The Multi-armed Bandit Problem

John Duchi

(Stochastic) Multi-armed bandit problems

Formal setting: there are d different means

$$\mu_1, \mu_2, \ldots, \mu_d$$

$\mu_{i^\star} \geq \mu_j$ for all $j \neq i^\star$. Proceed sequentially as follows: at round t

(1) Choose arm $A_t \in \{1, \ldots, d\}$
(2) Observe Y_{A_t} with $\mathbb{E}[Y_{A_t}] = \mu_{A_t}$

Goal: Make the expected regret

$$\mathsf{Reg}_T := \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i^\star} - \mu_{A_t}\right]$$

small

Motivation

- Two treatments for disease available
- Need to find treatment with best effect for population
- Don't want to give bad treatment to too many people

- Strong relationship with causality

Exploration vs. exploitation

- Exploration: figure out performance of different arms
- Exploitation: pull the best arm!

Tension between the two!
Idea: Some kind of confidence bounds

A few simple insights

Assume each arm i is σ^2-sub-Gaussian, i.e.

$$\mathbb{E}\left[\exp(\lambda(Y_i - \mu_i))\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

- $T_i(t) = \sum_{\tau \le t} \mathbf{1}\{A_\tau = i\}$ is count of arm i pulls at time t
- Have pretty good mean estimates

$$\widehat{\mu}_i(t) := \frac{1}{T_i(t)} \sum_{\tau \le t : A_\tau = i} Y_\tau$$

$$\mathbb{P}\left(|\widehat{\mu}_i(t) - \mu_i(t)| \ge \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{T_i(t)}}\right) \le 2\delta.$$

Upper confidence bound (UCB) algorithm

Input: Sub-gaussian parameter σ^2 and probabilities $\delta_1, \delta_2, \ldots$
Initialization: Play each arm $i = 1, \ldots, K$ ones
Repeat: play arm maximizing

$$\widehat{\mu}_i(t) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}}$$

Regret of UCB algorithm

Gaps $\Delta_i = \mu_{1^\star} - \mu_i$

$$\mathsf{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i^\star} - \mu_{A_t}\right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[T_i(T)\right]$$

Analysis goal: Show that $T_i(T)$ is small for all $i \neq i^\star$

Proposition (Arm pulls in UCB)

Assume $\delta_1 \geq \delta_2 \geq \ldots$. For all T and $i \neq i^\star$,

$$\mathbb{E}\left[T_i(T)\right] \leq \left\lceil \frac{4\sigma^2 \log \frac{1}{\delta_T}}{\Delta_i^2} \right\rceil + 2\sum_{t=2}^{T} \delta_t.$$

Showing regret bound

Assume w.l.o.g. that $i^\star = 1$

- Three problematic events

$$\mathcal{E}_1(t) : \ \widehat{\mu}_i(t) \geq \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}}$$

$$\mathcal{E}_2(t) : \ \widehat{\mu}_1(t) \leq \mu_1 - \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_1(t)}}$$

$$\mathcal{E}_3(t) : \ \Delta_i \leq 2\sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{T_i(t)}}$$

Counting the bad events

For any fixed $l \in \{1, \ldots, T\}$

$$\mathbb{E}[T_i(T)] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{A_t = i\right\}\right] \le l + \sum_{t=l+1}^{T} \mathbb{E}\left[\mathbf{1}\left\{A_t = i, T_i(t) > l\right\}\right]$$

Counting the bad events

$$\mathbb{E}[T_i(T)] \le \left\lceil \frac{4\sigma^2 \log \frac{1}{\delta_T}}{\Delta_i^2} \right\rceil + \sum_{t=l^\star+1}^{T} \mathbb{P}(A_t = i, \mathcal{E}_3(t) \text{ fails})$$

UCB Regret bound

Theorem (UCB regret)

Take $\delta_t = \frac{1}{T}$ for all t, then

$$\mathsf{Reg}_T \le O(1)\left[\frac{K\sigma^2 \log T}{\min_{i \neq i^\star} \Delta_i} + \sum_{i=1}^{K} \Delta_i\right]$$

Regret "smaller" than the number of arms

If $\Delta = \max_{i \neq i^\star} \Delta_i$ small, should not really matter...

