\lambda \geq 0}{=} Pr(\exp(\lambda(X - E[X])) \geq \exp(\lambda \epsilon)) \leq \frac{E[\exp(\lambda(X - E[X]))]}{\exp(\lambda \epsilon)} \rightarrow$ ~~MGF~~ $= \frac{E[\exp(\lambda X)]}{\exp(\lambda \epsilon + \lambda E}$

Chernoff Inequality $X_i \sim Pr(\cdot)$ $Pr(\sum_{i=1}^n X_i - nE[X] \geq \epsilon) \leq \min_{\lambda \geq 0} [\prod_{i=1}^n (E[\exp^{\lambda(X_i - E[X_i])}]) \exp^{-\lambda \epsilon}$

Proof: $MGF_{X_1 + X_2 + \cdots X_n}(\lambda) = \prod_{i=1}^n MGF_{X_i}(\lambda)$

Chernoff Bound: Gets the optimal λ for a peticular distribution

④ Sub-Gaussian

If $X \sim subG(\sigma^2)$, then we have $E[\exp^{\lambda X}] \le \exp(\frac{\lambda^2\sigma^2}{2})$ the MGF of $N(0,\sigma^2)$

扩展: If $X \sim N(\mu, \sigma^2)$, then we have $E[\exp^{\lambda X}] = e^{\lambda\mu + \lambda^2\sigma^2/2}$

Proof: $M_X(t) = E[\exp^{\lambda X}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2} + \lambda x)\, dx$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{[x-(\mu+\lambda\sigma^2)]^2}{2\sigma^2} + \lambda\mu + \frac{\lambda^2\sigma^2}{2})\, dx$$

$$= e^{\lambda\mu + \frac{\lambda^2\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{[x-(\mu+\lambda\sigma^2)]^2}{2\sigma^2}) = 1$$

$$= e^{\lambda\mu + \frac{\lambda^2\sigma^2}{2}}$$

⑤ Hoeffding Lemma → Hoeffding Bound

Hoeffding lemma: (Loose Version) $E[\exp^{\lambda(X-E[X])}] \le \exp(\frac{\lambda^2(b-a)^2}{2})$, $\lambda \in R$, $X \in [a,b]$

(Strict Version / Tight) $E[\exp^{\lambda(X-E[X])}] \le \exp(\frac{\lambda^2(b-a)^2}{8})$, $\lambda \in R$, $X \in [a,b]$

Proof of the Loose Version:

$$E[\exp^{\lambda(X-E[X])}] \overset{X' \sim P'}{=} E_X[\exp^{\lambda(X-E_{X'}[X'])}] \overset{\text{Jensen Inequality}}{\le} E_X[E_{X'}[\exp^{\lambda(X-X')}]]$$

X 是关于0对称的, Rademacher $= \begin{cases} 1, & p=\frac{1}{2} \\ -1, & p=\frac{1}{2} \end{cases}$, $Pr(X-X') = Pr(\varepsilon(X-X'))$

$$= E_X[E_{X'}[E_{\varepsilon \sim Rad}[\exp^{\lambda\varepsilon(X-X')}]]]$$

$$E_{\varepsilon \sim Rad}[\exp^{\lambda\varepsilon}] = \sum_{k=0}^{\infty} \frac{\lambda^k E[\varepsilon^k]}{k!}$$

$$= \sum_{k=0,2,4}^{\infty} \frac{\lambda^k E[\varepsilon^k]}{k!} \quad (\because E[\varepsilon^k] = 0 \text{ if } k \text{ is odd})$$

$$= \sum_{k=0,1,2}^{\infty} \frac{\lambda^{2k}}{(2k)!}$$

$$\le \sum_{k=0,1,2,\cdots}^{\infty} \frac{\lambda^{2k}}{2^k \times k!} = \sum_{k=0,1,2\cdots}^{\infty} (\frac{\lambda}{2})^{2k} \frac{1}{k!}$$

$$= \exp(\frac{\lambda^2}{2})$$

$$\le E_X[E_{X'}[\exp(\frac{\lambda^2(X-X')^2}{2})]]$$

$$\le E_X[E_{X'}[\exp(\frac{\lambda^2(a-b)^2}{2})]] \quad (\because a \le X, X' \le b)$$

$$= \exp(\frac{\lambda^2(a-b)^2}{2})$$

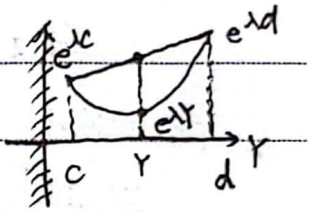
Proof of the Tight Version:

