LEC020 MAB II (Theory)

VG441 SS2021

Cong Shi
Industrial & Operations Engineering
University of Michigan

Coin Tossing Problem (Beta-Bernoulli Bandit)

1. Infinite horizon: $1, 2, \cdots$
2. Independent coins: $1, \cdots, K$
3. $\mathbb{P}(\text{head})$: $\theta_1, \cdots, \theta_K \in [0, 1]$
4. Action (index of the coin tossed at time t): $x_t \in \{1, \cdots, K\}$
5. Outcome of the coin tossed at time t: $y_t \in \{0, 1\}$ (head=1, tail=0)
6. Reward at time t: y_t
7. Time discount: $\gamma \in (0, 1)$

Bayes' Rule on Belief Update

$$
\begin{aligned}
f_{t+1}\left(\hat{\theta}\right) &= \frac{f_t\left(\hat{\theta}\right)\mathbb{P}\left(y_{t+1}|\theta=\hat{\theta}\right)}{\mathbb{P}\left(y_{t+1}|f_t\right)} \\
&= \begin{cases} \frac{f_t\left(\hat{\theta}\right)\hat{\theta}}{\int_{\theta'=0}^{1} f_t(\theta')\theta' d\theta'} & \text{if } y_{t+1}=1 \\ \frac{f_t\left(\hat{\theta}\right)\left(1-\hat{\theta}\right)}{\int_{\theta'=0}^{1} f_t(\theta')(1-\theta')d\theta'} & \text{if } y_{t+1}=0 \end{cases}
\end{aligned}
$$

Note that the normalizing constant is given by

$$
\mathbb{P}\left(y_{t+1}|f_t\right) = \begin{cases} \int_{\theta'=0}^{1} \theta' f_t\left(\theta'\right) d\theta' & \text{if } y_{t+1}=1 \\ \int_{\theta'=0}^{1} \left(1-\theta'\right) f_t\left(\theta'\right)\theta' d\theta' & \text{if } y_{t+1}=0 \end{cases}
$$

Beta (Conjugate Prior for Bernoulli LH)

$Beta(\alpha, \beta)$ has the p.d.f.

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

If $f_t \sim Beta(\alpha, \beta)$, then

$$f_{t+1} \sim Beta(\alpha + y_{t+1}, \beta + 1 - y_{t+1})$$

Note that the normalizing constant is given by

$$\mathbb{P}(y_{t+1}|f_t) = \begin{cases} \frac{\alpha}{\alpha+\beta} & \text{if } y_{t+1} = 1 \\ \frac{\beta}{\alpha+\beta} & \text{if } y_{t+1} = 0 \end{cases}$$

Beta-Bernoulli Bandit (as DP)

Belief on θ_k at the end of period t: $Beta\left(\alpha_k^t, \beta_k^t\right)$
Objective:

$$\max \ \mathsf{E}\left[\sum_{t=1}^{\infty} \gamma^t y_t \middle| \alpha_k^0, \beta_k^0, \forall k\right]$$

<u>DP formulation</u>

1. State: $s_t = (\alpha_k^t, \beta_k^t, \forall k) \in \mathbb{N}^{2K}$
2. Action: $x_t \in \{1, \cdots, K\}$
3. Outcome: $y_t \in \{0, 1\}$
4. Transition: $\alpha_{x_t}^{t+1} = \alpha_{x_t}^t + y_t$, $\beta_{x_t}^{t+1} = \beta_{x_t}^t + 1 - y_t$

Bellman equation:

$$J(s) = \max_{k \in \{1, \cdots, K\}} \frac{\alpha_k}{\alpha_k + \beta_k} + \gamma \frac{\alpha_k}{\alpha_k + \beta_k} J\left(s^{\alpha_k + 1}\right) + \gamma \frac{\beta_k}{\alpha_k + \beta_k} J\left(s^{\beta_k + 1}\right)$$

Optimal Policy: Gittin's Index Theorem

Define

$$J_M(\alpha,\beta) = \max\left\{M, \frac{\alpha}{\alpha+\beta} + \gamma\frac{\alpha}{\alpha+\beta}J_M(\alpha+1,\beta) + \gamma\frac{\beta}{\alpha+\beta}J_M(\alpha,\beta+1)\right\}$$

Gittin's index is defined as

$$M^*(\alpha,\beta) = \min\{M : J_M(\alpha,\beta) = M\}$$

Optimal policy:

$$x_t^* \in \arg\max_k M^*\left(\alpha_k^{t-1}, \beta_k^{t-1}\right)$$

Approximation Algorithms for MAB

1. Unknown parameters: θ
2. Finite horizon: $1, 2, \cdots, T$
3. Action at time t: $x_t \in \mathcal{X}$
4. Outcome at time t: $y_t \sim q_\theta\left(\cdot|x_t\right)$
5. Reward at time t: $r_t = R\left(y_t\right)$

Regret:

Define

$$g_\theta(x) \triangleq \mathsf{E}\left[R(y)|x\right] = \int R(y) dq_\theta(y|x).$$

