Algorithmic Foundations of Learning

Lecture 15
Stochastic Multi-Armed Bandit Problem and Algorithms

Patrick Rebeschini

Department of Statistics
University of Oxford

Online Statistical Learning

At every time step $t = 1, 2, \ldots, n$:

1. Choose an action $A_t \in \mathcal{A}$
2. A data point Z_t is sampled independently from an unknown distribution
   - **Full Information:** Z_t is revealed
   - **Bandit:** Z_t is not revealed
3. Suffer a loss $\ell(A_t, Z_t)$

Define the expected/population loss/risk as $\boxed{a \in \mathcal{A} \longrightarrow r(a) := \mathbf{E}\,\ell(a, Z)}$

Goal: Minimize the (normalized) pseudo-regret defined as

$$\boxed{\frac{1}{n}\sum_{t=1}^{n} r(A_t) - \inf_{a \in \mathcal{A}} r(a)}$$

A_t (possibly random) function of information available up to time t:

- **Full Information:** function of $\{A_1, ..., A_{t-1}\}$ and $\{Z_1, ..., Z_{t-1}\}$
- **Bandit:** function of $\{A_1, ..., A_{t-1}\}$ and $\{\ell(A_1, Z_1), ..., \ell(A_{t-1}, Z_{t-1})\}$

Multi-Armed Bandit Problem

Classical setup:

- **Arms:** $\mathcal{A} = \{1, \ldots, k\}$ and $|\mathcal{A}| = k$
- **Rewards:**
  - $Z_t = (Z_{t,1}, \ldots, Z_{t,k}) \in [0,1]^k$
  - $Z_{1,a}, \ldots, Z_{n,a} \in [0,1]$ is i.i.d. from unknown distrib. mean μ_a
- **Loss:** $\ell(A_t, Z_t) = -Z_{t,A_t}$

Expected loss: $r(a) = \mathbf{E}\,\ell(a, Z) = -\mathbf{E}Z_a := -\mu_a$

Pseudo-regret:

$$R_n := n\mu_{a^\star} - \sum_{t=1}^{n} \mu_{A_t}$$

$$a^\star \in \operatorname*{argmax}_{a \in [k]} \mu_a$$

A_t is function of $A_1, ..., A_{t-1}$ and $Z_{A_1}, ..., Z_{A_{t-1}}$

Note: Learning occurs when algorithm achieves **sub-linear** growth with n

Multi-Armed Bandit Problem

- Number of times arm a is pulled up to time t: $\boxed{N_{t,a} := \sum_{s=1}^{t} 1_{A_s=a}}$
- Sub-optimality gap of arm a: $\boxed{\Delta_a := \mu_{a^\star} - \mu_a}$

Proposition 15.1

$$\boxed{R_n = \sum_{a \in \mathcal{A}} \Delta_a N_{n,a}}$$

Proof. $n = \sum_{a \in \mathcal{A}} N_{n,a}$ and $\sum_{t=1}^{n} \mu_{A_t} = \sum_{a \in \mathcal{A}} \mu_a N_{n,a}$.

Q. How do we construct an algorithm?

A. Use sample means...

$$\boxed{M_{t,a} := \frac{1}{N_{t,a}} \sum_{s=1}^{t} Z_{s,a} 1_{A_s=a}}$$

Explore-Then-Commit

Algorithm 1: Explore-then-Commit(ε)

for $t = 1, \ldots, \varepsilon k$ **do**
  set $A_t = (t-1) \pmod k + 1$;
end
for $t = \varepsilon k + 1, \ldots, n$ **do**
  set $A_t \in \operatorname{argmax}_{a \in \mathcal{A}} M_{\varepsilon k, a}$;
end

Exploration/Exploitation tradeoff controlled by $\varepsilon \in \mathbb{N}_+$

Linear pseudo-regret for Explore-Then-Commit (Proposition 15.2)

