Week 9 - Lecture 43

$$R(x,a) = \langle \underbrace{\Psi(x,a)}_{\text{known}}, \theta^* \rangle$$

unknown

$\theta^* \in \mathbb{R}^d$

known

feature map

$\Psi(x,a) \in \mathbb{R}^d$

$\forall\ (x,a)$

Example: Recommendation Engine [User: context, Product: arm]

Products can be categorized : Sports, daily use, movie etc.

Users can be categorized: Sex, Age, location

User : young Girl

(sex, Age, location)

(1, 1, 0)

Product : DVD, tennis ball

(Sport, movie, joy)

(0, 1, 1)

(1, 0, 1)

Depending on user & product feature map can be formed.

Such feature maps can be created.

These feature maps can be in millions, but these are generated offline.

Thus the problem is to find θ^*, with interaction with env.

θ^* is a const., doesn't depend on context & action.

Fixed for an env., changes with change in env.

Assume dimension of θ^* is known, also assume norm is bounded.

$\|\theta^*\|_2^2 \leq L$ L is known

So,

$|R(x,a) - R(x',a')|$

$$\leq \|\theta^*\|_2 \, \|\Psi(x,a) - \Psi(x',a')\|$$

$$\boxed{\begin{array}{l}\text{Lipschitz function:} \\ |f(x) - f(x')| \leq L\,\|x - x'\|\end{array}}$$

Thus $|R(x,a) - R(x',a')|$ is Lipschitz with const. L.

This limits θ^* to a bounded space.

Thus, assuming that:

Rewards are linear & parametrized by θ^*.

Thus in the game:

$\boxed{1}$, $\boxed{2}$, $x_t \to \boxed{3}$, $\vdots$, $\boxed{K}$

$$\{\Psi(x_t, a)\}_{a \in [K]} = D_t$$ ← feature set (Decision set)

Choose the feature map with the given context.

What we have thus is:

$$R(x,a) = \langle \Psi(x,a), \theta^* \rangle$$

where we need to maximize this over the feature vectors.

Thus this can be seen as linear optimization problem over a feature set

$\{D_t\}_{t \geq 1}$ Now find:

$$d_t^* = \underset{d \in D_t}{\arg\max} \langle d, \theta^* \rangle$$

After making the assumtion that rewards are linearly parametrized by θ^* $\therefore$ arms have no real significance, but we need to maximize the features.

Thus: Stochastic Linear Bandits

$\{D_t\}_{t \geq 1}$ are sequence of decision sets choose d_t^* s.t.

$$d_t^* = \underset{d \in D_t}{\operatorname{argmax}} \langle d, \theta^* \rangle$$

$d_t \in D_t$ in round t,

$$r_t = \langle d_t, \theta^* \rangle + \eta_t$$

<u>Regret</u> :

$$\hat{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \max_{d_t \in D_t} \langle d, \theta^* \rangle - \sum_{t=1}^{T} \langle d_t, \theta^* \rangle\right]$$

If D_t is a set of unit vectors

$$D_t = \{e_1, e_2, e_3 \ldots\ldots e_d\}$$

where
$$e_1 = \{1, 0, 0 \ldots\ldots 0\}$$
$$e_2 = \{0, 1, 0 \ldots\ldots 0\}$$
$$\vdots$$
$$e_d = \{0, 0, 0 \ldots\ldots 1\}$$

Assume θ^* is a vector with each component being the mean of each arm.

$\therefore \quad e_i^T \theta^* = \theta_i$

Thus the stochastic Linear Bandits captures the Stochastic D-armed Bandits.

(if θ^* has D-dimensions)

But D_t need not be unit vectors

$$\boxed{D_t \subset \mathbb{R}^d}$$