CSE599s, Spring 2014, Online Learning	Lecture 14 - 05/14/2014

Stochastic Bandits

Lecturer: Ofer Dekel — *Scribe: Matthias W. Smith*

1 Stochastic Bandits

This is a special case of the adversarial bandits covered in a previous lecture.

Stochastic Bandits Game
for rounds $1..T$
player pulls a single arm per round $\{1...d\}$
player receives reward x_i

The game bakes in a few assumptions which distinguish it from the general adversarial bandits case.

- $\nu_1, \nu_2, ..., \nu_d$ are unknown distributions, supported on $[0, 1]$, over reward
- pulling arm i for the s'th time results in reward $x_{is} \sim \nu_i$ and $(x_{i1}, x_{i2}, ...)$ are independent

Essentially, the adversary generates the following table.

	ν_1	ν_2	$\dots$	ν_d
1	x_{11}	x_{21}	$\dots$	x_{d1}
2	x_{12}	x_{22}	$\dots$	x_{d2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T	x_{1T}	x_{2T}	$\dots$	x_{dT}

2 Probability Theory

Before we dive into the stochastic bandits problem, we need to take a bit of a detour through some probability theory.

2.1 Weak Law of Large Numbers

We start by requiring that $x_1, ..., x_n$ are iid and "well-behaved" random variables. Then we can define the emprical mean as $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. If we denote the expected mean $\mu = \mathsf{E}[x]$, then $\hat{\mu}$ converges to μ in probability.

Theorem 1.

$$\forall \epsilon > 0, \lim_{n \to \infty} \mathsf{P}\left(|\hat{\mu} - \mu| \geq \epsilon\right) = 0$$

The theorem is readily present in nearly all statistical learning theory. In order to study *sample complexity*, we need a non-asymptotic version of the law of large numbers (LLN).

2.2 Markov's Inequality

Theorem 2. *If x is a non-negative r.v. and $\epsilon > 0$, then* $\mathsf{P}(x \geq \epsilon) \leq \frac{\mathsf{E}[x]}{\epsilon}$

Proof. Start with the following assertion

$$\epsilon \cdot \mathbb{I}_{\{x \geq \epsilon\}} \leq x.$$

We have two possible cases, $x \geq \epsilon$ and $x < \epsilon$. It is fairly easy to convince yourself that the statement must always be true. Now we proceed by taking the expectation of each side.

$$\epsilon \mathsf{E}[\mathbb{I}_{\{x \geq \epsilon\}}] \leq \mathsf{E}[x]$$

$$\epsilon \mathsf{P}[x \geq \epsilon] \leq \mathsf{E}[x]$$

Thus proving Markov's Inequality. □

2.3 Chebychev's Inequality

Theorem 3. *If x is r.v. with $\mu = \mathsf{E}[x] < \infty$ and $\epsilon > 0$, then $\mathsf{P}(|x - \mu|) > \epsilon) \leq \frac{Var(x)}{\epsilon^2}$.*

Proof. Define $z = (x - \mu)^2$. By definition we can write

$$\mathsf{P}(|x - \mu|) \equiv \mathsf{P}(z \geq \epsilon^2),$$

and by Markov's Inequality

$$\mathsf{P}(z \geq \epsilon^2) \leq \frac{\mathsf{E}[z]}{\epsilon^2}.$$

□

Assume without losing generality that $\mu = 0$. Noting that we could always define a new random variable $y = x - \mu$ where $\mathsf{E}[y] = 0$ now.

Theorem 4. *Assume $\mu = 0$, then $z = x^2$ and*

$$\mathsf{P}(|x| \geq \epsilon) = \mathsf{P}(z \geq \epsilon^2) \leq \frac{\mathsf{E}[x^2]}{\epsilon^2}.$$

Proof. Use the Weak *LLN* with $\mu = 0$, assuming $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $x_1 ... x_n$ are iid and $\mathsf{E}[x_i^2] < \infty$. By definition we get the first relation

$$\mathsf{E}[\hat{\mu}^2] = \mathsf{E}[(\frac{1}{n}\sum_{i=1}^{n} x_i)^2].$$

Now use the fact that x_i is a generic random variable and recast its squared sum as multiplication of two random variables to get

$$\mathsf{E}[\hat{\mu}^2] = \frac{1}{n^2}\mathsf{E}[\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j].$$

Separating terms in which $i = j$ and $i \neq j$,

$$\mathsf{E}[\hat{\mu}^2] = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{E}[x_i^2] + \sum_{i \neq j}\mathsf{E}[x_i x_j].$$

Now we can eliminate the second term by using the linearity of expectation and the fact that we have defined $\mu = 0$.

$$\mathsf{E}[x_i x_j] = \mathsf{E}[x_i]\mathsf{E}[x_j] = 0 \cdot 0$$

$$\mathsf{E}[\hat{\mu}^2] = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{E}[x_i^2] = \frac{1}{n}\mathsf{E}[x^2]$$

Plugging back into the bound given by Chebychev's Inequality we get

$$\Rightarrow \mathsf{P}(|\hat{\mu} - \mu| \geq \epsilon) \leq \frac{\mathsf{E}[x^2]}{n\epsilon^2}.$$

