Multi-Armed Bandits

Andrew Nobel

October, 2023

Stochastic Multi-Armed Bandits

Stochastic Multi-Armed Bandits

Setting: Casino with K slot machines (one-armed bandits)

- Reward for playing machine/arm k is random with distribution P_k, and expected value α_k (both unknown)
- Successive plays of same or different arms are independent

Notation: Let $\alpha^* = \max_k \alpha_k$ and $k^* \in \operatorname{argmax}_k \alpha_k$

Goal: Maximize expected return over multiple rounds of play

- Optimal strategy plays best arm(s) k^* at each round
- Can we find an adaptive strategy that does almost as well?

Background, Exploration-Exploitation

Multi-armed bandit problem is an example of sequential decision making, simple case of reinforcement learning

- Result of playing arm k gives information only about arm k
- To identify optimal arm(s) a strategy must repeatedly sample every arm

At each round t a good strategy needs to balance

- Exploitation: play arm with largest average reward (greedy action)
- Exploration: play other arms to better estimate their expected rewards

Reinforcement vs Supervised Learning (from Sutton and Barto)

The most important feature distinguishing reinforcement learning from other types of learning is that it uses training information that evaluates the actions taken rather than instructs by giving correct actions. This is what creates the need for active exploration, for an explicit search for good behavior. Purely evaluative feedback indicates how good the action taken was, but not whether it was the best or the worst action possible. Purely instructive feedback, on the other hand, indicates the correct action to take, independently of the action actually taken. This kind of [instructive] feedback is the basis of supervised learning, which includes large parts of pattern classification, artificial neural networks, and system identification. In their pure forms, these two kinds of feedback are quite distinct: evaluative feedback depends entirely on the action taken, whereas instructive feedback is independent of the action taken.

Stochastic Multi-armed Bandit Problem

Preliminaries

- Let $X_{k,1}, X_{k,2}, \ldots \in [0,1]$ iid $\sim P_k$ be the reward sequence for arm k
- Assume reward sequences for different arms are independent

At each round $t \geq 1$

- Forecasting strategy f selects arm $f(t) \in [K]$ based on rewards received in rounds $1, \ldots, t-1$
- Forecaster receives reward $X_{f(t),t}$ independent of previous rewards

Pseudo-Regret

Definition: The pseudo-regret of a forecasting strategy f at time n is

$$\overline{R}_n \;=\; n\alpha^* - \mathbb{E}\left(\sum_{t=1}^{n} \alpha_{f(t)}\right)$$

Note: $\overline{R}_n$ is the difference between the expected reward of the optimal strategy and the expected reward of the strategy f

Counts and Mean Estimates

Definition: For $t \geq 1$ and $k \in [K]$ let the count

$$T_k(t) \;=\; \sum_{s=1}^{t} \mathbb{I}(f(s) = k)$$

be the number of times arm k is played by strategy f in rounds $1, \ldots, t$

Note: Natural estimate of α_k after t rounds of play is the average

$$\hat{\alpha}_{k,t} \;=\; \frac{1}{T_k(t)} \sum_{s \geq 1} X_{k,s}\, \mathbb{I}(f(s) = k \wedge T_k(s) \leq t)$$

of all rewards f receives when playing arm k

Suboptimality and Pseudo-Regret

Definition: Suboptimality parameter of arm k is $\Delta_k = \alpha^* - \alpha_k$

Fact: For each $n \geq 1$ the pseudo-regret may be written as

$$\overline{R}_n = \sum_{k=1}^{K} \Delta_k \mathbb{E} T_k(n)$$

Upper Confidence Bound (UCB)

Confidence bound: Fix $\gamma > 2$. For each round $t \geq 1$ and arm k define

$$\hat{U}_{k,t} = \hat{\alpha}_{k,T_k(t-1)} + \sqrt{\frac{\gamma \log t}{2T_k(t-1)}}$$