Set $Y = X - E[X]$, we have $E[Y] = 0$, $Y \in [\underbrace{a - E[X]}_{c}, \underbrace{b - E[X]}_{d}] = [c, d]$

What we need to prove is: $E[\exp^{\lambda Y}] \le \exp\left(\frac{\lambda^2 (c-d)^2}{8}\right)$, $\lambda \in R$, $Y \in [c,d]$, $E[Y] = 0$

① Proof One (Easy to understand):

$\because e^{\lambda Y}$ is convex of Y $\quad \therefore e^{\lambda Y} \le \boxed{\frac{b-Y}{b-a} e^{\lambda a} + \frac{Y-a}{b-a} e^{\lambda b}} \quad \frac{d-Y}{d-c} e^{\lambda c} + \frac{Y-c}{d-c} e^{\lambda d}$



$$E[\exp^{\lambda Y}] \le \frac{d - E[Y]}{d-c} e^{\lambda c} + \frac{E[Y] - c}{d-c} e^{\lambda d}$$

$$= \frac{d}{d-c} e^{\lambda c} + \frac{-c}{d-c} e^{\lambda d}$$

$$= \left(\frac{-c}{d-c}\right) e^{\lambda c} \left(-\frac{d}{c} + e^{\lambda (d-c)}\right)$$

$\because E[Y] = 0 \quad \therefore c \le 0, d \ge 0$ and c, d don't equal to 0 at the same time $\therefore c < 0, d > 0$

Set $\theta = -\frac{c}{d-c} > 0$,

$$E[\exp^{\lambda Y}] \le \theta e^{-\lambda\theta(d-c)} \left(\frac{1}{\theta} - 1 + e^{\lambda(d-c)}\right)$$

$$= \underbrace{(1 - \theta + \theta e^{\lambda(d-c)})}\, e^{-\lambda\theta(d-c)}$$

$\because 1 - \theta + \theta e^{\lambda(d-c)} = \theta\left(\frac{1}{\theta} - 1 + e^{\lambda(d-c)}\right) = \theta\left(-\frac{a}{b} + e^{\lambda(d-c)}\right) > 0$

$\therefore E[\exp^{\lambda Y}] \le e^{\ln(1 - \theta + \theta e^{\lambda(d-c)})} e^{-\lambda\theta(d-c)}$

Set $u = \lambda(d-c)$: $\quad E[e^{\lambda Y}] \le e^{\ln(1 - \theta + \theta e^{u}) - \theta u}$

Define $\varphi: R \to R$, $\varphi(u) = \ln(1 - \theta + \theta e^{u}) - \theta u$. $\quad E[e^{\lambda Y}] \le e^{\varphi(u)}$

According to Taylor's Theorem, $\exists \xi \in [0, u]$, $\varphi(u) = \varphi(0) + u\varphi'(0) + \frac{1}{2} u^2 \varphi''(\xi)$

$\varphi(0) = 0$, $\varphi'(0) = \left(\frac{\theta e^{u}}{1 - \theta + \theta e^{u}} - \theta\right)\Big|_{u=0} = 0$, $\varphi''(\xi) = \frac{\theta e^{\xi}}{1 - \theta + \theta e^{\xi}}\left(1 - \frac{\theta e^{\xi}}{1 - \theta + \theta e^{\xi}}\right) = t(1-t) \le$

$\therefore \varphi(u) \le 0 + 0 + \frac{1}{2} u^2 \times \frac{1}{4} = \frac{1}{8}\lambda^2 (d-c)^2$

Then we gets $\quad E[\exp^{\lambda Y}] \le \exp\left(\frac{\lambda^2 (d-c)^2}{8}\right)$

② Proof Two (Hard to understand): $M_Y(\lambda) = E[e^{\lambda Y}]$

Set $\Psi_Y(\lambda) = \ln M_Y(\lambda) = \ln E[e^{\lambda Y}]$.