Define

$$x^* \in \arg\max_{x \in \mathcal{X}} g_\theta(x)$$

In the setting that the decision maker does not know θ, under any policy π,

$$\text{Regret}\left(T, \pi, \theta\right) = T g_\theta\left(x^*\right) - \mathsf{E}\left[\sum_{t=1}^{T} g_\theta\left(x_t^\pi\right)\right].$$

Upper Confidence Bound (UCB) Algorithm

UCB algorithm:

1. At each time t, define an upper confidence expected reward under each action x, $U_t(x)$, where $U_t(\cdot)$ may depend on the history $\{(x_s, y_s) : s = 1, \cdots, t-1\}$.
2. Apply action:
$$x_t^{\mathrm{UCB}} \in \arg\max_{x \in \mathcal{X}} U_t(x).$$
3. Observe y_t^{UCB}.

Suppose $|\mathcal{X}| = K < \infty$. At time $t \in \{1, \cdots, \min\{K, T\}\}$, the decision maker takes the tth action. At time $t \in \{\min\{K, T\} + 1, \cdots, T\}$, the decision maker applies the UCB algorithm with

$$U_t(x) \triangleq \min\left\{\hat{\mu}_{t-1}(x) + \beta\sqrt{\frac{\log T}{N_{t-1}(x)}}, 1\right\},$$

where

$$N_{t-1}(x) = \sum_{s=1}^{t-1} \mathbf{1}\{x_s = x\}, \quad \hat{\mu}_{t-1}(x) = \frac{\sum_{s=1}^{t-1} \mathbf{1}\{x_s = x\} r_s}{N_{t-1}(x)}.$$

$$\mathrm{Regret}\left(T, \pi^{\mathrm{UCB}}, \theta\right) \quad \leq \quad \min\{K, T\} + 2\sqrt{T} + 2\sqrt{KT\log T}.$$

Analysis of UCB

Our goal is to prove

$$\text{Regret}\left(T, \pi^{\text{UCB}}, \theta\right) \quad \leq \quad \min\{K, T\} + 2\sqrt{T} + 2\sqrt{KT \log T}.$$

Consider the first scenario that $T \leq K$. We have

$$\begin{aligned} \text{Regret}\left(T, \pi^{\text{UCB}}, \theta\right) &\leq T g_\theta\left(x^*\right) \\ &\leq T. \end{aligned}$$

The first inequality follows from the property that $r_t \in [0, 1]$.

Analysis of UCB

Consider the second scenario $T > K$. We define a lower confidence bound

$$L_t(x) \triangleq \max\left\{\hat{\mu}_{t-1}(x) - \beta\sqrt{\frac{\log T}{N_{t-1}(x)}}, 0\right\}.$$

$$\begin{aligned}
\text{Regret}\left(T, \pi^{\text{UCB}}, \theta\right) &= T g_\theta\left(x^*\right) - \mathsf{E}\left[\sum_{t=1}^{T} g_\theta\left(x_t^{\text{UCB}}\right)\right] = \sum_{t=1}^{T} \mathsf{E}\left[g_\theta\left(x^*\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right] \\
&= \sum_{t=1}^{K} \mathsf{E}\left[g_\theta\left(x^*\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right] + \sum_{t=K+1}^{T} \mathsf{E}\left[g_\theta\left(x^*\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right] \\
&\leq K + \sum_{t=K+1}^{T} \mathsf{E}\left[g_\theta\left(x^*\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right] \\
&= K + \sum_{t=K+1}^{T} \mathsf{E}\left[g_\theta\left(x^*\right) - U_t\left(x_t^{\text{UCB}}\right) + U_t\left(x_t^{\text{UCB}}\right) - L_t\left(x_t^{\text{UCB}}\right) + L_t\left(x_t^{\text{UCB}}\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right] \\
&\leq K + \sum_{t=K+1}^{T} \mathsf{E}\left[g_\theta\left(x^*\right) - U_t\left(x^*\right) + U_t\left(x_t^{\text{UCB}}\right) - L_t\left(x_t^{\text{UCB}}\right) + L_t\left(x_t^{\text{UCB}}\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right] \\
&= K + \underbrace{\sum_{t=K+1}^{T} \mathsf{E}\left[g_\theta\left(x^*\right) - U_t\left(x^*\right)\right]}_{A} + \underbrace{\sum_{t=K+1}^{T} \mathsf{E}\left[U_t\left(x_t^{\text{UCB}}\right) - L_t\left(x_t^{\text{UCB}}\right)\right]}_{B} + \underbrace{\sum_{t=K+1}^{T} \mathsf{E}\left[L_t\left(x_t^{\text{UCB}}\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right]}_{C}.
\end{aligned}$$

The first inequality follows from the property that $r_t \in [0, 1]$. The second inequality follows from the definition of x_t^{UCB}.

