Statistical Techniques in Robotics (16-831, F11) Lecture#23 (November 20, 2014)

Bandit Problems

Lecturer: Venkatraman Narayanan & Karthik Lakshmanan Scribes: Nate Otten & Hanbyul Joo[1]

1 Bandits and Online Learning

In Generalized Weighted Majority (GWM), we have lots of experts and pick one to minimize regret:

$$R_T = \sum_{t=1}^{T} l_t(alg) - \min_i \sum_t l_t(e_i)$$

$$\frac{R_T}{T} \to 0$$

where e_i is the best expert in hindsight. But now let's say that at time step T, we make a prediction and only see the loss for the expert we picked instead of the losses for all experts.

$$\sqrt{\ln NT}$$

$$\frac{1}{\beta} \ln N + \beta \sum_t l_t(e*)$$

We need order $\sqrt{n}$ time/data.

1.1 Stochastic Bandit Setting

The word "bandit" refers to "one-armed bandits," another name for slot machines. The problem is that initially we don't know the reward distribution of any of the bandits, and we can only try them one at a time, so based on the outcomes so far we must choose to exploit the current bandit or explore others. For this problem, we'll think about rewards instead of losses and assume the distributions doesn't change with time (i.i.d.). At time t, our reward from bandit i is a random variable r_i^t.

$$R_t = \left(\max_i \sum_{t=1}^{T} r_i^t \right) - \sum_{t=1}^{T} r_{I_t}^t$$

$$I_t \in 1, 2, \ldots, N$$

The problem is that r_i^t is drawn from a distribution, meaning we need to compute the expected value, not the sum; only this is too hard to compute exactly, so we will use the pseudo regret

[1]Some content adapted from previous scribe: Bradford Neuman (16-899 ACRL, S10)

instead.

$$\begin{aligned}
E[R_t] &= E\left[\max_i \sum_t r_i^t - \sum_t r_{I_t}^t\right] \\
&= \max_i E\left[\sum_t r_i^t\right] - E\left[\sum_t r_{I_t}^t\right] \quad \text{pseudo regret} \\
&= \max_i \sum_t E\left[r_i^t\right] - E\left[\sum_t r_{I_t}^t\right] \\
&= \max_i T\mu_i - E\left[\sum_t r_{I_t}^t\right] \quad \text{dist mean } \mu_i, \quad i = 1, \ldots, N \\
&= T\mu_i^* - E\left[\sum_t \mu_{I_t}\right]
\end{aligned}$$

N_i^T is the number of times arm i is pulled in T time steps.

$$\begin{aligned}
\bar{E}[R_T] &= T\mu_i^* - \sum_{i=1}^N E[N_i^T]\mu_i \\
&= \sum_{i=1}^N E[N_i^T]\mu_i^* - \sum_{i=1}^N E[N_i^T]\mu_i \\
&= \sum_{i=1}^N E[N_i^T](\mu_i^* - \mu_i) \\
&= \sum_{i=1}^N E[N_i^T]\Delta i
\end{aligned}$$

1.2 Upper Confidence Bound (UCB)

The sample mean is given by

$$\hat{\mu}_i^t = \frac{1}{N_i^t}\sum r_i$$

We can't be greedy and need to keep track of the sample confidence. The upper confidence bound (UCB) is given by the estimated sample mean plus the confidence.

$$UCB = \hat{\mu}_i^t + \sqrt{\frac{\alpha \ln T}{2N_i^t}}$$

If the total number of arm pulls T is unknown, we can replace it with t. Some combination of high mean and high uncertainty makes us want to pull the arm. In other words, we are drawn to the bandits that are paying out large rewards and those that we know little about. We want to upper bound the number of times we will pull arm i, so we will attempt to compute $E[N_i^T]$. We will do so using the following lemma.

