Upper confidence bound strategy on stochastical bandits

Multiarmed bandit: K arms, at each step we can choose one arm to be pulled while the other K-1 arms stay frozen (no reward).

- **Stochastic bandit:** Each arm has fixed distribution in all rounds.
- **Adversarial bandit:** Bandits can change payout in each round.
- **Markovian bandit:** Activated arm changes in a 'Markovian style'.

We are only looking at stochastic bandits and Markovian bandits.

Stochastic bandits

K arms with an unknown, fixed probability distribution $\nu_1, ..., \nu_K$ on $[0,1]$. At each step $t = 1, 2, ...$ choose arm $I_t \in \{1, ..., K\}$ and draw reward $X_{I_t,t} \sim \nu_{I_t}$ independent of the past.

Let μ_i be the mean of ν_i, $\mu^* = \max\limits_{i=1,...,K} \mu_i$ and $i^* \in \operatorname*{argmax}\limits_{i=1,..,K} \mu_i$.

The **regret** after n rounds is defined as $R_n := \max\limits_{i=1,...,K} \sum_{t=1}^n X_{i,t} - \sum_{t=1}^n X_{I_t,t}$

The **pseudo-regret** is $\overline{R}_n := \max\limits_{i=1,...,K} \mathbb{E}[\sum_{t=1}^n X_{i,t} - \sum_{t=1}^n X_{I_t,t}] = n\mu^* - \sum_{t=1}^n \mathbb{E}[\mu_{I_t}]$

By defining $N_n(i) = \sum_{t=1}^s \mathbb{1}_{I_t=i}$, i.e number of times arm i is pulled up to time n, and let $\triangle_i = \mu^* - \mu_i$ we can rewrite the pseudo-regret as

$$\overline{R}_n = \sum_{i=1}^K \mathbb{E}[N_n(i)]\mu^* - \sum_{i=1}^K \mathbb{E}[N_n(i)\mu_i] = \sum_{i=1}^K \triangle_i \mathbb{E} N_n(i)$$

The upper confidence bound strategy (UCB)

For the UCB strategy we need the following assumption:
There is a convex function ψ on $\mathbb{R}$ such that, $\forall \lambda \geq 0$:

$$\ln \mathbb{E}\mathrm{e}^{\lambda(X-\mathbb{E}[X])} \leq \psi(\lambda), \quad \text{and} \quad \ln \mathbb{E}\mathrm{e}^{\lambda(E[X]-X)} \leq \psi(\lambda) \tag{1}$$

Note that if $X \in [0,1]$ we can take $\psi(\lambda) = \lambda^2/8$. (Hoeffding's lemma)
The Legendre-Fenchel (also known as the convex conjugate) of ψ is defined as

$$\psi^*(\epsilon) = \sup_{\lambda\in\mathbb{R}}(\lambda\epsilon - \psi(\lambda))$$

Note that for $\psi(\lambda) = \lambda^2/8$ we have $\psi^*(\epsilon) = 2\epsilon^2$
Let $\hat{\mu}_{i,s}$ be the sample mean of the rewards, i.e $\hat{\mu}_{i,s} = \frac{1}{s}\sum_{t=1}^s X_{i,s}$ in distribution since the rewards are i.i.d.
By Markov's inequality and by equation (1) we obtain

$$\mathbb{P}(\mu_i - \hat{\mu}_{i,s} > \epsilon) \leq \mathrm{e}^{-s\psi^*(\epsilon)} \tag{2}$$

And by defining $\delta = \mathrm{e}^{-s\psi^*(\epsilon)}$ we have, with probability at least $1-\delta$

$$\hat{\mu}_{i,s} + (\psi^*)^{-1}(\frac{1}{s}\ln(\frac{1}{\delta})) > \mu_i$$

Hence, for a parameter $\alpha > 0$ the (α, ψ)-UCB strategy is to select the arm

$$I_t \in \operatorname*{argmax}_{i=1,...,K} \left[\hat{\mu}_{i,N_{t-1}(i)} + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{N_{t-1}(i)}\right)\right]$$