$\Psi_Y(0) = \ln E[e^{0\cdot Y}] = 0$, $\Psi_Y'(0) = \frac{M_Y'(0)}{M_Y(0)} = \frac{E[Ye^{0\cdot Y}]}{E[e^{0\cdot Y}]} = E[Y] = 0$

For any $\xi \in [0, \lambda]$,

$$\Psi_Y''(\xi) = \left(\frac{M_Y'(\xi)}{M_Y(\xi)}\right)' = \frac{M_Y''(\xi)}{M_Y(\xi)} - \frac{M_Y'(\xi)^2}{M_Y(\xi)^2} = E\left[\frac{Y^2 e^{\xi Y}}{M_Y(\xi)}\right] - E\left[\frac{Ye^{\xi Y}}{M_Y(\xi)}\right]^2$$

Define a Z with CDF:

$$F_Z(z) = \int_{-\infty}^{z} \frac{e^{\xi y}}{M_Y(\xi)}\, dF_Y(y) = \int_c^z \frac{e^{\xi y}}{E[e^{\xi Y}]}\, dF_Y(y)$$

F_Z satisfies $F_Z(d) = \int_c^d \frac{e^{\xi y}}{E[e^{\xi Y}]}\, dF_Y(y) = 1$, so F_Z is well-defined, and $c \le Z \le d$

$$E\left[\frac{Ye^{\xi Y}}{M_Y(\xi)}\right] = \int_c^d y \cdot \frac{e^{\xi y}}{M_Y(\xi)}\, dF_Y(y) = E[Z]$$

$$E\left[\frac{Y^2 e^{\xi Y}}{M_Y(\xi)}\right] = \int_c^d y^2 \cdot \frac{e^{\xi y}}{M_Y(\xi)}\, dF_Y(y) = E[Z^2]$$

$\Psi_Y''(\xi) = E[Z^2] - E[Z]^2 = Var[Z]$

$\because c \le Z \le d \quad \therefore \quad |Z - \frac{c+d}{2}| \le \frac{d-c}{2}$

$\therefore Var[Z] = Var[Z - \frac{c+d}{2}] \le E[(Z - \frac{c+d}{2})^2] \le \frac{(d-c)^2}{4}$

$\therefore \exists \xi \in [0, \lambda], \quad \Psi_Y(\lambda) = \frac{\lambda^2}{2} Var[Z] \le \frac{\lambda^2 (d-c)^2}{8}$

Hoeffding Inequality:

For $X_1, X_2, \cdots X_n$, if X_i can be bounded, $P(X_i \in [a_i, b_i]) = 1$, Then for (Independent, r.v.) $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$

it satisfies

$$P(\bar{X} - E[\bar{X}] \ge t) \le \exp\left(-\frac{2t^2 n^2}{\sum_{i=1}^{n} (b_i - a_i)^2}\right)$$

$$P(|\bar{X} - E[\bar{X}]| \ge t) \le 2\exp\left(-\frac{2t^2 n^2}{\sum_{i=1}^{n} (b_i - a_i)^2}\right) \quad (t \ge 0)$$

Proof: Using Chernoff Bound we have

$$P(\bar{X}-E[\bar{X}]\geq t) \leq \min_{\lambda\geq 0} \frac{E[\exp(\lambda(\bar{X}-E[\bar{X}]))]}{\exp(\lambda t)}$$

According to Hoeffding Lemma, for $\bar{X}\in[\frac{1}{n}\sum_{i=1}^{n}a_i, \frac{1}{n}\sum_{i=1}^{n}b_i]$, we have

$$E[\exp(\lambda(\bar{X}-E[\bar{X}]))] \leq \exp\left(\frac{\lambda^2\frac{1}{n^2}(\sum_{i=1}^{n}b_i-\sum_{i=1}^{n}a_i)^2}{8}\right)$$

So we get

$$P(\bar{X}-E[\bar{X}]\geq t) \leq \min_{\lambda\geq 0} \exp\left(\underbrace{\frac{\lambda^2}{8n^2}(\sum_{i=1}^{n}b_i-\sum_{i=1}^{n}a_i)^2-\lambda t}_{g(\lambda)}\right) \overset{g'(\lambda)=0}{=\!=} \exp\left(-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i-\right.$$

Inversly, $P(E[\bar{X}]-\bar{X}\geq t) \leq \exp\left(-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$

Finally, we get

$$P(|\bar{X}-E[\bar{X}]|\geq t) \leq 2\exp\left(-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

Notice: Although Hoeffding Lemma is to prove the Sub-Gaussian property bounded random variables, we don't need the assumption of bounded Hoeffding Inequality. What we need is just the Sub-Gaussian property, a

$$E[\exp^{\lambda X}] \leq f(x) \qquad X\sim \text{subG}(\frac{(b-a)^2}{4})$$

5) Bernstein's Inequality (A Sharper bound using the variance of X_i)

Let $\{X_i\}$ be independent random variables, where with probability 1. $|X_i-E[X$

Let $\text{Var}(X_i)\leq \sigma_i^2$. Then, for all $t\geq 0$,

$$P\left(\sum_{i=1}^{n}(X_i-E[X])\geq t\right) \leq \exp\left(\frac{-t^2}{2\sum\sigma_i^2+\frac{2}{3}Rt}\right)$$

nce. this inequality is less used. we omits the proof.