Bound A

We bound A. We have

$$
\begin{aligned}
A &= \sum_{t=K+1}^{T} \mathsf{E}\left[\left(g_\theta\left(x^*\right) - U_t\left(x^*\right)\right) \mathbf{1}\left\{g_\theta\left(x^*\right) > U_t\left(x^*\right)\right\}\right] + \sum_{t=K+1}^{T} \mathsf{E}\left[\left(g_\theta\left(x^*\right) - U_t\left(x^*\right)\right) \mathbf{1}\left\{g_\theta\left(x^*\right) \leq U_t\left(x^*\right)\right\}\right] \\
&\leq \sum_{t=K+1}^{T} \mathsf{E}\left[\left(g_\theta\left(x^*\right) - U_t\left(x^*\right)\right) \mathbf{1}\left\{g_\theta\left(x^*\right) > U_t\left(x^*\right)\right\}\right] \\
&\leq \sum_{t=K+1}^{T} \mathbb{P}\left(g_\theta\left(x^*\right) > U_t\left(x^*\right)\right) \\
&\leq \sum_{t=K+1}^{T} e^{-2\beta^2 \log T} \\
&= \sum_{t=K+1}^{T} \frac{1}{T^{2\beta^2}} \\
&\leq T^{1-2\beta^2}.
\end{aligned}
$$

The second inequality follows from the property that $r_t \in [0,1]$. The third inequality follows from the definition of $U_t(\cdot)$ and Hoeffding's Inequality.

Bound C

We bound C. We have

$$
\begin{aligned}
C &= \sum_{t=K+1}^{T} \mathsf{E}\left[\left(L_t\left(x_t^{\text{UCB}}\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right)\mathbf{1}\left\{L_t\left(x_t^{\text{UCB}}\right) > g_\theta\left(x_t^{\text{UCB}}\right)\right\}\right] \\
&\quad + \sum_{t=K+1}^{T} \mathsf{E}\left[\left(L_t\left(x_t^{\text{UCB}}\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right)\mathbf{1}\left\{L_t\left(x_t^{\text{UCB}}\right) \leq g_\theta\left(x_t^{\text{UCB}}\right)\right\}\right] \\
&\leq \sum_{t=K+1}^{T} \mathsf{E}\left[\left(L_t\left(x_t^{\text{UCB}}\right) - g_\theta\left(x_t^{\text{UCB}}\right)\right)\mathbf{1}\left\{L_t\left(x_t^{\text{UCB}}\right) > g_\theta\left(x_t^{\text{UCB}}\right)\right\}\right] \\
&\leq \sum_{t=K+1}^{T} \mathbb{P}\left(L_t\left(x_t^{\text{UCB}}\right) > g_\theta\left(x_t^{\text{UCB}}\right)\right) \\
&\leq \sum_{t=K+1}^{T} e^{-2\beta^2 \log T} \\
&= \sum_{t=K+1}^{T} \frac{1}{T^{2\beta^2}} \\
&\leq T^{1-2\beta^2}.
\end{aligned}
$$

The second inequality follows from the property that $r_t \in [0,1]$. The third inequality follows from the definition of $L_t(\cdot)$ and Hoeffding's Inequality.

Bound B

We bound B. Define $\mathcal{T}_x \triangleq \left\{t : t \in \{K+1, \cdots, T\}, x_t^{\mathrm{UCB}} = x\right\}$. We have

$$
\begin{aligned}
B &\leq \sum_{t=K+1}^{T} \mathsf{E}\left[2\beta\sqrt{\frac{\log T}{N_{t-1}(x_t^{\mathrm{UCB}})}}\right] \\
&= \mathsf{E}\left[\sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}_x} 2\beta\sqrt{\frac{\log T}{N_{t-1}(x)}}\right] \\
&= 2\beta\sqrt{\log T}\mathsf{E}\left[\sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}_x} \frac{1}{\sqrt{N_{t-1}(x)}}\right] \\
&= 2\beta\sqrt{\log T}\mathsf{E}\left[\sum_{x\in\mathcal{X}}\sum_{n=1}^{|\mathcal{T}_x|} \frac{1}{\sqrt{n}}\right] \\
&\leq 2\beta\sqrt{\log T}\mathsf{E}\left[\sum_{x\in\mathcal{X}}\int_{n=0}^{|\mathcal{T}_x|} \frac{1}{\sqrt{n}}dn\right] \\
&= 2\beta\sqrt{\log T}\mathsf{E}\left[\sum_{x\in\mathcal{X}} 2\sqrt{|\mathcal{T}_x|}\right] \\
&\leq 4\beta\sqrt{K\,(T-K)\log T} \\
&\leq 4\beta\sqrt{KT\log T}.
\end{aligned}
$$

The third inequality follows from Cauchy Schwarz's inequality.

Putting Everything Together

Therefore,

$$\text{Regret}\left(T, \pi^{\text{UCB}}, \theta\right) \quad \leq \quad K + 2T^{1-2\beta^2} + 4\beta\sqrt{KT\log T}.$$

By taking $\beta = 1/2$, we have

$$\text{Regret}\left(T, \pi^{\text{UCB}}, \theta\right) \quad \leq \quad K + 2\sqrt{T} + 2\sqrt{KT\log T}.$$