And as $n \to \infty$

$$\frac{\mathsf{E}[x^2]}{n\epsilon^2} \to 0$$

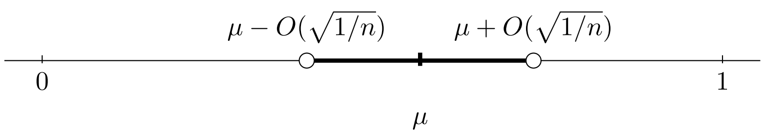
□

2.4 Confidence Intervals

These are all over the place in Machine Learning. Define $\delta = \frac{\mathsf{E}[x^2]}{n\epsilon^2}$, and similarly $\epsilon = \sqrt{\frac{\mathsf{E}[x^2]}{\delta n}}$. The term is denoted the confidence and it is defined on $\forall \delta \in [0, 1]$. With probability $(w.p.) \geq 1 - \delta$

$$|\hat{\mu} - \mu| \leq \sqrt{\frac{\mathsf{E}[x^2]}{\delta n}}$$

In diagram form the relation is shown below.



2.5 Hoeffding-Azuma Inequality

Theorem 5. *Instead of plugging in the 2nd moment ($\mathsf{E}[x^2]$) into Markov's Inequality use the "exponential moment", $z = e^{\lambda \sum_{i=0} n x_i}$, to get*

$$\mathsf{P}(\hat{\mu} > \epsilon) = \mathsf{P}(z \geq e^{n\lambda\epsilon})$$

Proof. Using Markov's Inequality on the righthand side we have

$$\mathsf{P}(z \geq e^{n\lambda\epsilon}) \leq \mathsf{E}[z]e^{-n\lambda\epsilon}.$$

We can substitute for z and turn the sum into a *pi*-product since it is in the exponent to get

$$\mathsf{E}[z]e^{-n\lambda\epsilon} = \mathsf{E}[\prod_{i=1}^{n} e^{\lambda x_i}] \cdot e^{-n\lambda\epsilon}.$$

$$\mathsf{E}[z]e^{-n\lambda\epsilon} = \prod_{i=1}^{n} \mathsf{E}[e^{\lambda x_i}] \cdot e^{-n\lambda\epsilon}$$

Using convexity and a Taylor Expansion we get derive

$$\mathsf{E}[e^{\lambda x}] \leq e^{\lambda^2/\delta}.$$

If $x \in [a, b]$ with $b - a = 1$, then $\mathsf{E}[x] = 0$.

$$\mathsf{E}[z]e^{-n\lambda\epsilon} = e^{n\lambda^2/\delta - n\lambda\epsilon}$$

$$\mathsf{P}(\hat{\mu} > \epsilon) \leq \min_{\lambda > 0} e^{n(\frac{\lambda^2}{\delta} - \lambda\epsilon)}$$

Taking the derivative to minimize we find that $\lambda = 4\epsilon$, so

$$\mathsf{P}(\hat{\mu} \geq \epsilon) \leq e^{-2n\epsilon^2}$$

and symmetrically

$$\mathsf{P}(\hat{\mu} \leq -\epsilon) \leq e^{-2n\epsilon^2}.$$

$$\Rightarrow \mathsf{P}(|\hat{\mu}| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

□

2.6 Confidence Intervals Revisited

We can defined a new confidence interval

$$\delta = 2e^{-2n\epsilon^2}.$$

and

$$\epsilon = \sqrt{\frac{\log 2/\delta}{2n}}.$$

We have shown: $\forall \delta \in [0, 1]$, $w.p. \geq 1 - \delta$

$$|\hat{\mu} - \mu| \leq \sqrt{\frac{\log 2/\delta}{2n}}$$

Compare to our previous result we now have $\sqrt{\log 1/\delta}$ versus $\sqrt{1/\delta}$. This result is heavily predicated by $x \in [a, b]$ where $b - a = 1$.

3 Stochastic Bandits Revisited

First we want to define some convenient variables: $\mu_i = \mathsf{E}_{x\sim\nu_i}[x]$ is the expected reward of arm i, $T_i(t)$ is the number of times arm i is pulled on rounds $1...t$, and $\Delta_i = \mu^* - \mu_i$.
Comments:

- the exact time that the arm was pulled doesn't matter, only $T_i(t)$ matters
- there is a "best arm", the one with the largest expetec reward $\mu^* = \max_{1\leq i\leq d} \mu_i$

We can now take our regret

$$\text{Regret} = T_{\mu^*} - \mathsf{E}[\sum_{t=1}^{T} X_{I_t, T_{I_t}(t)}]$$

and recast it as

$$\text{Regret} = T_{\mu^*} - \mathsf{E}[\sum_{t=1}^{T} \mu_{I_t}]$$

$$\text{Regret} = \sum_{i=1}^{d} \Delta_i \mathsf{E}[T_i(t)].$$

The goal for stochastic bandits is to bound $\mathsf{E}[T_i(t)]$ for all i with $\Delta_i > 0$.

3.1 Algorithm

The simplest feasible algorithm is dubbed "ϵ-first". If we know Δ, a lower bound of $\{\Delta_i : \Delta_i > 0\}$. We sample each arm $O\left(\frac{\log(1/\delta)}{\Delta^2}\right)$ times, estimate each μ_i to within $\Delta/2$, and stick to the emperical best arm henceforth.

$$\text{Regret} \leq \sum_{i=1}^{d} \Delta_i \cdot \left(\frac{\log 1/\delta}{\Delta^2}\right) + \Theta(\delta T) \sim \Theta(\log T)$$