⑦ Azuma - Hoeffding Inequality (Give a bound of martingale difference)

Azuma - Hoeffding aims at solving the dependent variables

Martingale Difference:

In RL, the independent r.v. is very hard to achieve, we usually get the dependent r.v.. For dependent r.v., we usually have the martingale difference

Example: Consider two random variables $X_1 \in \{0,1\}$, $X_2 \in \{-1,0,1\}$ with the following distribution:

$X_1=0$	value	probability
	-1	0.50
X_2	0	0
	1	0.50

$X_1=1$	value	probability
	-1	0
X_2	0	1
	1	0

We have $E[X_2|X_1=0] = E[X_2|X_1=1] = E[X_2]$, then we call $\{X_i - E[X_i]\}_{i=1}^{2}$ is a Martingale difference. Simply explaining, the variables and randomness of the past may change the distribution of the variable I am going to take but can't change the expectation.

Azuma - Hoeffding's Inequality: Suppose $F_1, F_2, \cdots, F_n$ are filtrations s.t. $F_{i-1} \subset F_i$, and $X_1, X_2, \cdots, X_n$ are random variables s.t. $X_i \in F_i$. If the following conditions hold,

- $E[X_i - EX_i | F_{i-1}] = 0$ (Martingale Difference)
- $|X_i - E[X_i]| \leq R_i$

then we have

$$P\left(\sum_{i=1}^{n}(X_i - EX_i) \geq t\right) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^{n} R_i^2}\right), \quad \forall t \geq 0$$

Proof: What we need is focusing on $E[\exp(\lambda \sum_{i=1}^{n}(X_i - EX_i))]$ (Need)

Proof: Sets $S_n = \sum_{i=1}^{n} X_i$, $Z_n = S_n - ES_n$, $W_n = e^{\lambda Z_n}$, $\mathcal{F}_1$

(Using the proof② of hoeffding lemma)

$$E[W_n | \mathcal{F}_{n-1}] = W_{n-1} E[e^{\lambda(Z_n - Z_{n-1})} | \mathcal{F}_{n-1}]$$

$$\leq W_{n-1} \frac{b_n e^{\lambda a_n} - a_n e^{\lambda b_n}}{b_n - a_n}$$

$E[W_n]$ ↑

$$\therefore E[E[W_n | \mathcal{F}_{n-1}]] \leq E[W_{n-1}] \cdot \frac{b_n e^{\lambda a_n} - a_n e^{\lambda b_n}}{b_n - a_n}$$

$$\leq \prod_{i=1}^{n} \frac{b_i e^{\lambda a_i} - a_i e^{\lambda b_i}}{b_i - a_i}$$

$$\forall t > 0, \quad P(S_n - ES_n \geq t) = P(W_n \geq e^{\lambda t}) \leq \frac{E[W_n]}{e^{\lambda t}}, \forall t > 0$$

$$\leq e^{-\lambda t} \prod_{i=1}^{n} \frac{b_i e^{\lambda a_i} - a_i e^{\lambda b_i}}{b_i - a_i}$$

$$\leq e^{-\lambda t} \prod_{i=1}^{n} e^{\lambda^2 (b_i - a_i)^2 / 8}$$

$$\leq \exp\left(\frac{-t^2}{2 \sum_{i=1}^{n} R_i^2} \right)$$

⑧ Uniform Concentration

1. (Union Bound) $P(A \cup B) \leq P(A) + P(B)$

2. Suppose $\{X_i\}_{i=1}^{n}$ are random variables chosen i.i.d. and $\text{supp}(X_i) = \{s | P(X_i = s) > 0, \forall s \in S\}$ where S is a discrete set with $|S| = S$.

Consider two cases shown below.

