July 30, 2020

Stochastic Bandits w/ finitely many arms and Algorithms (2)

- UCB : Asymptotic optimality

- input : k
- choose each arm once
- then,

$$A_t = \arg\max_i \hat{\mu}_i(t-1) + \sqrt{\frac{2 \log f(t)}{T_i(t-1)}}$$

where $f(t) = 1 + t \log^2 t$

Thm : $R_n \leq C \sum_{i \mid \Delta_i > 0} \left(\Delta_i + \frac{\log n}{\Delta_i} \right)$

using same idea as before,

$$R_n \leq C \sum_i \Delta_i + 2\sqrt{Ck \log n}$$

more complex

$$R_n \leq \sum_{i \mid \Delta_i > 0} \inf_{\epsilon \in [0, \Delta_i)} \Delta_i \left(1 + \frac{5}{\epsilon^2} + \frac{2\left(\log f(n) + \sqrt{\pi \log f(n)} + 1\right)}{(\Delta_i - \epsilon)^2} \right)$$

$\epsilon = \frac{\Delta_i}{2} \Rightarrow$

$$R_n \leq \sum_{i \mid \Delta_i > 0} \left(\Delta_i + \frac{1}{\Delta_i} \left(8 \log f(n) + 8\sqrt{\pi \log f(n)} + 28 \right) \right)$$

→ does not depend on knowledge of n

- Minimax optimality in Stochastic case (MOSS)
[Audibert, Bubeck '09]

→ cannot be improved except wrt constants-

- input : n, k
- Choose each arm once
- Then,

$$A_t = \arg\max_i \hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+\left(\frac{n}{k\, T_i(t-1)}\right)}$$

where $\log^+(x) = \log(\max\{1, x\})$

Thm: for any 1-sub Gaussian bandit, Moss satisfies

$$R_n \leq 39\sqrt{nk} + \sum_i \Delta_i$$

Problems with Moss

- Suboptimality wrt UCB

ex: $n = k^3$, $\nu \sim N(0,1)$, $\mu_1 = 0$, $\mu_2 = -\sqrt{\frac{k}{n}}$, $\mu_i = -1$, $i > 2$

$$R_n^{UCB} = O(k \log k)$$

$$R_n^{Moss} = \Omega(k^2)$$

Adversarial Bandits w/ finitely many arms and Algorithms

- no assumptions on how rewards are generated
  - ⇒ environment ≡ adversary
    - ability to examine algo, choose rewards accordingly.

- A k-armed adversarial bandit:

arbitrary sequence $(x_t)_{t=1}^{n}$, $x_t \in [0,1]^k$

- in each round, $t = 1, \ldots, n$, learner chooses a distribution over actions $P_t \in \mathcal{P}_{k-1}$
- then action $A_t \in [k]$ sampled from P_t, recieves reward $x_{tA_t} = X_t$
- policy, $\pi : \left([k] \times [0,1]\right)^* \to \mathcal{P}_{k-1}$

expected regret: $R_n(\pi, x) = \max_{i \in [k]} \sum_{t=1}^{n} x_{ti} - E\left[\sum_t x_{tA_t}\right]$

worst case: $R_n^*(\pi) = \sup_{x \in [0,1]^{n \times k}} R_n(\pi, x)$

Q: can one achieve $R_n^*(\Pi) = o(n)$?

A: for deterministic, $R_n^*(\Pi) \geq n(1-1/k)$

Remark:
(A): deterministic strategies lead to sub-optimal regret for adversarial case
(B) will adv. bandit strategy have small "expected regret" in stochastic setting?

- let Π be an adversarial bandit policy and $\nu = (\nu_1, \ldots, \nu_k)$ be stochastic bandit with $\text{supp}(\nu_i) = [0,1]$
let $X_{ti} \sim \nu_i \quad \forall\, i \in [k],\ t \in [n]$, indep

$$R_n(\Pi, \nu) = \max_{i \in [k]} E\left[\sum_{t=1}^{n} (X_{ti} - X_{tA_t})\right]$$

$$\leq E\left[\max_{i \in [k]} \sum_{t=1}^{n} (X_{ti} - X_{tA_t})\right] \quad \text{Jensen}$$

$$= E[R_n(\Pi, X)] \leq R_n^*(\Pi)$$

- Importance-weighted estimators (IWE)

- need to estimate reward of unplayed arms

let $P_{ti} = \Pr(A_t = i \mid A_1, X_1, \ldots, A_{t-1}, X_{t-1})$

assume $P_{ti} > 0$ almost surely, then, IWE of x_i

$$\hat{X}_{ti} = \frac{\mathbb{1}\{A_t = i\} X_t}{P_{ti}}$$

Let $E_t[\cdot] \triangleq E[\cdot \mid A_1, X_1, \ldots, A_{t-1}, X_{t-1}]$

Then $E_{t-1}[\hat{X}_{ti}] = x_{ti}$

→ Let $V_{t-1}[U] = E_{t-1}\left[(U - E_{t-1}[U])^2\right]$

$$V_{t-1}[\hat{X}_{ti}] = \frac{x_{ti}^2 (1 - P_{ti})}{P_{ti}}$$

- (Exp3) exponential-weight algorithm for exploration & exploitation

Let $\hat{S}_{ti} = \sum_{s=1}^{t} \hat{X}_{si}$

for some $\eta > 0$,

$$P_{ti} = \frac{\exp(\eta \hat{S}_{t-1,i})}{\sum_{j=1}^{k} \exp(\eta \hat{S}_{t-1,j})}$$

learning rate, $f(n,k)$

Exp3:

- input n, k, η
- let $\hat{S}_{0i} = 0 \quad \forall i$
- for $t = 1 \ldots n$,
  - compute P_{ti}
  - sample $A_t \sim P_t$, observe reward X_t
  - calculate $\hat{S}_{ti} = \hat{S}_{t-1,i} + 1 - \frac{\mathbb{1}_{\{A_t = i\}}(1 - X_t)}{P_{ti}}$

Thm: Let $x \in [0,1]^{n \times k}$, π be policy of Exp3 with $\eta = \sqrt{(\log k)/nk}$, then,

$$R_n(\pi, x) \leq 2\sqrt{nk \log k}$$