Finite armed bandits $\leq$ UCB, successive elimination — $\sqrt{KT}$
K

Continuous armed bandits / Structured bandits
$S \subseteq \mathbb{R}^d$ - Lipschitz $\leq$ reduction to finite arm — $T^{\frac{d+1}{d+2}}$
zooming (adaptive binning) — $T^{\frac{d'+1}{d'+2}}$

- Linear — UCB — $d\sqrt{T}$
θ $\quad X_t \leftarrow \arg\max_x UCB_t(x) \rightarrow \max_{\theta \in C_t} \langle x, \theta \rangle$

$C_t = \{\theta : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}} \leq \sqrt{\beta_t}\}$

How to leverage prior information?
- Finite
- Continuous

$\begin{pmatrix} \mu(1) \\ \vdots \\ \mu(K) \end{pmatrix} = \mu$ - mean reward vector $\quad \mu \sim P$ (prior distr^n) $\quad \mu \in [0,1]^K$

Bayesian Regret $\quad E_{\mu \sim P}\left[\mu^* T - \sum_{t=1}^{T} \mu(a_t) \right]$

for simplicity,
- Single parameter reward family $\quad \mu \sim$ Bernoulli(θ), unit-var Gaussian, zero-mean Gaussian
- finite support for reward & $P(\mu)$
- a^* is unique for each μ.

Fact: History $H_t = \{a_1, r_1, \ldots a_t, r_t\}$
Posterior distribution $P_H(\mu) = P(\mu \mid H_t = H)$ doesn't depend on algorithm given history.

Thompson Sampling

Observe $H_{t-1} = H$

Sample mean vector from posterior $\mu_t \sim P_H(\mu)$

Choose best arm according to μ_t, $a_t = \arg\max_a \mu_t(a)$

Update $H_t \leftarrow (a_t, r_t) \cup H_{t-1}$

If prior is correct, $O(\sqrt{KT\log T})$

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

Thm: Bayesian regret

Proof: At round t, $BR_t = E_{\mu\sim P}[\mu(a^*) - \mu(a_t)]$

$= E_{\mu\sim P}[E_{H_t}[\mu(a^*) - \mu(a_t) \mid H_t]]$

$= E_{H_t}[E_{\mu\sim P}[\mu(a^*) - \mu(a_t) \mid H_t]]$ ★

Claim: a_t and a^* have same distribution given H_t.

$a_t = \arg\max_a \mu_t(a)$, $\mu_t \sim Pr(\mu \mid H_{t-1}) \propto P(\mu)\, Pr(H_{t-1} \mid \mu)$

a_t & a^* have same dist given H_{t-1}

$Pr(a_t = a \mid H_{t-1}) = Pr(\mu_t(a) \geq \mu_t(a') \;\forall a' \mid H_{t-1})$ $\quad \mu_t \sim Pr(\mu \mid H_{t-1})$

$= Pr(\mu(a) \geq \mu(a') \;\forall a' \mid H_{t-1})$

$= Pr(a^* = a \mid H_{t-1})$

★ $BR_t = E_{H_{t-1}}\{E_{\mu\sim P}[\mu(a^*) - UCB(a^*, H_{t-1}) + UCB(a_t, H_{t-1}) - \mu(a_t) \mid H_{t-1}]$

$\because E[UCB(a^*, H_{t-1}) \mid H_{t-1}] = E[UCB(a_t, H_{t-1}) \mid H_{t-1}]$

as a^*, a_t have same dist given H_{t-1}

$= E[\mu(a^*) - UCB(a^*, H_t)] + E[UCB(a_t, H_t) - \mu(a_t)]$

≤ 0

a^*, a are random

whp $\quad \mu(a^*) - UCB(a^*, H_{t-1}) \le 0$

$$UCB(a_t, H_{t-1}) - \mu(a_t) = \sum_a 1_{a_t=a} \left(UCB(a, H_{t-1}) - \mu(a)\right)$$

$$\le \sum_a 1_{a_t=a} \sqrt{\frac{\log T}{n_{t-1}(a)}}$$

$$BR \le E\left[\sum_t \sum_a 1_{a_t=a} \sqrt{\frac{\log T}{n_{t-1}(a)}}\right]$$

$$= E\left[\sum_a \sum_t 1_{a_t=a} \sqrt{\frac{\log T}{n_{t-1}(a)}}\right]$$

$$= E\left[\sum_a \sum_{j=1}^{n_T(a)} \frac{1}{\sqrt{j}} \sqrt{\log T}\right]$$

$$= O\left(E\left[\sum_a 1 \cdot \sqrt{n_T(a)}\sqrt{\log T}\right]\right)$$

$$\le O\left(\sqrt{\log T}\sqrt{\underbrace{\sum_a 1}_{K} \cdot \underbrace{\sum_a n_T(a)}_{T}}\right) \quad \text{whp}.$$

misspecified ε $\quad$ (standard cum regret) $+ \varepsilon T$

(standard cum regret) = sublinear $\quad$ εT = linear

Note: Can be improved to $O\left(\sqrt{KT \cdot \text{entropy}(a^*)}\right)$

$\sqrt{KT \log K}$ $\quad$ $a^* = \arg\max_a P(\mu)$

K – information gain

Continuous action space

$$\mu(a) \sim GP(m, K)$$

m: mean function; K: covariance function $= K(a, a')$

$$\begin{pmatrix} \mu(1) \\ \vdots \\ \mu(K) \end{pmatrix} = \mu$$

Posterior :

$$m_t(x) = m_{t-1}(x) + K(x, H_{t-1}) \, K^{-1}(H_{t-1}, H_{t-1}) \left(\begin{bmatrix} r_1 \\ \vdots \\ r_{t-1} \end{bmatrix} - m(H_{t-1}) \right)$$

$$K_t(x,x) = K(H_{t-1}, H_{t-1}) - K(x, H_{t-1}) \, K^{-1}(H_{t-1}, H_{t-1}) \, K(H_{t-1}, x)$$

μ, Σ

$$\mu(x) \to \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$\mu(H) \to$

$\mu(x|H)$ $\quad \Sigma_{x|H}$

$\mu_{1|2}$ $\quad \Sigma_{1|2}$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

$$\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (r_2 - \mu_2)$$

observed value