Case 1: f fixed

$f: S \to [0,1]$ where $f(X_i)$ are random variables chosen i.i.d. By Hoeffding Ineq

$$P\left(\left|\sum_{i=1}^{n} f(x_i) - Ef(x_i)\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{n}\right)$$

Case 2: f not fixed

For example, $\hat{f} = \operatorname{argmin}_{f \in F} L(f, \{x_i\}_{i=1}^n)$, $\hat{f}$ depends on the whole dataset

$$P\left(\left|\sum_{i=1}^{n} (\hat{f}(x_i) - E\hat{f}(x_i)\right| \geq t\right) \leq P\left(\exists f \in F, \left|\sum_{i=1}^{n} (f(x_i) - Ef(x_i))\right| \geq t\right)$$

$$= P\left(\bigcup_{f \in F} \left|\sum_{i=1}^{n} (f(x_i) - Ef(x_i))\right| \geq t\right)$$

Case 2a: Finite F

If F is finite, we can use lemma 1 to get the following:

$$P\left(\bigcup_{f \in F} \left|\sum_{i=1}^{n} (f(x_i) - Ef(x_i))\right| \geq t\right) \leq \sum_{f \in F} P\left(\left|\sum_{i=1}^{n} (f(x_i) - Ef(x_i))\right| \geq t\right)$$

$$\leq 2|F|\exp\left(\frac{-2t^2}{n}\right) \quad (*)$$

$(*)$ is only true if the following holds with probability $1-\delta$.

$$\sup_{f \in F} \left|\frac{1}{n}\sum_{i=1}^{n} (f(x_i) - Ef(x_i))\right| \leq \sqrt{\frac{1}{2n}\log\frac{2|F|}{\delta}}$$

$$= O\left(\sqrt{\frac{1}{n}\log\frac{|F|}{\delta}}\right)$$

Case 2b: Infinite F

We can discretize it using ϵ-grid G_ϵ s.t. $G_\epsilon = \{0, \epsilon, 2\epsilon, 3\epsilon, \ldots, \lfloor\frac{1}{\epsilon}\rfloor\epsilon\}$

$F = G_\epsilon^S = \{f \mid f: S \to G_\epsilon\}$. We have $\forall f \in F, \exists f_\epsilon \in F_\epsilon$ s.t. $\sup_{x \in S} |f(x) - f_\epsilon(x)| \leq \epsilon$

w.p. $1-\delta$, $\forall f \in F$, $\left|\frac{1}{n}\sum_{i=1}^{n} (f(x_i) - Ef(x_i))\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n} f_\epsilon(x_i) - Ef_\epsilon(x_i) + (f - f_\epsilon)(x_i) - E[(f - f_\epsilon)(x_i)]\right|$

⑨ Bellman Equation

Recall MDP (S, A, r, P, H) where:

environment : P, r

policy : (history independent) $\pi : S \to A / \Delta A$ $\quad \pi_h : H \to A / \Delta A$

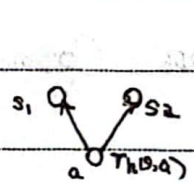
H : $(s_1, a_1, r_1, s_2, a_2, r_2, \cdots)$

State-Value Function : $V_h^{\pi}(s) = E_{\pi}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right]$

Action-Value Function : $Q_h^{\pi}(s,a) = E_{\pi}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right]$

Bellman Equations :

$$\begin{cases} V_h^{\pi}(s) = \sum_{a \in A} Q_h^{\pi}(s,a)\, \pi_h(a|s) \\ Q_h^{\pi}(s,a) = r_h(s,a) + E_{s' \sim P_h(\cdot|s,a)} V_{h+1}^{\pi}(s') \end{cases}$$



Optimal Policy

Optimality Equation :

$$\begin{cases} V_h^*(s) := \max_{\pi} V_h^{\pi}(s) & \forall (s,h) \\ Q_h^*(s,a) := \max_{\pi} Q_h^{\pi}(s,a) & \forall (h,s,a) \end{cases}$$

Optimal policy π^* :

$$\begin{cases} V_h^*(s) := V_h^{\pi^*}(s) & \forall (s,h) \\ Q_h^*(s,a) := Q_h^{\pi^*}(s,a) & \forall (s,a,h) \end{cases}$$

Theorem 1 (Optimality Existence): There exists a history independent and deterministic policy π^* that is optimal.

Proof: Let $\pi_h^*(s) = \operatorname{argmax}_a Q^*(s,a)$ be history independent and deterministic. We can now prove by induction:

