Monday, November 2, 2020

Lecture 21

CS 59000 - RL

MDPs

Agenda

- UCRL2 (Pebe, josch, Ronald 2010)

Aim: $\hat{R}_T = \underline{T\rho} - \underbrace{\sum_{t=1}^{T} r(X_t, A_t)}$

what is oracles long term return — what our algorithm acheives

---

- $\bar{r}(x,a) = r(x,a)$ deterministic, known

- Diameter $D < \infty$.

- $0 \leq r(x,a) \leq 1$

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

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

Side note

- binomial: $p \to$ w.p. $p \to$ outcome $\begin{bmatrix}1\\0\end{bmatrix}$
  w.p. $1-p \to$ outcome $\begin{bmatrix}0\\1\end{bmatrix}$

- multinomial distribution $p \in \Delta^d \to$ outcome e_i w.p. p_i

Monday, November 2, 2020

$\rightarrow P \in \Delta^d$ means $P \in \mathbb{R}^d_+$ and $\sum_{i=1}^d p_i = 1$

$\uparrow$ simplex.

This is a die with d sides, side i is the outcome with probability p_i

If we roll this die n times what is p?

$$\boxed{\frac{1}{n} \sum_{t=1}^n e_{x_t} = \hat{p}} \Rightarrow \hat{p}_i = \frac{\sum_{t=1}^n I(x_t = i)}{n}$$

we hope $\hat{p} \simeq p$

$e_i = \begin{bmatrix} \vdots \\ 1 \\ \vdots \end{bmatrix} \leftarrow 1$ at ith index

---

Theorem (Concentration on simplex)

[Weissman et al. 2003]

Let $X_1, \ldots X_n$ be i.i.d sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_1 : \Omega \to \{1, \ldots m\}$

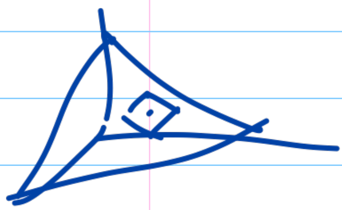
Let $p_i = \mathbb{P}(X_1 = i)$ and $\hat{p}_i = \frac{1}{n} \sum_{t=1}^n I\{X_t = i\}$

Then:

$$\mathbb{P}\left(\|P - \hat{P}\|_1 > \sqrt{\frac{2m \log \frac{2}{\delta}}{n}} \right) < \delta$$

Monday, November 2, 2020

Now back to MDPs and UCRL2:



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

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

(If $T_t(x,a) = 0 \to$ set $L_{t-1}(x,a) = 1$)

$$\to \quad C_t^\delta = \left\{ P' = \{P'(\cdot \mid x,a)\} \;;\; P'(\cdot \mid x,a) \in C_t^\delta(x,a), \forall (x,a) \in \mathcal{X}\times\mathcal{A} \right\}$$

The set of plausable model.

For $\quad L_t(x,a) \leq L := 2\log\left(\frac{4|\mathcal{X}||\mathcal{A}|\, T(T+1)}{\delta}\right)$

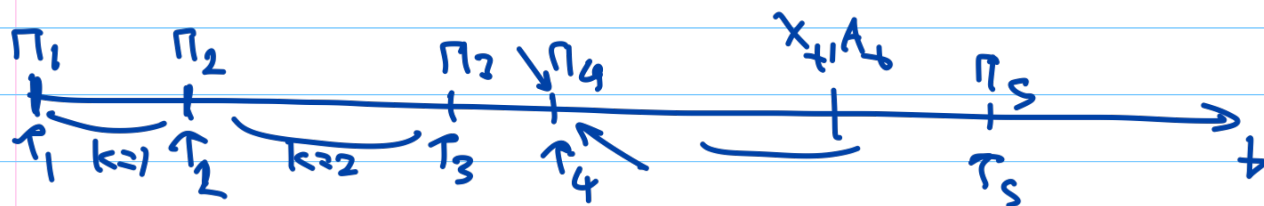
We want the true probability kernel P to be in C_t^δ for all T, with probability at least $1-\delta$.

Monday, November 2, 2020

UCRL 2:
- start with a policy
- Collect samples
- stop when the number samples at least for a pair of (x,a) is doubled,
- Compute $C_t^{\delta}()$ and P_t
- compute optimistic model $\tilde{P}_t$

- Compute π_k, the optimistic policy

- Then start new epoch $k+1$



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

otherwise stop the epoch.

