Monday, November 9, 2020

Lecture 22

CS 59000 _ RL

MDP

Agenda:

- Finish UCRL2 proof
- other models.

$$R_K \leq D + \sum_{t=\tau_k}^{\tau_{k+1}-1} \overbrace{E\left[V_k(X_{t+1}) \mid \mathcal{F}_t\right] - V_k(X_{t+1})}^{\text{mean zero}} \leftarrow A_1$$

$$+ \frac{D\sqrt{L|X|}}{2} \sum_{x,a \in X \times A} \frac{T_k(x,a)}{\sqrt{\max(1, T_{\tau_k - 1}(x,a))}} \leftarrow A_2$$

number of x,a samples at epoch k

$$\Rightarrow \hat{R}_T \leq KD + \sum_{t=1}^{T} E\left[V_{k_t}(X_{t+1}) \mid \mathcal{F}_t\right] - V_{k_t}(X_{t+1})$$

k_t: mean epoch number at time t

$$+ \frac{D\sqrt{L|X|}}{2} \sum_{x,a \in X \times A} \sum_{k=1}^{K} \frac{T_k(x,a)}{\sqrt{1 \vee T_{\tau_{k-1}}(x,a)}}$$

Using Azuma inequality

we have:
$$\mathbb{P}\left(F \text{ and } \sum_{t=1}^{n} E\left[V_{k_t}(X_{t+1})\right] - V_{k_t}(X_{t+1}) \leq \frac{D\sqrt{T \log \frac{2}{\delta}}}{2}\right) \leq 1 - \frac{\delta}{2}$$

Monday, November 9, 2020

Now: $\displaystyle\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \frac{T_k(x,a)}{\sqrt{1\vee T_{\tau_{k-1}}(x,a)}}$

Lemma 19 (UCR22)

For any sequence of numbers $z_i \in \mathbb{N}\cup\{0\}$

with $0 \leqslant z_k \leqslant Z_{k-1} = 1 \vee \sum_{i=1}^{k-1} z_i$

$$\sum_{k=1}^{K} \frac{z_k}{\sqrt{Z_{k-1}}} \leqslant (\sqrt{2}+1)\sqrt{Z_K}$$

Proof. HW.

Using this lemma:

$$\sum_{k=1}^{K} \frac{T_k(x,a)}{\sqrt{1\vee T_{\tau_{k-1}}(x,a)}} \leqslant (\sqrt{2}+1)\sqrt{T_T(x,a)}$$

Since $\displaystyle\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} T_T(x,a) = T$

$$\Rightarrow \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \sqrt{T_T(x,a)} \leqslant \sqrt{|\mathcal{X}||\mathcal{A}|T}$$

Monday, November 9, 2020

$$\Rightarrow \hat{R}_T \leq KD + \frac{D\sqrt{2|\mathcal{X}|}}{2}\sqrt{|\mathcal{X}||A|T} + \frac{D\sqrt{T\log\frac{2}{\delta}}}{2}$$

with probability at least $1-\frac{\delta}{2}$

what is K ?

How many times for (x,a), we can double the epoch?

$$\underbrace{1+\log_2 T_T(x,a)}$$

$$\Rightarrow K \leq \sum_{\underbrace{(x,a)\in\mathcal{X}\times A}} \left(1+\underbrace{\log_2\left(T_T(x,a)\right)}\right)$$

since $\sum_{(x,a)\in\mathcal{X}\times A} T_T(x,a) = \underline{T}$ $\Rightarrow$ the max for the right hand side:

$$K \leq |\mathcal{X}||A|\left(1+\log_2\left(\frac{T}{|\mathcal{X}||A|}\right)\right)$$

we just proved UCRL2 acheives regret of

$$\rightarrow D|\mathcal{X}||A|\left(1+\log_2\frac{T}{|\mathcal{X}||A|}\right)+\frac{D\sqrt{2|\mathcal{X}|}}{2}\sqrt{|\mathcal{X}||A|T}+\frac{D\sqrt{T\log\frac{2}{\delta}}}{2}$$

i.e. $\underline{\tilde{O}\left(D|\mathcal{X}|\sqrt{|A|T}\right)}$

Monday, November 9, 2020

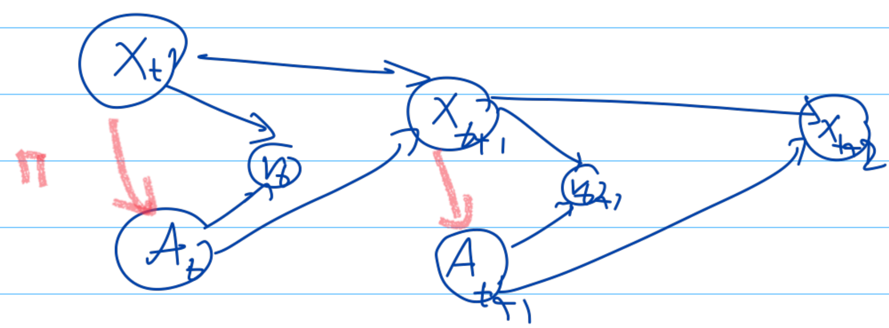
Note: the lowers bound is $\sqrt{D|X||A|T}$, $\Rightarrow$ we are off with $\sqrt{D|X|}$