1. Base case: H: $Q_H^*(s,a) = r_H(s,a) = Q_H^{\pi^*}(s,a)$
2. $Q_h^* = Q_h^{\pi^*} \Rightarrow V_h^* = V_h^{\pi^*}$;

$$V_h^*(s,a) = \max_{\pi_h} \max_{\pi_{h+1}\cdots\pi_H} \sum_{a\in A} Q_h^{\pi}(s,a)\cdot \pi_h(a|s)$$

$$= \max_{\pi_h} \sum_{a\in A} Q_h^*(s,a)\pi_h(a|s)$$

$$= \max_a Q_h^*(s,a) = Q_h^*(s,\pi_h^*(s)) = V_h^{\pi^*}(s)$$

③ $V_{h+1}^* = V_{h+1}^{\pi^*} \Rightarrow Q_h^* = Q_h^{\pi^*}$

$$Q_h^*(s,a) = r_h(s,a) + \max_{\pi} (PV_{h+1}^{\pi})(s,a) = r_h(s,a) + (PV_{h+1}^*)(s,a) = r_h(s,a) + (PV_{h+1}^{\pi^*})(s,a) = Q$$

Bellman Optimality Equation: $\begin{cases} V_h^*(s) = \max_a Q_h^*(s,a) \\ Q_h^*(s,a) = r_h(s,a) + (PV_{h+1}^*)(s,a) \end{cases}$

Goal of RL: Find the $\pi_h^*(s) = \operatorname{argmax}_a Q_h^*(s,a)$

Bellman Equation: $\begin{cases} V^{\pi}(s) = \sum_{a\in A} Q^{\pi}(s,a)\pi(a|s) \\ Q^{\pi}(s,a) = r(s,a) + \gamma(PV^{\pi})(s,a) \end{cases}$

Discounting Factor γ MDP(S, A, r, P, γ)

② Planning in MDP

Algorithm 1 Policy Evaluation (DP)

for $h = H, H-1, \cdots, 1$ do

$Q_h^{\pi}(s,a) = r_h^{\pi}(s,a) + (P_h V_{h+1}^{\pi}(s,a))$

$V_h^{\pi}(s) = \sum_{a\in A} Q_h^{\pi}(s,a)\pi(a|s)$

until $\max_s |V_{h+1}^{\pi}(s) - V_h^{\pi}(s)| < \theta$ ×)

gorithm 2 Policy Iteration

for $t = 1, 2, \cdots, H$ do

Run Policy Evaluation to evaluate $Q_h^{\pi^{t+1}}(s,a)$ $\forall s,a,h$

$\pi_h^t(s) \leftarrow \operatorname{argmax}_{a\in A} Q_h^{\pi^{t+1}}(s,a)$

(Infinite Horizon)

Algorithm 1 Policy Evaluation (Fixed Point I

$Q_0 \in [0, \beta]^{SA}$ $\beta (\gamma + \gamma^2 + \cdots \gamma^{\infty} < \frac{1}{1-\gamma} = \beta)$

for $t = 0, 1, \cdots$ do

$Q_{t+1} = T^{\pi} Q_t = r + \gamma P^{\pi} Q^{\pi}$

Lemma: $\|Q_t - Q^{\pi}\|_{\infty} \le \gamma^t \|Q_0 - Q^{\pi}\|_{\infty} \le \gamma^t \beta$

Algorithm 2 Value It

$Q_0 \in [0, \beta]^{SA}$

for $t = 0, 1, \cdots$ do

$Q_{t+1} = T^* Q_t$

(Finite Horizon)

Algorithm 3 Value Iteration

for $h = H, H-1, \dots, 1$ do

$Q_h^*(s,a) = r_h(s,a) + (P_h V_{h+1}^*(s,a))$

$V_h^*(s) = \max_{a \in A} Q_h^*(s,a)$

$\pi_h^*(s) = \operatorname{argmax}_{a \in A} Q_h^*(s,a)$

(☆) Theorem: π^{t_h} is optimal $\forall t \geq H$

Proof: Using induction

Base case: $t=0$ $Q_H^{\pi^0}(s,a) = r_H(s,a) = Q$

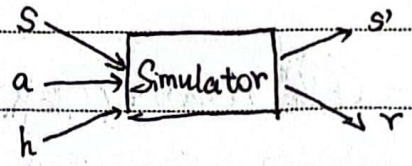
If t is true, then at $t+1$, $\forall h \geq H-t$,

$\pi_h^{t+1}(s) = \operatorname{argmax}_a Q_h^{\pi^t}(s,a) = \operatorname{argmax}_a Q_h^*($

$= \pi_h^*(s)$

(II) Generative Models

MDP(S, A, r, P, H) with P, r unknown, but we have access to a simu



Goal is to find some policy $\hat{\pi}$ so that $V_1^*(s_1) - V_1^{\hat{\pi}}(s_1) \leq \epsilon$ for all i, and the sample complexity is how many queries needed to ask the simulator to find ϵ-optimal $\hat{\pi}$.