<u>Theorem</u> (Pseudo-regret for UCB strategy):
Assume that the ν_i satisfy the convex assumption (1). Then the pseudo-regret for a (α, ψ)-UCB stategy with $\alpha > 2$ satisfies

$$\overline{R}_n \leq \sum_{i:\triangle_i > 0} \left(\frac{\alpha \triangle_i}{\psi^*(\triangle_i/2)} \ln n + \frac{\alpha}{\alpha - 2} \right)$$

If we have $X \in [0, 1]$, using $\psi^*(\epsilon) = 2\epsilon^2$, then

$$\overline{R}_n \leq \sum_{i:\triangle_i > 0} \left(\frac{2\alpha}{\triangle_i} \ln n + \frac{\alpha}{\alpha - 2} \right)$$

<u>Lower bound for Bernoulli-distributed rewards</u>

For the following result, we are assuming that $X_{i,t} \sim \mathit{Bernoulli}(p, q)$ with $p, q \in [0, 1]$

<u>Theorem</u> (Lower bound):
Assume $\mathbb{E}N_n(i) = o(n^a)$ for $a > 0$ and that $\triangle_i > 0 \ \forall i$. Then we have

$$\liminf_{n \to \infty} \frac{\overline{R}_n}{\ln n} \geq \sum_{i:\triangle_i > 0} \frac{\triangle_i}{kl(\mu_i, \mu^*)}$$

where $kl(\mu_i, \mu^*) = \mu_i \ln \left(\frac{\mu_i}{\mu^*} \right) + (1 - \mu_i) \ln \left(\frac{1 - \mu_i}{1 - \mu^*} \right)$ is the Kullback-Leibler divergence.

<u>Comparision of lower & upper bound</u>

We have that

$$kl(\mu_i, \mu^*) \leq \frac{(\mu^* - \mu_i)^2}{\mu^*(1 - \mu^*)}$$

which follows from *ln x* $\leq$ *x − 1*. Hence, the lower bound satisfies

$$\liminf_{n \to \infty} \frac{\overline{R}_n}{\ln n} \geq \sum_{i:\mu^* - \mu_i > 0} \frac{\mu^*(1 - \mu^*)}{(\mu^* - \mu_i)}$$

Comparing this with the upper bound

$$\overline{R}_n \leq \sum_{i:\mu^* - \mu_i > 0} \left(\frac{2\alpha}{\mu^* - \mu_i} \ln n + \frac{\alpha}{\alpha - 2} \right)$$

we see that the difference between upper and lower bound for a Bernoulli-distributed reward is given by some constants.

Markovian bandits

Again we consider K arms, at each step we can choose one arm to be pulled while the remaining K-1 arms stay frozen. But now the rewards of the pulled arm can change its state in a 'Markovian style', i.e the arm produces reward $r(x_t)$ and changes start to x_{t+1} according to a Markov dynamic $x \to y$ with $\mathbb{P}(x, y)$

The goal now it to maximize a **β-discounted reward**

$$\mathbb{E}\left[\sum_{t=0}^{\infty} r_{i_t}(x_{i_t}(t))\beta^t\right]$$

where i_t is the arm pulled at time t and $0 < \beta < 1$ is the discounting factor. This discounted reward is maximized by forward induction.

It can be shown (not part of the talk) that the biggest Gittins index

$$G_i(x_i) = \sup_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t | x_i(0) = x_i\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t | x_i(0) = x_i\right]}, \text{ where } \tau \text{ is a stopping time,}$$

is enough to determine which arm is to be pulled.

Note that the numerator denotes the discounted rewards up to τ and the denumerator represents the discounted time up to τ.

Hence, we can find the best strategy by computing the Gittins Index for all arms, where each index is independent of all other arms. Thus, we only need to solve a K-dimensional problem in each step, which greatly reduces the computational work.