18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet Lecture 18
Scribe: Haihao (Sean) Lu Nov. 2, 2015

3. STOCHASTIC BANDITS

3.1 Setup

The stochastic multi-armed bandit is a classical model for decision making and is defined as follows:

There are K arms(different actions). Iteratively, a decision maker chooses an arm $k \in \{1, \dots, K\}$, yielding a sequence $X_{K,1}, \dots, X_{K,t}, \dots$, which are i.i.d random variables with mean μ_k. Define $\mu_* = \max_j \mu_j$ or $* \in \arg\max$. A policy π is a sequence $\{\pi_t\}_{t \geq 1}$, which indicates which arm to be pulled at time t. $\pi_t \in \{1, \dots, K\}$ and it depends only on the observations strictly interior to t. The regret is then defined as:

$$\begin{aligned} R_n &= \max_k \mathbb{E}[\sum_{t=1}^n X_{K,t}] - \mathbb{E}[\sum_{t=1}^n X_{\pi_t,t}] \\ &= n\mu_* - \mathbb{E}[\sum_{t=1}^n X_{\pi_t,t}] \\ &= n\mu_* - \mathbb{E}[\mathbb{E}[\sum_{t=1}^n X_{\pi_t,t} \mid \pi_t]] \\ &= \sum_{k=1}^K \Delta_k \mathbb{E}[T_k(n)] \ , \end{aligned}$$

where $\Delta_k = \mu_* - \mu_k$ and $T_k(n) = \sum_{t=1}^n \mathbb{1}(\pi_t = k)$ is the number of time when arm k was pulled.

3.2 Warm Up: Full Info Case

Assume in this subsection that $K = 2$ and we observe the full information $\begin{pmatrix} X_{1,t} \\ \vdots \\ X_{K,t} \end{pmatrix}$ at time t after choosing π_t. So in each iteration, a normal idea is to choose the arm with highest average return so far. That is

$$\pi_t = \operatorname*{argmax}_{k=1,2} \bar{X}_{k,t}$$

where

$$\bar{X}_{k,t} = \frac{1}{t} \sum_{s=1}^t X_{k,s}$$

Assume from now on that all random variable $X_{k,t}$ are subGaussian with variance proxy σ^2, which means $\mathbb{E}[e^{ux}] \leq e^{\frac{u^2\sigma^2}{2}}$ for all $u \in \mathbb{R}$. For example, $N(0, \sigma^2)$ is subGaussian with variance proxy σ^2 and any bounded random variable $X \in [a, b]$ is subGaussian with variance proxy $(b-a)^2/4$ by Hoeffding's Lemma.

Therefore,

$$R_n = \Delta \mathbb{E}[T_2(n)] \ , \tag{3.1}$$

where $\Delta = \mu_1 - \mu_2$. Besides,

$$\begin{aligned} T_2(n) &= 1 + \sum_{t=2}^{n} \mathbb{I}(\bar{X}_{2,t} > \bar{X}_{1,t}) \\ &= 1 + \sum_{t=2}^{n} \mathbb{I}(\bar{X}_{2,t} - \bar{X}_{1,t} - (\mu_2 - \mu_1) \geq \Delta) \ . \end{aligned}$$

It is easy to check that $(\bar{X}_{2,t} - \bar{X}_{1,t}) - (\mu_2 - \mu_1)$ is centered subGaussian with variance proxy $2\sigma^2$, whereby

$$\mathbb{E}[\mathbb{I}(\bar{X}_{2,t} > \bar{X}_{1,t})] \leq e^{-\frac{t\Delta^2}{4\sigma^2}}$$

by a simple Chernoff Bound. Therefore,

$$R_n \leq \Delta(1 + \sum_{t=0}^{\infty} e^{-\frac{t\Delta^2}{4\sigma^2}}) \leq \Delta + \frac{4\sigma^2}{\Delta} \ , \tag{3.2}$$

whereby the benchmark is

$$R_n \leq \Delta + \frac{4\sigma^2}{\Delta} \ .$$

3.3 Upper Confidence Bound (UCB)

Without loss of generality, from now on we assume $\sigma = 1$. A trivial idea is that after s pulls on arm k, we use $\hat{\mu}_{k,s} = \frac{1}{s} \sum_{j \in \{\text{pulls of } k\}} X_{K,j}$ and choose the one with largest $\hat{\mu}_{k,s}$. The problem of this trivial policy is that for some arm, we might try it for only limited times, which give a bad average and then we never try it again. In order to overcome this limitation, a good idea is to choose the arm with highest upper bound estimate on the mean of each arm at some probability lever. Note that the arm with less tries would have a large deviations from its mean. This is called Upper Confidence Bound policy.

