July 29, 2020

Stochastic Bandits with finitely many arms and Algorithms (1)

Intro + Notation :

- game b/w learner and environment
- in each round $t \in [n]$, learner chooses action $A_t \in \mathcal{A}$, environment reveals reward $X_t \in \mathbb{R}$
- history : $H_{t-1} \triangleq (A_1, X_1, \ldots, A_{t-1}, X_{t-1})$
- policy : mapping from histories to actions
- environment : mapping from actions to rewards

Objective : choose actions to maximize cumulative reward, $\sum_{t=1}^{n} X_t$

Regret (Informal) : difference b/w total expected reward & total expected reward collected. evaluated w.r.t policy π.

regret relative to set of Π is max over all regrets, $\pi \in \Pi$

- **Example** : (Stochastic Bernoulli Bandit)

let $A = \{1, \dots, k\}$, $X_t \in \{0,1\}$ and

there exists $\mu \in [0,1]^k$ s.t

$$\Pr(X_t = 1 \mid A_t = a) = \mu_a.$$

if $\vec{\mu}$ were known, optimal policy is to

play fixed action $a^* = \arg\max_{a \in A} \mu_a$.

$$R_n = n \max_{a \in A} \mu_a - E\left[\sum_{t=1}^{n} X_t\right]$$

Question : how does R_n scale with n?

Answer : "good learner" achieves sub-linear regret, ie, $R_n = o(n)$ $\left[\lim_{n \to \infty} \frac{R_n}{n} \to 0\right]$

Q2: under wat circumstances is $R_n \in O(\sqrt{n})$ or $R_n \in O(\log n)$ etc

A2: In above example, $R_n = \Omega(\sqrt{n})$ and there exist policies for which $R_n = O(\sqrt{n})$

Stochastic Bandits

- collection of distributions, $\nu = (P_a \mid a \in \mathcal{A})$
- environment samples reward $X_t \in \mathbb{R}$ from P_{A_t}, reveals to learner
- horizon : # rounds, n, generally finite

Some constraints

[...uped] ⓐ $P(X_t \mid A_1, X_1, \ldots, A_{t-1}, X_{t-1}) = P_{A_t}$

[causality] ⓑ conditional law of action A_t given H_{t-1} is $\Pi_t(\cdot \mid H_{t-1})$ where $\Pi_1, \ldots$ is sequence of probabilities that characterize learner

- use $\mathcal{E}$ to denote environment class
- in sequel use $\mathcal{E}^k_{SG}(1)$ ie, all P_a's are 1- subGaussian,

$$\Pr_{X \sim P_a}(|X| > \epsilon) \leqslant \exp\left(-\frac{\epsilon^2}{2}\right)$$

Algorithms

1. Explore-then-commit (ETC)

- characterized by $\underline{m} \in \mathbb{N}$:

  # times each arm is explored.

let $\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{1}_{\{A_s = i\}} X_s$

where $T_i(t) = \sum_{s=1}^{t} \mathbb{1}_{\{A_s = i\}}$

Algo 1 : ETC

- input m
- in round t, choose action

$$A_t = \begin{cases} (t \bmod k)+1, & t \leq mk \\ \arg\max_i \hat{\mu}_i(mk), & t > mk. \end{cases}$$

Def : μ_i - mean reward for action i

$\Delta_i = \mu^* - \mu_i$, sub optimality gap

Thm : Consider ETC, $\mathcal{E}^k_{sg}(1)$, $1 \leq m \leq n/k$

$$R_n \leq \underbrace{m \sum_{i=1}^{k} \Delta_i}_{\text{explore}} + \underbrace{(n-mk) \sum_{i=1}^{k} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)}_{\text{exploit}}$$

how to pick m?

- let $k=2$, $\mu^* = \mu_1$, $\Delta_2 = \Delta$, $\Delta_1 = 0$

$$R_n \leq m\Delta + (n-2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right)$$

$$\leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right)$$

$\Rightarrow$ for large enough n, r.h.s is minimized if

$$m = \max\left\{1, \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right)\right\rceil\right\}$$

Implications

- w/ above value of m, $R_n \leq \Delta + C\sqrt{n}$
- if n, Δ were unkown (generally true), then $R_n = O(n^{2/3})$

:

~~Algo 2~~

2. Upper Confidence Bound (UCB)

- drawbacks of ETC: – needs Δ;
  - $k > 2$ is "hard" (?)
  - depends on n

- based on "Optimism in the face of uncertainty"

[intuition]

- based on observed data, to each arm assign UCB

s.t. w.h.p UCB $\geq$ mean

- if UCB_{opt} is an overestimate, a different arm is played only if $UCB_i > UCB_{opt} > \mu_{opt}$
  - but this cannot happen "too many times" since after enough rounds, $UCB_i < UCB_{opt}$

Def^n: let $\{X_t\}_{t=1}^{n}$ iid 1-subGaussian, $E[X_t] = \mu$.

let $\hat{\mu} = \frac{1}{n}\sum X_t$ sample mean, then,

$$Pr\left(\mu > \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{n}} \right) \leq \delta \quad \forall\ \delta \in (0,1)$$

in context of algo, at round t, learner has seen $T_i(t-1)$ samples of arm i, sample mean $\hat{\mu}_i(t-1)$, then,

$$UCB_i(t-1, \delta) = \begin{cases} \infty & \text{if } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}, & \text{ow} \end{cases}$$

↳ actually an r.v, but generally ok

Algo 2: UCB

- input: k, δ
- for $t = 1, \ldots, n$

  pick $A_t = \arg\max_i UCB_i(t-1, \delta)$

  observe reward X_t, update UCB_i

[more intuition]

- algo should explore more ~~arms~~ if
  (a) $\hat{\mu}_i(t-1)$ is large
  (b) not well explored ie $T_i(t-1)$ is small

- assume at round t, arm 1 played >> others,
- hope 1 - optimal arm, $\hat{\mu}_1(t-1) \approx \mu_1$
- ensure that i-th arm worse than 1 if

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} \leq \mu_1 < \hat{\mu}_1(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_1(t-1)}}$$

How to pick δ?

- next lectures, but need $\delta < 1/n$

Thm: consider UCB. for any n, if $\delta = 1/n^2$, then

$$R_n \leq 3 \sum_{i=1}^{k} \Delta_i + \sum_{i|\Delta_i>0} \frac{16 \log n}{\Delta_i}$$

Corollary: if $\delta = 1/n^2$, then for any $v \in \mathcal{E}_{SG}^k(1)$, UCB regret,

$$R_n \leq 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k} \Delta_i$$

$\Rightarrow$ No algo can do better than

$$R_n = O(\sqrt{nk}) \text{ over } v \in \mathcal{E}_{SG}^k(1).$$