- $\hat{\alpha}_{k,T_k(t-1)}$ is an estimate of α_k with sample size $T_k(t-1)$
- $\hat{U}_{k,t} =$ estimate of α_k plus uncertainty, depending on sample size
- By Hoeffding $\mathbb{P}(\alpha_k \geq \hat{U}_{k,t}) \leq t^{-\gamma}$
- $\hat{U}_{k,t}$ is a $(1 - t^{-\gamma})$-UCB for α_k

UCB Strategy

γ-UCB stategy: For each round $t \geq 1$ select arm

$$f(t) \in \operatorname*{argmax}_{k \in [K]} \hat{U}_{k,t}$$

- γ-UCB treats $\hat{U}_{k,t}$ as an optimistic estimate of α_k
- γ-UCB uses value and uncertainty of mean estimates to trade off exploration and exploitation
- If $T_k(t-1) = 0$ then $\hat{U}_{k,t} = \infty$
- In first K rounds each arm selected once

Upper Bound on Pseudo-Regret of γ-UCB

Theorem: Assume all rewards take values in $[0, 1]$ and $\gamma > 2$. For each $n \geq 1$ the γ-UCB strategy f satisfies

$$\overline{R}_n \leq \sum_{k:\Delta_k>0} \left(\frac{2\gamma}{\Delta_k} \log n + \frac{\gamma}{\gamma - 2} \right)$$

Kullback-Liebler (KL) Divergence

Recall: The KL-divergence between distributions P, Q on a countable set $\mathcal{X}$

$$D(P,Q) \;=\; \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} \;=\; \mathbb{E}_P\left[\log\frac{P(X)}{Q(X)}\right]$$

- $D(P,Q) \geq 0$ with equality iff $P = Q$. Possible that $D(P,Q) = +\infty$
- $D(P,Q)$ is not a metric: in general $D(P,Q) \neq D(Q,P)$
- Tensorization: $D(\otimes_{i=1}^n P_i, \otimes_{i=1}^n Q_i) \;=\; \sum_{i=1}^n D(P_i,Q_i)$
- $D(P,Q)$ is jointly convex in its first and second arguments
- Pinsker's Inequality: $\sum_{x\in\mathcal{X}} |P(x) - Q(x)| \leq \sqrt{D(P,Q)/2}$

Preliminaries

Bernoulli case: When $P = \text{Bern}(p)$ and $Q = \text{Bern}(q)$ write $D(P, Q)$ as

$$D(p, q) \;=\; p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

If $0 \leq p, q \leq 1$ then we have upper and lower bounds

$$2(p-q)^2 \;\leq\; D(p, q) \;\leq\; \frac{(p-q)^2}{q(1-q)}$$

First inequality follows from Pinsker, second from $\log(x) \leq x - 1$

Fact: If $X_1, X_2, \ldots, X \in \mathbb{R}$ are iid with $\mathbb{E}|X| < \infty$ and $\mathbb{E}X > 0$ then

$$\lim_{n \to \infty} \frac{1}{n} \max_{1 \leq t \leq n} \sum_{i=1}^{t} X_i \;=\; \mathbb{E}X$$

Lower Bound on Pseudo-Regret of γ-UCB

Suppose $P_k = \text{Bern}(\alpha_k)$ for $k \in [K]$. Let f be a selection strategy such that

$$\lim_{n\to\infty} \frac{\mathbb{E}T_k(n)}{n^c} = 0$$

for every every $c > 0$ and any arm k with $\Delta_k > 0$

Theorem: For each $n \geq 1$ the strategy f satisfies

$$\liminf_{n\to\infty} \frac{\overline{R}_n}{\log n} \geq \sum_{k:\Delta_k>0} \frac{\Delta_k}{D(\alpha_k, \alpha^*)} \geq \sum_{k:\Delta_k>0} \frac{\alpha^*(1-\alpha^*)}{\Delta_k}$$