Lemma 1. *If arm i is pulled, then at least one of the following must be true.*

1. $\hat{\mu}_i^{t-1*} \leq \mu^* - \sqrt{\frac{\alpha \ln T}{2N_i^{t-1*}}}$
2. $\hat{\mu}_i^{t-1} \geq \mu_i - \sqrt{\frac{\alpha \ln T}{2N_i^{t-1}}}$
3. $N_i^{t-1} \leq \frac{2\alpha \ln T}{\Delta i^2}$

To prove this, we only need to assume all three are false and show that it leads to a contradiction.

$$\begin{aligned}
\hat{\mu}_i^{t-1*} + \sqrt{\frac{\alpha \ln T}{2N_i^{t-1*}}} &> \mu^* \\
&= \mu_i + \Delta i \quad \text{where } \Delta i = \mu^* - \mu_i \\
&\geq \mu_i + \sqrt{\frac{2\alpha \ln T}{N_i^{t-1}}} \\
&> \hat{\mu}_i^{t-1} + \sqrt{\frac{2 \ln T}{2N_i^{t-1}}}
\end{aligned}$$

Now we will bound the number of times the arm i is pulled:

$$\begin{aligned}
E[N_i^T] &= E\left[\sum_{t=1}^{T} \mathbb{1}(I_t = i)\right] \\
&= E\left[\sum_{t=1}^{T} \mathbb{1}(I_t = i, N_i^{t-1} \leq t_0) + \sum_{t=1}^{T} \mathbb{1}(I_t = i, N_i^{t-1} > t_0)\right] \\
&\leq t_0 + E\left[\sum_{t=t_0}^{T} \mathbb{1}(I_t = i, N_i^{t-1} > t_0)\right]
\end{aligned}$$

$$t_0 = \sqrt{\frac{2\alpha \ln T}{\Delta i^2}}$$

$$P(1 \wedge 2 \wedge 3) \leq t_0 + \sum_{t=t_0+1}^{T} P(I_t = i, N_i^{t-1} > t_0)$$

$$P(\neg 3) = P(1 \wedge 2) = t_0 + \sum_{t=t_0+1}^{T} P(1 \wedge 2) \leq t_0 + \sum_{t=t_0+1}^{\infty} \frac{1}{T\alpha}$$

$$P\left(\hat{\mu}_i^{t-1*} \leq \mu^* - \sqrt{\frac{\alpha \ln T}{2N_i^{t-1*}}}\right) \leq \exp\left(-2 \times (t-1) \times \frac{\alpha \ln T}{2(t-1)}\right)$$

$$P(\tilde{x} - x > \epsilon) < e^{-2n\epsilon^2} \leq \frac{1}{T\alpha}$$

Finally...

$$
\begin{aligned}
E[N_i^T] &\leq t_0 + \frac{\alpha}{\alpha - 2} \\
\bar{E}[R_i] &\leq \sum_{i=\Delta i>0} \left(\frac{2\alpha}{\Delta i}\right) \ln T + \frac{\alpha}{\alpha - 2} \sum \Delta i \\
&\leq \sqrt{\alpha N T \ln T}
\end{aligned}
$$

No regret.

2 Exp3 (Adversarial bandit setting)

In the non-stochastic (or adversarial case) we cannot use UCB directly. EXP3 can be used in this case. Intuitively, as shown in 1, EXP3 exploits Generalized Weighted Majority (GWM) by passing a unbiased loss vector $\hat{l}_t^i$ to the experts of GWM as follows:

$$
\hat{l}_t^i = \begin{pmatrix} \cdots \\ 0 \\ 0 \\ \frac{l_t^i}{P_t^i} \\ 0 \\ 0 \\ \cdots \end{pmatrix}, \tag{1}
$$

And, in Exp3,

$$
E\left[R_T\right] \leq \Sigma_{1:\triangle>0} \frac{2\alpha}{\triangle_i} \ln T + \frac{\alpha}{\alpha - 2}
$$

We assume **N** experts:

$$
e_i, i \in \{1, N\}
$$

At time t, the algorithm is

$$
\begin{aligned}
\text{Pick } e_t \propto P_t^i &= \frac{w_t^i}{\Sigma_j W_t^j} \\
\text{Receive } w_{t+1}^i &= w_t^i e^{-\epsilon l_t^i}
\end{aligned}
$$