Algorithm 1 Upper Confidence Bound (UCB)

```
for $t = 1$ to $K$ do
  $\pi_t = t$
end for
for $t = K + 1$ to $n$ do
  $T_k(t) = \sum_{s=1}^{t-1} \mathbb{1}(\pi_t = k)$
  (number of time we have pull arm $k$ before time $t$)
  $\hat{\mu}_{k,t} = \frac{1}{T_k(t)} \sum_{s=1}^{t-1} X_{K,t \wedge s}$
  $\pi_t \in \underset{k \in [K]}{\operatorname{argmax}} \left\{ \hat{\mu}_{k,t} + 2\sqrt{\frac{2\log(t)}{T_k(t)}} \right\}$ ,
end for
```

Theorem: The UCB policy has regret

$$R_n \leq 8 \sum_{k, \Delta_k > 0} \frac{\log n}{\Delta_k} + (1 + \frac{\pi^2}{3}) \sum_{k=1}^{K} \Delta_k$$

Proof. From now on we fix k such that $\Delta_k > 0$. Then

$$\mathbb{E}[T_k(n)] = 1 + \sum_{t=K+1}^{n} \mathbb{P}(\pi_t = k) \ .$$

Note that for $t > K$,

$$\begin{aligned}\{\pi_t = k\} &\subseteq \{\hat{\mu}_{k,t} + 2\sqrt{\frac{2\log t}{T_k(t)}} \leq \hat{\mu}_{*,t} + 2\sqrt{\frac{2\log t}{T_*(t)}}\} \\ &\subseteq \left\{ \{\mu_k \geq \hat{\mu}_{k,t} + 2\sqrt{\frac{2\log t}{T_k(t)}}\} \bigcup \{\mu_* \geq \hat{\mu}_{*,t} + 2\sqrt{\frac{2\log t}{T_*(t)}}\} \bigcup \{\mu_* \leq \mu_k + 2\sqrt{\frac{2\log t}{T_k(t)}}, \pi_t = k\} \right\}\end{aligned}$$

And from a union bound, we have

$$\begin{aligned}\mathbb{P}(\hat{\mu}_{k,t} - \mu_k < -2\sqrt{\frac{2\log t}{T_k(t)}}) &= \mathbb{P}(\hat{\mu}_{k,t} - \mu_k < 2\sqrt{\frac{2\log t}{T_k(t)}}) \\ &\leq \sum_{s=1}^{t} \exp(\frac{-s\frac{8\log t}{s}}{2}) \\ &= \frac{1}{t^3}\end{aligned}$$

Thus $\mathbb{P}(\mu_k > \hat{\mu}_{k,t} + 2\sqrt{\frac{2\log t}{T_k(t)}}) \leq \frac{1}{t^3}$ and similarly we have $\mathbb{P}(\mu_* > \hat{\mu}_{*,t} + 2\sqrt{\frac{2\log t}{T_*(t)}}) \leq \frac{1}{t^3}$, whereby

$$\begin{aligned}
\sum_{t=K+1}^{n} \mathbb{P}(\pi_t = k) &\leq 2\sum_{t=1}^{n} \frac{1}{t^3} + \sum_{t=1}^{n} \mathbb{P}(\mu_* \leq \mu_k + 2\sqrt{\frac{2\log t}{T_k(t)}}, \pi_t = k) \\
&\leq 2\sum_{t=1}^{\infty} \frac{1}{t^3} + \sum_{t=1}^{n} \mathbb{P}(T_k(t) \leq \frac{8\log t}{\Delta_k^2}, \pi_t = k) \\
&\leq 2\sum_{t=1}^{\infty} \frac{1}{t^3} + \sum_{t=1}^{n} \mathbb{P}(T_k(t) \leq \frac{8\log n}{\Delta_k^2}, \pi_t = k) \\
&\leq 2\sum_{t=1}^{\infty} \frac{1}{t^3} + \sum_{s=1}^{\infty} \mathbb{P}(s \leq \frac{8\log n}{\Delta_k^2}) \\
&\leq 2\sum_{t=1}^{\infty} \frac{1}{t^2} + \frac{8\log n}{\Delta_k^2} \\
&= \frac{\pi^2}{3} + \frac{8\log n}{\Delta_k^2} \ ,
\end{aligned}$$