Generative Value Iteration:

Input: n

For all $(s,a,h) \in S \times A \times H$ do

Query (s,a,h) and collect n samples $\{s_1', s_2', \dots, s_n'\}$

Estimate transition probability by

$$\hat{P}_h(s'|s,a) = \frac{1}{n} \sum \mathbb{1}[s_i' = s']$$

for all $s' \in S$.

For $h = H$ to 1 do

Compute $\hat{Q}_h^*(s,a) = r_h(s,a) + (\hat{P}_h V_{h+1}^*)(s,a)$

Compute $V_h^*(s) = \max_{a \in A} \hat{Q}^*(s,a)$

Output: $\hat{\pi}_h(s) = \operatorname{argmax}_a \hat{Q}_h(s,a)$

(Coarse Analysis)

Analysis 1:

Theorem: If we chose $n \geq C\frac{H^4 S}{\epsilon^2} l$, where $l = \log\frac{HSA}{p}$, then with probabili the policy resulting from value iteration will be ϵ-optimal.

Lemma 1. For any policy π and any $(s,h) \in S \times [H]$,

$$\underbrace{V_h^{\pi}(s)}_{\text{true}} - \underbrace{\hat{V}_h^{\pi}(s)}_{\text{estimated}} = E_{M,\pi}\left[\sum_{i=h}^{H} ((P_i - \hat{P}_i) V_{i+1}^{\pi})(s_i, a_i) \mid s_h = s\right]$$

Proof. Note that

$$Q_h^{\pi}(s,a) - \hat{Q}_h^{\pi}(s,a) = (PV_{h+1}^{\pi})(s,a) - (\hat{P}\hat{V}_{h+1}^{\pi})(s,a) = [(P_h - \hat{P}_h)V_{h+1}^{\pi}) + (\hat{P}_h(V_{h+1}^{\pi} -$$

$$\Rightarrow V_h^{\pi}(s) - \hat{V}_h^{\pi}(s) = E_{\pi}[(P_h - \hat{P}_h) V_{h+1}^{\pi}(s_h, a_h) \mid s_h = s] + E_{M,\pi}[V_{h+1}^{\pi}(s_{h+1}) - \hat{V}_{h+1}^{\pi}(s_{h+1}) \mid s_h =$$

$$= E_{M,\pi}\left[\sum_{i=h}^{H} (P_i - \hat{P}_i) V_{i+1}^{\pi})(s_i, a_i) \mid s_h = s\right]$$

mma 2. With probability $\geq 1-p$, we have $\forall (s,a,h) \in S \times A \times [H]$, $\forall V \in [0,H]^S$,

$$|(P_h - \hat{P}_h) V(s,a)| \leq C \cdot H \cdot \frac{Sl}{n}, \text{ where } l = \log\frac{HSA}{p}$$

sing Lemma 1, 2. Theorem can be get.

Refined Analysis)

lysis 2.

Theorem: If we chose $n \geq c \cdot \frac{H^4 l}{\epsilon^2}$ where $l = \log\frac{HSA}{p}$, then ...

⑫ Q-learning

$$Q_h^t(s,a) = (1-\alpha_t) Q_h^{t-1}(s,a) + \alpha_t (r_h(s,a) + V_{h+1}^{t-1}(s_i'))$$

α_t: learning rate

Algorithm Q-learning

Input: learning rate $\{\alpha_t\}_{t=1}^{\infty}$

Initialize: $Q_h^0(s,a) = 0,\ \forall (s,a,h) \in S \times A \times [H]$

for $t = 1, \dots$ do

  for $h = 1, \dots, H$ do

    for $(s,a) \in S \times A$ do

      sample next state s_t' from $P_h(\cdot|s,a)$

      Update

$$Q_h^t(s,a) = (1-\alpha_t) Q_h^{t-1}(s,a) + \alpha_t (r_h(s,a) + V_{h+1}^{t-1}(s_i'))$$

  for $s \in S$ do

    $V_h^t(s) = \max_{a \in A} Q_h^t(s,a)$

Remark:

① 1. Model-free learning

2. Incremental update using one noisy sample. Space complexity: $O(HSA)$

ng the update rule we get:

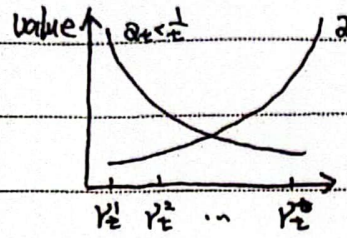
$$Q_h^t(s,a) = \sum_{i=1}^{t} \gamma_t^i (r_h(s,a) + V_{h+1}^{i-1}(s_i')) = r_h(s,a) + \sum_{i=1}^{t} \gamma_t^i V_{h+1}^{i-1}(s_i')$$

$$\gamma_t^i := \alpha_i \prod_{j=i+1}^{t} (1-\alpha_j) \quad \text{and} \quad \sum_{i=1}^{t} \gamma_t^i = 1$$

