Bandits

Stochastic - play action a_t (choice/decision)

observe reward $r_t = r(a_t)$

$E[r_t] = \mu(a_t)$ ← $r_t = \mu(a_t) + \eta_t$

r, μ - bounded $[0,1]$ or subgaussian

concentration $\hat{\mu} \to \mu$

hoeffdings / bernstein - finite, Lipschitz } μ

" martingale - linear, GP

$wp \geq 1-\delta$ $R(T) = \sum_{t=1}^{T} (\mu(a^*) - \mu(a_t)) \leq \varepsilon(\delta, T, A)$

↳ action space

$E[R(T)] \asymp O(\sqrt{KT\log T})$ K - finite

$\asymp O(d\sqrt{T\log T}),\ T^{\frac{d+1}{d+2}}(\log T)^{\frac{1}{d+2}}$ d - linear, Lipschitz

d, K, ν - dim & kernel hyperparameters

Algorithms: Explore then exploit

ε-greedy

UCB (Upper Confidence Bound) $|\mu - \hat{\mu}| \leq \varepsilon$

Thompson

- Cumulative Regret $\sum_{t=1}^{T} \mu(a^*) - \mu(a_t) = R(T)$
- Simple Regret $\mu(a^*) - \mu(\hat{a}_T) = S(T)$ $\hat{a}_T$ - recommended action after T rounds

Can get $E[S(T)] = \frac{E[R(T)]}{T}$ at time T let $\hat{a}_T = a$ wp $\frac{n_a(T)}{T}$

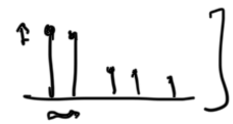
$E[R(T)] = E[\sum_t \mu(a^*) - \mu(a_t)] = \sum_a (\mu(a^*) - \mu(a)) E[n_a(T)]$

$E[S(T)] = E[\mu(a^*) - \mu(\hat{a}_T)] = \sum_a (\mu(a^*) - \mu(a)) E[\frac{n_a(T)}{T}]$

Conversely, $E[S(T)] \geq \min_{i \neq i^*} \Delta_i \, e^{-c E[R(T)]}$

$\underbrace{\mu_i - \mu_{i}}$

for all Bernoulli bandits.



- Best arm identification $\quad \underline{P(\hat{a}_T \neq a^*)} = e(T)$

$\rightarrow \quad \underbrace{\min_{i \neq i^*} \Delta_i} \cdot e(T) \leq E[S(T)] \leq e(T) \qquad$ if $\mu \in [0,1]$

Fixed confidence

given $\delta \qquad \min_{\tau} E[\tau] \quad$ s.t. $P(\hat{a}_\tau \neq a^*) \geq 1-\delta$

↳ stopping time (random) "anytime" valid

Fixed budget

given T, $\quad \min P(\hat{a}_T \neq a^*)$

- Reward estimation $\quad \hat{\mu}(i)$
- Average treatment effect. (ATE) $\qquad K=2$

$$ATE = \mu(1) - \mu(0)$$

Randomized control trials

Fixed allocation $\pi = P(a=1)$

1) Inverse Probability Weighting (IPW)
Horvitz - Thompson estimator

$$\Rightarrow \hat{\mu}(1) = \frac{1}{T}\sum_{t=1}^{T} r_t \frac{\mathbb{1}_{a_t=1}}{\pi} \qquad \hat{\mu}(0) = \frac{1}{T}\sum_{t=1}^{T} r_t \frac{\mathbb{1}_{a_t=0}}{1-\pi}$$

$$\rightarrow \widehat{ATE} = \frac{1}{T}\sum_{t=1}^{T} r_t \left(\frac{\mathbb{1}_{a_t=1}}{\pi} - \frac{\mathbb{1}_{a_t=0}}{1-\pi} \right)$$

$E[\hat{\mu}(i)] = \mu(i)$ unbiased.

$$\text{var} \quad \frac{1}{T}\left(\frac{\overset{m_1^2}{E[r^2(1)]}}{\pi} + \frac{\overset{m_0^2}{E[r^2(0)]}}{1-\pi} - \Delta^2 \right) \qquad \Delta = \mu(1) - \mu(0)$$

$\min$ var ? $\equiv \min_{\pi}$ MSE

π "

Neyman policy $\pi_{Ney} = \frac{m_1}{m_1+m_0}$ $= \frac{1}{2}$ iff $m_1 = m_0$

$\left.\frac{\hat{m}_1}{\hat{m}_1+\hat{m}_0}\right\} \leftarrow$ clipped $0 < \hat{m}_0, \hat{m}_1 < 1$

(2) Direct method

$\hat{\mu} \leftarrow \min_{f \in F} (f(i) - \mu(i))^2$ - misspecification

(3) AIPW (unbiased even if $\hat{\mu}$ is biased / misspecified)

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1_{a_t=1}(r_t - \hat{\mu}_t(1))}{\pi} + \hat{\mu}_T(1) - \frac{1_{a_t=0}}{1-\pi}(r_t - \hat{\mu}_t(0)) - \hat{\mu}_T(0)\right)$$

$E[\hat{\mu}] \neq \mu$

unbiased

var $\frac{\sigma^2(1)}{\pi} + \frac{\sigma^2(0)}{1-\pi} + E[(\mu(1)-\mu(0) - \Delta)^2]$

$\pi^*_{AIPW} = \frac{\sigma(1)}{\sigma(1)+\sigma(0)}$

how to estimate in finite sample setting.
dependence.

Take away: so far considered bounded or equal var actions
need to account for different variances (std dev.

Adversarial setting - play action a_t
observe reward $r_t(a_t)$ bandit / $\{r_t(a)\}_{a \in A}$ full feedback

reward can change at ~~$r(a_t)$~~ each step t for any a.

$R(T) = \sum_t [r_t(a^*) - r_t(a_t)]$ $a^* = \arg\min_a \sum_t r_t(a)$

Pseudo regret $a^* = \arg\min_a \sum_t E[r_t(a)]$

Algos.: Weighted Maj Vote (rewards binary)] full feedback

Hedge

Exp(4) - bandit

Sample arms / experts acc to $\omega_t(a) / \omega_t(e)$

multiplicative / exp update $\omega_{t+1}(a) = \omega_t(a)(1-\varepsilon)^{c_t(a)} = e^{-\varepsilon c_t(a)}$

$e^{-\varepsilon} \simeq 1 + \varepsilon + \frac{\varepsilon^2}{2} + \ldots$

$$E[R(T)] \lesssim \sqrt{\underbrace{\sum_t c_t(a^*)}_{KT} \ln K}$$

or e^*, $K \to N$

Contextual bandits - middle ground.

env choose context x_t visible to player

play action a_t

observe reward $r_t = r(a_t, x_t)$ $E[r_t] = \mu(a_t | x_t)$

$$R(T) = \sum_{t=1}^{T} [\underbrace{\max_a \mu(a | x_t)}_{\pi^*(x_t)} - \mu(a_t | x_t)]$$

$$E[R(T)] = O(\sqrt{|A||X| T \log T})$$

$$= O(d \sqrt{T \log T})$$

$A \times X \to R^d$

control, don't control