where s is the counter of pulling arm k. Therefore we have

$$R_n = \sum_{k=1}^{K} \Delta_k \mathbb{E}[T_k(n)] \leq \sum_{k,\Delta_k>0} \Delta_k(1 + \frac{\pi^2}{3} + \frac{8\log n}{\Delta_k^2}) \ ,$$

which furnishes the proof. □

Consider the case $K = 2$ at first, then from the theorem above we know $R_n \sim \frac{\log n}{\Delta}$, which is consistent with intuition that when the difference of two arm is small, it is hard to distinguish which to choose. On the other hand, it always hold that $R_n \leq n\Delta$. Combining these two results, we have $R_n \leq \frac{\log n}{\Delta} \wedge n\Delta$, whereby $R_n \leq \frac{\log(n\Delta^2)}{\Delta}$ up to a constant. Actually it turns out to be the optimal bound. When $K \geq 3$, we can similarly get the result that $R_n \leq \sum_k \frac{\log(n\Delta_k^2)}{\Delta_k}$. This, however, is not the optimal bound. The optimal bound should be $\sum_k \frac{\log(n/H)}{\Delta_k}$, which includes the harmonic sum and $H = \sum_k \frac{1}{\Delta_k^2}$. See [Lat15].

3.4 Bounded Regret

From above we know UCB policy can give regret that increases with at most rate $\log n$ with n. In this section we would consider whether it is possible to have bounded regret. Actually it turns out that if there is a known separator between the expected reward of optimal arm and other arms, there is a bounded regret policy.

We would only consider the case when $K = 2$ here. Without loss of generality, we assume $\mu_1 = \frac{\Delta}{2}$ and $\mu_2 = -\frac{\Delta}{2}$, then there is a natural separator 0.

Algorithm 2 Bounded Regret Policy (BRP)

```
$\pi_1 = 1$ and $\pi_2 = 2$
for $t = 3$ to $n$ do
  if $\max_k \hat{\mu}_{k,t} > 0$ then
    then $\pi_t = \text{argmax}_k \hat{\mu}_{k,t}$
  else
    $\pi_t = 1$, $\pi_{t+1} = 2$
  end if
end for
```

Theorem: BRP has regret

$$R_n \le \Delta + \frac{16}{\Delta} \ .$$

Proof.

$$\mathbb{P}(\pi_t = 2) = \mathbb{P}(\hat{\mu}_{2,t} > 0, \pi_t = 2) + \mathbb{P}(\hat{\mu}_{2,t} \le 0, \pi_t = 2)$$

Note that

$$\begin{aligned}
\sum_{t=3}^{n} \mathbb{P}(\hat{\mu}_{2,t} > 0, \pi_t = 2) &\le \mathbb{E} \sum_{t=3}^{n} \mathbb{1}(\hat{\mu}_{2,t} > 0, \pi_t = 2) \\
&\le \mathbb{E} \sum_{t=3}^{n} \mathbb{1}(\hat{\mu}_{2,t} - \mu_2 > 0, \pi_t = 2) \\
&\le \sum_{s=1}^{\infty} e^{-\frac{s\Delta^2}{8}} \\
&= \frac{8}{\Delta^2} \ ,
\end{aligned}$$

where s is the counter of pulling arm 2 and the third inequality is a Chernoff bound. Similarly,

$$\begin{aligned}
\sum_{t=3}^{n} \mathbb{P}(\hat{\mu}_{2,t} \le 0, \pi_t = 2) &= \sum_{t=3}^{n} \mathbb{P}(\hat{\mu}_{1,t} \le 0, \pi_{t-1} = 1) \\
&\le \frac{8}{\Delta^2} \ ,
\end{aligned}$$

Combining these two inequality, we have

$$R_n \le \Delta(1 + \frac{16}{\Delta^2}) \ ,$$

□

References

[Lat15] Tor Lattimore, *Optimally confident UCB : Improved regret for finite-armed bandits*, Arxiv:1507.07880, 2015.