Wednesday, October 28, 2020

Lecture 20

CS59000_RL

MDP

- Infinite horizon MDP – Undiscounted
- Optimism

we defined

$$\rho^\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_\pi^{t-1} r_\pi$$

$$\to \text{let } \bar{P}_\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_\pi^{t-1}$$

$$\Rightarrow \rho^\pi = \bar{P}_\pi r_\pi$$

$$V_\pi^T = \sum_{t=1}^{T} P_\pi^{t-1} \left(r_\pi - \overset{\bar{\rho}^\pi \mathbb{1}}{\rho^\pi} \right)$$

$$\Rightarrow V_\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{T>0} V_\pi^T$$

Hardy 1945

Lemma: $V_\pi = \left(\left(I - P_\pi - \bar{P}_\pi \right)^{-1} - \bar{P}_\pi \right) r_\pi$

A value function $V: \mathcal{X} \to \mathbb{R}$

$$\Rightarrow \operatorname{span}(V) = \max_{x\in\mathcal{X}} V(x) - \min_{x\in\mathcal{X}} V(x)$$

Lemma: For any memoryless policy π

$$\bar{P}_\pi V_\pi = 0$$

why ?

Proof: note that $\boxed{\bar{P}_\pi P_\pi = \bar{P}_\pi}$

$$\Rightarrow \bar{P}_\pi V_\pi^T = \sum_{t=1}^{T} \underbrace{\bar{P}_\pi P_\pi^{t-1}}_{\bar{P}_\pi} (r_\pi - \rho^\pi) = T\left(\bar{P}_\pi (r_\pi - \rho^\pi)\right)$$

$$= T(\underbrace{\bar{P}_\pi r_\pi}_{\rho^\pi} - \rho^\pi)$$

$\boxed{\rho^\pi 1}$

$$= T(0) = 0$$

$$\Rightarrow V_\pi^T = 0 \Rightarrow V_\pi = 0$$

Lemma: For any memoryless policy, we have

$$\rho^\pi + V_\pi = r_\pi + P_\pi V_\pi$$

Historically this equation is known as Poisson equation
Nowadays we call it Bellman equation.

---

proof: $V_\pi = (I - P_\pi + \bar{P}_\pi)^{-1} r_\pi - \underbrace{\bar{P}_\pi r_\pi}_{\rho^\pi}$

$$V_\pi + \rho^\pi = \underbrace{(I - P_\pi + \bar{P}_\pi)^{-1}}\, r_\pi$$

Wednesday, October 28, 2020

$$\underbrace{(I - P_\pi + \bar{P}_\pi)}(\underbrace{V_\pi} + \rho^\pi) = r_\pi$$

$$V_\pi - P_\pi V_\pi + \underbrace{\bar{P}_\pi V_\pi}_{0} + \rho^\pi - \underbrace{P_\pi \rho^\pi}_{-\rho^\pi} + \underbrace{\bar{P}_\pi \rho^\pi}_{\rho^\pi} = r_\pi$$

$$\Rightarrow V_\pi - P_\pi V_\pi + \rho^\pi = r_\pi$$

Bellman Optimality equation

inner product.

$$\rightarrow \quad \rho + V(x) = \max_{a \in A} \left(\bar{r}(a,x) + \langle P(\cdot|x,a), V \rangle \right)$$

Bellman operator: $T(V) = \max_a \left(\bar{r}(a,x) + \langle P(\cdot|x,a), V \rangle \right)$

For V, define greedy policy $\underbrace{\pi_V}(x) \in \arg\max_{a \in A} \bar{r}(a,x) + \langle P(\cdot|x,a), V \rangle$

Theorem:

(i) Bellman Optimality equation has a solution

(ii) A solution (ρ, V) satisfies $\rho^* = \rho$, $\pi^* = \pi_V$

Why not unique? if V^π is solution $\Rightarrow V^\pi + \alpha 1$ is also a solution.

Now, how to solve it?

- Policy iteration
- Value iteration.
- Linear program.

Wednesday, October 28, 2020

$$\underset{\rho \in \mathbb{R}, V \in \mathbb{R}^{|\mathcal{X}|}}{\text{minimize}} \ \rho$$

$$\text{s.t.: } \rho + V(x) \geq \bar{r}(x,a) + \langle P(\cdot|x,a), V \rangle \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A}$$

$\Rightarrow$ The solution is ρ^*. But, the V and Π_V might be useless.

having ρ^*, now go for V^*:

$$\underset{V \in \mathbb{R}^{|\mathcal{X}|}}{\text{minimize}} \ \langle V, \mathbb{1} \rangle$$

$$\text{s.t. } \begin{cases} \rho^* + V(x) \geq \bar{r}(x,a) + \langle P(\cdot|x,a), V \rangle & \forall (x,a) \in \mathcal{X} \times \mathcal{A} \\ V(x_0) = 0 & x_0 \Rightarrow \text{a state that } \Pi^* \text{ visit.} \end{cases}$$

Now consider, we neither have P, nor R.

We need to intract with the environment, explore it, learn it, and exploit what we have.