Theorem (Alternate form of regret)

Let $\delta_t = \frac{1}{T}$ for all t, then

$$\mathsf{Reg}_T \le O(1) \cdot \sqrt{K\sigma^2 T \log T}.$$

Mirror descent for bandit problems

- $x \in \Delta_d = \{x \in \mathbb{R}_+^d : \mathbf{1}^T x = 1\}$
- At round t, draw $i \sim x_t$ (treat $x_t \in \Delta_d$ as distribution)
- Loss

$$f(x) = \langle -\mu, x \rangle = -\sum_{j=1}^{d} \mu_j x_j = \mathbb{E}_{i \sim x}[-\mu_i]$$

- Regret: let $x^\star = e_{i^\star}$, then

$$\mathrm{Reg}_T = \sum_{t=1}^{T} [f(x_t) - f(x^\star)] = \sum_{t=1}^{T} \left[\mu_{i^\star} - \mathbb{E}_{i \sim x_t}[\mu_i]\right]$$

A more careful regret bound for mirror descent

Proposition

Let $X = \Delta_d$ and play exponentiated gradient algorithm

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \left\{ \langle g_t, x \rangle + \frac{1}{\alpha} D_h(x, x_t) \right\}$$

where $h(x) = \sum_{j=1}^d x_j \log x_j$. Then

$$\sum_{t=1}^T \langle g_t, x_t - x^\star \rangle \le \frac{\log d}{\alpha} + \frac{\alpha}{2} \sum_{t=1}^T \sum_{j=1}^d x_{t,j} g_{t,j}^2$$

Applying mirror descent: EXP3

Repeat: for $t = 1, 2, \ldots$

- Choose action $A_t = i$ with probability $x_{t,i}$
- Receive loss $Y_i(t)$ and set

$$g_{t,j} := \begin{cases} -Y_j(t)/x_{t,j} & \text{if } A_t = j \\ 0 & \text{otherwise} \end{cases}$$

- Update for $i = 1, \ldots, d$

$$x_{t+1,i} = \frac{\exp(-\alpha g_{t,i})}{\sum_{j=1}^d \exp(-\alpha g_{t,j})}$$

Regret bounds with mirror descent

Theorem

Assume $Y_i \geq 0$ *and* $\mathbb{E}[Y_i^2] \leq \sigma^2$*. The expected regret of the exponentiation weights (EXP3) algorithm is*

$$\mathsf{Reg}_T = \sum_{t=1}^{T} \mathbb{E}\left[\mu_{i^\star} - \mu_{A_t}\right] \leq \frac{\log d}{\alpha} + \frac{\alpha}{2}\sigma^2 KT.$$

Proof of regret bound

Have

$$\mathbb{E}[\mu_{A_t} \mid x_t] = \sum_{j=1}^{d} \mu_j x_{t,j} = \mathbb{E}[\langle g_t, x_t \rangle \mid x_t]$$

so

$$\mathsf{Reg}_T = \sum_{t=1}^{T} \mathbb{E}\left[\langle g_t, x_t - x^\star \rangle\right]$$

Extensions and other approaches

1. Thompson sampling—putting a Bayesian prior on the means, draw random mean according to posterior belief, play best arm according to random mean
2. Adversarial bandits—sequence $f_t : X \to \mathbb{R}$ chosen adversarially (arbitrarily), observe only $f_t(x_t)$
3. Contextual bandits—some side information availabe
4. "Batched" bandits—only a small number of rounds, but in each round, a large sample is available (e.g. in FDA trials)

Reading and bibliography

1. N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006

2. S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012

3. P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a

4. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b

5. J. C. Duchi. Stats311/EE377: Information theory and statistics. Course at Stanford University, Fall 2015. URL `http://web.stanford.edu/class/stats311`