,where l_t^i is elements of the loss vector $\bar{l}_t$. Then,

$$
E_{p_t^i}\left[\hat{l}_t^i\right] = P_t^i \left(\frac{l_t^i}{P_t^i}\right) + (1 - P_t^i) = l_t^i
$$

And, this is still no regret. The whole algorithm is:

At time t,
Pick $e_t^i \propto P_t^i$
Receive l_t
for *j=1 to N* **do**
  if *j=i* **then**
    $w_{t+1}^i = w_t^i e^{-\epsilon \hat{l}_t^i}$
  else
    $w_{t+1}^i = w_t^i$
  end
end

Algorithm 1: Exp3

Proof Sketch

1. $R \leq E_{algo}[R]$
2. Use GWM to bound $E_{algo}[R]$

Proof of 1:

$$\hat{R} = \Sigma_{t=1}^T \left\langle P_t, \hat{l}_t \right\rangle - \min_i \Sigma_{t=1}^T \hat{l}_t^i$$

$$= \Sigma_{t=1}^T \langle P_t, l_t \rangle - \min_i \Sigma_{t=1}^T l_t^i \qquad \text{(By Jensen's inequailty)}$$

$$E_{P_1,\ldots,P_T}\left[\hat{R}\right] = E_{P_1 \ldots P_T} \left\langle P_t, \hat{l}_t \right\rangle - \min_i \Sigma_{t=1}^T \hat{l}_t^i$$

$$= \Sigma_{t=1}^T l_t - E_{P_1 \ldots P_T} \min_i \Sigma_{t=1}^T \hat{l}_t^i$$

$$\geq \Sigma_{t=1}^T \langle P_t, l_t \rangle - \min_i \Sigma_{t=1}^T l_t^i$$

$$\text{Thus, } E\left[\hat{R}\right] \geq \hat{R}$$

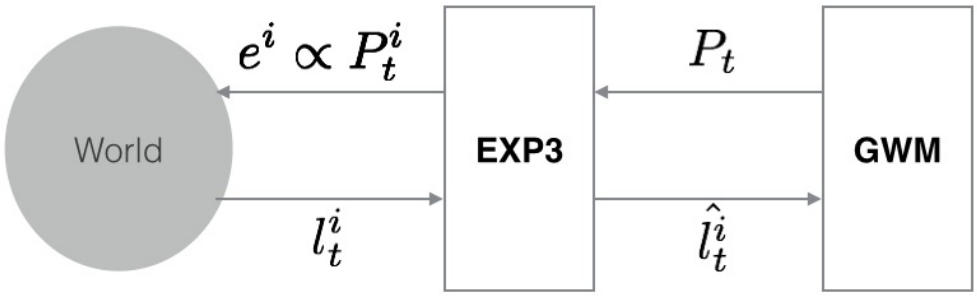

Figure 1:

Proof of 2:

$$\hat{R} \leq \frac{1}{\epsilon}\log N + \epsilon\Sigma_{t=1}^{T}\|\hat{l}_t\|^2$$
$$\leq \frac{1}{\epsilon}\log N + \frac{\epsilon}{2}\Sigma_{t=1}^{T}\Sigma_{t=1}^{N}P_t\hat{l_t^i}^2$$

$$E\left[\hat{R}\right] \leq \frac{logN}{\epsilon} + \frac{\epsilon}{2}NT$$

Let's pick ϵ as follows:

$$\epsilon = \sqrt{\frac{2\log N}{NT}}$$

Then,

$$\frac{\log N\sqrt{NT}}{\sqrt{2\log N}} + \frac{1}{2}\sqrt{\frac{2logN}{NT}}NT = \sqrt{2TN\log N}$$

In a conclusion, Exp3 and UCB are algorithms for different settings, and UCB has better regret in terms of number of experts, and in terms of the time horizon, Exp3 is better than UCB.