$\hookrightarrow$ Tradeoff exploration and exploitation

Define regret as : $\hat{R}_T = T\rho^* - \sum_{t=1}^{T} r(X_t, A_t)$

Find an algorithm with reasonable upper bound on regret.

Wednesday, October 28, 2020

We use optimism

Alg: Upper Confidence bound for reinforcement learning 2 (UCRL2)

How it works?

$\|P - \hat{P}\|$

$\|\bar{r} - \hat{r}\|$

$\rightarrow$ Most optimistic model $\rightarrow \tilde{\pi}$

True model M: $P, \bar{r}$

Estimated model $\hat{M}$: $\hat{P}, \hat{r}$

For now: $0 \leq r \leq 1$, we assume $\bar{r} = r$ and we know $\bar{r}$

At time t, define $T_t(x,a) = \sum_{i=1}^{t} \mathbb{I}\{X_i = x, A_i = a\}$

Then $$P_t(x'|x,a) = \frac{\sum_{i=1}^{t} \mathbb{I}\{X_i = x, A_i = a, X_{i+1} = x'\}}{1 \vee T_t(x,a)}$$

$\Rightarrow$ i.e. empirical estimate.

Define Confidence set for x,a

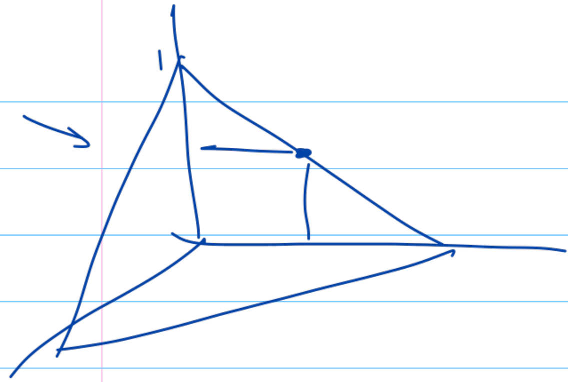
$$\rightarrow C_t^{\delta}(x,a) = \left\{ P' \in \Delta^{|X|} : \| P' - P_{t-1}(\cdot|x,a) \|_1 \leq \sqrt{\frac{|X| \, L_{t-1}(x,a)}{1 \vee T_{t-1}(x,a)}} \right\}$$

For $$L_{t-1}(x,a) = 2 \log\left(\frac{4|X||A| \, T_t(x,a)\,(1 + T_t(x,a))}{\delta} \right)$$

Wednesday, October 28, 2020

$P_{t-1} \in \mathbb{R}^{|X|}$ $\sum_{i=1}^{|X|} P_{t-1}(i) = 1 \Rightarrow P_{t-1} \in \Delta^{|X|}$

Δ^3



UCRL2:

For $k=1, \ldots$

- Set $\tau_k = t+1$
- Find an optistic model, return, value its optimal memory less, deterministic policy

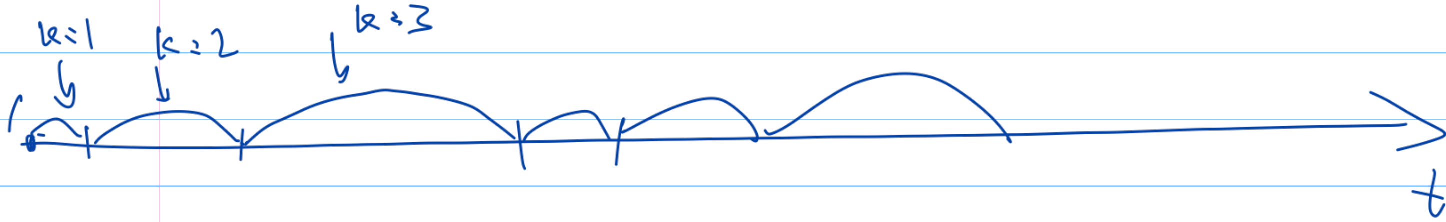
$$\rho_k = \max_{x \in X} \left(\max_{\pi} \max_{P' \in C_{\tau_k}} \rho_x^{\pi}(P') \right)$$

- optimistic model P_k, $\pi_k \rightarrow$ optimistic policy

- do

$t \leftarrow t+1$, observe X_t, take $A_t = \pi_k(X_t)$

while $T_t(X_t, A_t) < 2\, T_{\tau_k - 1}(X_t, A_t)$



Wednesday, October 28, 2020

Theorem: UCRL2 acheives regret of

$$\hat{R}_T \leq C D |X| \sqrt{|A| T \log\left(\frac{T |X| |A|}{\delta}\right)}$$

universal constant

with probability at least $1-\delta$

Lower bound on MDP

For $|X| \geq 3$, $|A| \geq 2$, $D \geq 6 + 2 \log_{|A|} |X|$,
and $T \geq D |X| |A|$,

then for any algorithm, there exists an MDP, such that

$$E[\hat{R}_T] \geq C' \sqrt{D |X| |A| T}$$

universal constant

UCRL2 is order optimal; But off $\sqrt{|X|} D$

up to log factors.