How to come up with optimistic model.

we want $\rho_k = \max_{x \in \mathcal{X}} \max_{\pi} \max_{P' \in C_{\tau_k}} \rho_x^{\pi}(P')$

The best $\hat{\rho}_k$

and $\rho_k + V_k(x) = \max_{a \in \mathcal{A}} r(x,a) + \langle \tilde{P}_k(\cdot \mid x, a), V_k \rangle$

$\forall (x,a) \in \mathcal{X} \times \mathcal{A}$

Monday, November 2, 2020

Define $\tilde{M}_k$, the extended MDP with the same statespace as M, but extended action space

$$\tilde{A}_x = \{(a, P') : a \in A, P' \in C_{P_k}(x,a)\}$$

with the reward $r(x,(a,P')) = r(x,a)$

and transition $P(\cdot|x,(a,P')) = P'(\cdot|x,a)$

For any policy π in M, there exist $\tilde{\pi}$ on $\tilde{M}$

such that for any $x \Rightarrow a = \pi(x)$ on M

$\pi(x), P = \tilde{\pi}(x)$ on $\tilde{M}$

$$\Rightarrow \boxed{D(\tilde{M}) \leq D}$$

Solve for ρ and V of $\tilde{M}$, then π_k follows

$$\pi_k(x) \in \arg\max_{a \in A} r(x,a) + \max_{P'(x,a) \in C^{\delta}_{P_k}(x,a)} \langle P'(x,a), V \rangle$$

There is an efficient alg to solve it (UCRL2)

Monday, November 2, 2020

Lemma: For a (ρ, V), satisfying Bellman optimality equation:

$$\text{span}(V) \leq D \qquad \text{[UCRL2, Chp 38 Bandit book]}$$

Solving Bellman optimality for $\tilde{M}_k$ give $\rho_k, \tilde{P}_k, V_k$. since V_k is not unique and $\text{span}(V_k) \leq D$, we choose V_k s.t. $\|V_k\|_\infty \leq \frac{D}{2}$

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

$$\hat{R}_T = \sum_{t=1}^{T} \left(\rho - r(X_t, A_t)\right) = \sum_{k=1}^{K} \sum_{t=\tau_k}^{\tau_{k+1}-1} \left(\rho - r(X_t, A_t)\right)$$

Optimism step:

$$\leq \sum_{k=1}^{K} \underbrace{\sum_{t=\tau_k}^{\tau_{k+1}-1} \left(\rho_k - r(X_t, A_t)\right)}_{R_k}$$

with probability at least $1-\delta$.

Also we know: $\rho_k = r_{\pi_k}(X_t) - V_k(X_t) + \langle \tilde{P}_k(\cdot | X_t, \pi_k(X_t)), V_k \rangle$

Monday, November 2, 2020

$$R_k \leq \sum_{t=\tau_k}^{\tau_{k+1}-1} \left(-V_k(x_t) + \langle \tilde{P}_k(\cdot|x_t, \pi_k(x_t)), V_k \rangle\right)$$

$$= \sum_{t=\tau_k}^{\tau_{k+1}-1} \left(-V_k(x_t) + \underline{\langle P(\cdot|x_t, A_t), V_k \rangle}\right) \rightarrow A_1$$

$$+ \sum_{t=\tau_k}^{\tau_{k+1}-1} \underline{\langle \tilde{P}_k(\cdot|x_t, A_t) - P(\cdot|x_t, A_t), V_k \rangle} \rightarrow A_2$$

$$(A_1) = \underbrace{\sum_{t=\tau_k}^{\tau_{k+1}-1} \left(V_k(x_{t+1}) - V_k(x_t)\right)}_{V_k(x_{\tau_{k+1}}) - V_k(x_{\tau_k}) \leq D} + \underline{\langle P(\cdot|x_t, A_t), V_k \rangle - V_k(x_{t+1})}$$

$$(A_1) \leq D + \sum_{t=\tau_k}^{\tau_{k+1}-1} \left(E[V_k(x_{t+1})|F_t] - V_k(x_t)\right)$$

$\hookrightarrow$ mean zero martingale sequence, and bounded adapted to F_t

using Holder inequality

$$A_2 \leq \sum_{t=\tau_k}^{\tau_{k+1}-1} \underbrace{\|P_k(\cdot|x_t, A_t) - P(\cdot|x_t, A)\|_1}_{\text{from the confidence interval}} \underbrace{\|V_k\|_\infty}_{D/2}$$

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

Monday, November 2, 2020

$$T_{(k)}(x,a) = \sum_{t=\tau_k}^{\tau_{k+1}} I(X_t = x, A_t = a)$$