Subsequent work show $\sqrt{D|X||A|T} + poly(|X|, |A|)$

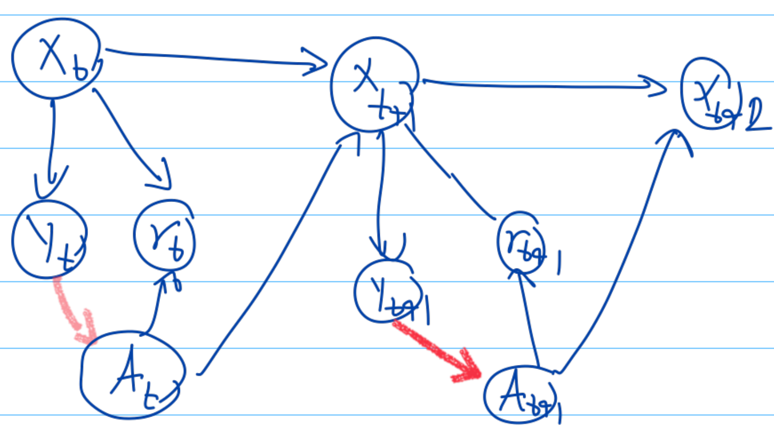
$\hookrightarrow \tilde{O}(\sqrt{D|X||A|T})$

Rich observable MDP (ROMDP)

— MDP



— ROMDP



There is a onto mapping from $Y \to X$

i.e. given Y we know the underlying X

Monday, November 9, 2020

in other words, the exist a function $f: \mathcal{Y} \to \mathcal{X}$

ROMDP: $M := (\mathcal{X}, \mathcal{Y}, \mathcal{A}, \underline{P}, P_1, \underline{R}, \underline{O})$

Infinite case $- X_1 \sim P_1$
$X_{t+1} \sim P(\cdot | X_t, A_t)$
$Y_t \sim O(\cdot | X_t)$
$r_t \sim R(X_t, A_t)$

O is a stochastic mapping from X to Y (surjective map)
f is its invers map. which onto.

ROMDP → results in MDP on $\mathcal{Y}$ space with $|\mathcal{Y}|$ number of state and $D_{\mathcal{Y}}$ as its diameter.

⇒ UCB style algorithm give regret of $\tilde{O}(|\mathcal{X}| D_{\mathcal{X}} \sqrt{|A| T})$
However UCRL2 gaurantees $\tilde{O}(|\mathcal{Y}| D_{\mathcal{Y}} \sqrt{|A| T})$

where $|\mathcal{Y}| \gg |\mathcal{X}|$ $D_{\mathcal{Y}} \gg D_{\mathcal{X}}$

Azizzadenesheli 2017

Monday, November 9, 2020

Partially Observable MDPs (POMDP)

simple case

$\rightarrow M := (\mathcal{X}, \mathcal{A}, \mathcal{Y}, P, P_1, O, R, \gamma)$

where

$\mathcal{X}$:= State space

$\mathcal{A}$:= Action space

$\mathcal{Y}$:= Observation space

P := Transition kernel

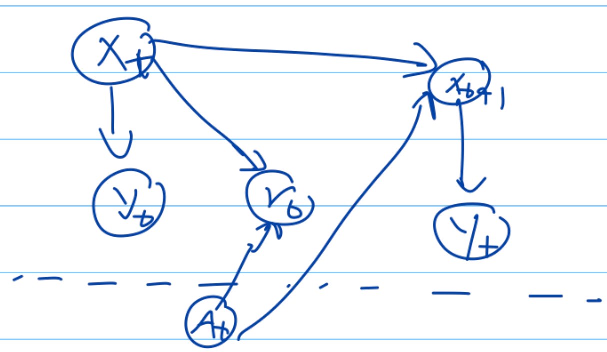
P_1 := Initial state measure

R := Reward Kernel

γ := Discount factor.

The process:

- $X_1 \sim P_1$



$Y_t \sim O(X_t)$

$r_t \sim R(X_t, A_t)$

$X_t \sim P(X_t, A_t)$

How A_t is generated?

Monday, November 9, 2020

Policy:

- History dependent policy $A_t \sim \Pi(h_t)$

history $(\ldots, A_{t}, Y_t)$

- Memory less policy $A_t \sim \Pi(Y_t)$

- Markovian policy $A_t \sim \Pi(Y_t, t)$

$\Rightarrow$ Note MDP $\subset$ ROMDP $\subset$ POMDP