(Generative)

Comparing to Value Iteration: $Q_n(s,a) = r_n(s,a) + \sum_{i=1}^{n} \frac{1}{n} \hat{V}_{n-1}(s_i')$

1. When $\alpha_t = \frac{1}{t}$ and $\gamma_t^i = \frac{1}{t}$, Equal
2. When $\alpha_t > \frac{1}{t}$, Q-learning favors later samples.
3. When $\alpha_t < \frac{1}{t}$, Q-learning favors earlier samples.



Usually, we choose $\alpha_t > \frac{1}{t}$.

⑭ Multi-arm bandit (MAB)

Exploration vs. Expoloitation

$$\text{regret}(T) = T \cdot r_{i^*} - \sum_{t=1}^{T} r_{i_t}$$

$$0 \le \text{regret}(T) \le T$$

Conversion:

lemma 1. If algorithm A has regret $c \cdot T^{1-\alpha}$, $\alpha \in (0,1]$. Then A finds an ϵ-opt
distribution of arm in $(\frac{c}{\epsilon})^{\frac{1}{\alpha}}$ samples. Conversely, $c \cdot \epsilon^{-\beta}$ samples $\Rightarrow$ $2 \cdot c^{\frac{1}{1+\beta}} \cdot T^{\frac{\beta}{1+\beta}}$

$\sqrt{T}$ Regret $\rightarrow \frac{1}{\epsilon^2}$ samples

$\frac{1}{\epsilon^2}$ samples $\rightarrow T^{\frac{1}{2}}$

Proof: a. $\text{regret}(T) = T \cdot r_{i^*} - \sum_{t=1}^{T} r_{i_t} \le c \cdot T^{1-\alpha}$

$$\frac{1}{T}\sum_{t=1}^{T} r_{i_t} \ge r_{i^*} - \frac{c}{T^{\alpha}}$$

$$\frac{c}{T^{\alpha}} = \epsilon \rightarrow T = (\frac{c}{\epsilon})^{\frac{1}{\alpha}}$$

b. $\text{regret}(T) \le c \cdot \epsilon^{-\beta} \times 1 + (T - c \cdot \epsilon^{-\beta}) \times \epsilon$

$$\le c \cdot \epsilon^{-\beta} + T\epsilon$$

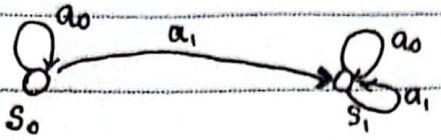
$$= 2\, c^{\frac{1}{1+\beta}} T^{\frac{\beta}{1+\beta}}$$

⑮ Exploration in MDPs

A contextual MAB can be viewed as an MDP without transit

exploration $\begin{cases} \text{MABs: } \tilde{\Theta}(|A|\epsilon^{-2}) \\ \text{MDP: } \Omega(2^H) \text{ (worst)} \end{cases}$

Combinatorial lock:



$r_H(s_0)=1$, $r_h(s)=0$

$\{\underbrace{a_0, a_0, \cdots, a_0}_{H}\}$ is the optimal, however, we need $\Omega(2^H)$ to get the p

MPP is more hard for exploration than MAB.

) UCB-VI (Upper Confidence Bound Value Iteration)

lgorithm UCB-VI

Initialize: dataset $D := \phi$; $Q_h(s,a) := H$ for $h \in [H]$; $V_{H+1}(s)=0$.

for episode $k=1$ to K do

(Part 1: Value iteration with bonus)

for $h = H$ to 1 do

for $(s,a) \in S \times A$ do

if $(s,a) \in D$ then

Compute $\hat{P}_h(s'|s,a) := \frac{N_h(s,a,s')}{N_h(s,a)}$ for all $s' \in S$, where $N_h(s,a,s')$ and N

$Q_h(s,a) := \min\{H, r_h(s,a) + (\hat{P}_h V_{h+1}(s,a) + b(N_h(s,a))\}$;

for $s \in S$ do

$V_h(s) := \max_{a \in A} Q_h(s,a)$;

for $h=1$ to H do

Take action $a_h := \operatorname{argmax}_{a \in A} Q_h(s_h, a)$