There exists a stochastic multi-armed bandit problem such that

$$\boxed{\mathbf{E}R_n = cn + \tilde{c}} \qquad \textbf{(i.e., bad algorithm)}$$

for some constants $c, \tilde{c} \in \mathbb{R}_+$ that do not depend on n

Proof of Proposition 15.2

- Two arms, $k = 2$
- The optimal arm $a^\star$ has a Bernoulli (0 or 1) reward with mean $\mu_{a^\star} > \mu_a$
- The suboptimal arm a have a fix reward equal to $\mu_a < 1$
- The probability of not choosing the best arm in the explor. phase is positive:

$$p = \mathbf{P}(M_{2\varepsilon,a^\star} < M_{2\varepsilon,a}) = \mathbf{P}(\text{Binomial}(\varepsilon, \mu_{a^\star}) < \varepsilon\mu_a) > 0.$$

- The number of times that the sub-optimal arm is played after the exploration phase is equal to $n - \varepsilon k$ with probability p (the suboptimal arm is played ε times during the exploration phase)
- By Proposition 15.1,

$$\mathbf{E}R_n = \Delta_a \mathbf{E}N_{n,a} = \Delta_a(\varepsilon + (n - \varepsilon k)p) = \Delta_a pn + \Delta_a \varepsilon(1 - kp)$$

ε-Greedy

IDEA: Keep exploration on

Algorithm 2: Greedy(ε)

---

for $t = 1, \ldots, k$ **do**
  set $A_t = t$;
end
for $t = k+1, \ldots, n$ **do**
  set $A_t \begin{cases} \in \operatorname{argmax}_{a \in \mathcal{A}} M_{t-1,a} & \text{with probability } 1-\varepsilon \\ A_t \sim \text{Unif}\{1, \ldots, k\} & \text{with probability } \varepsilon \end{cases}$
end

Exploration/Exploitation tradeoff controlled by $\varepsilon \in (0, 1)$

Linear pseudo-regret for ε-Greedy (Proposition 15.3)

There exists a stochastic multi-armed bandit problem such that

$$\boxed{\mathbf{E}R_n = cn + \tilde{c}} \qquad \textbf{(i.e., bad algorithm)}$$

for some constants $c, \tilde{c} \in \mathbb{R}_+$ that do not depend on n

Upper Confidence Bound (UCB) Algorithm

- For any arm $a \in \mathcal{A}$:

$$\mathbf{E}N_{n,a} \geq 1 + \frac{\varepsilon}{k}(n-k)$$

  In fact, after the initial phase when each arm is played once, there are still $n-k$ plays to be made and at every time step the probability that each arm is played is at least ε/k

- By Proposition 15.1,

$$\mathbf{E}R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbf{E}N_{n,a} \geq \frac{\varepsilon}{k}(n-k) \sum_{a \in \mathcal{A}} \Delta_a$$

Upper Confidence Bound (UCB) Algorithm

IDEA: Have exploration to depend on **confidence** of estimates

---

Algorithm 3: UCB(ε)

---

for $t = 1, \ldots, k$ **do**
  set $A_t = t$;
end
for $t = k+1, \ldots, n$ **do**
  set $A_t \in \operatorname{argmax}_{a \in \mathcal{A}} U_{t-1,a} := M_{t-1,a} + \sqrt{\frac{\varepsilon \log(t-1)}{2N_{t-1,a}}}$;
end

Exploration/Exploitation tradeoff controlled by $\varepsilon \in \mathbb{R}_+$

Logarithmic pseudo-regret for UCB — distribution-dependent (Theorem 15.4)

For any $\varepsilon > 1$

$$\mathbf{E}R_n \leq \log n \sum_{a \in \mathcal{A}} \frac{2\varepsilon}{\Delta_a} + \frac{2}{\varepsilon - 1} \sum_{a \in \mathcal{A}} \Delta_a$$

(i.e., good algorithm!)

Distribution-Dependent Bounds for UCB: Proof Ideas

Proposition 15.5

For any non-decreasing sequence $s_1 \leq \ldots \leq s_n$ in $\mathbb{R}_+$ and any $a \in \mathcal{A}$, we have

$$\mathbf{E}N_{n,a} \leq s_n + \sum_{t=k}^{n-1} \mathbf{P}(A_{t+1} = a | N_{t,a} \geq s_t)$$

Lemma 15.6

Let $A_{t+1} = \operatorname{argmax}_{a \in \mathcal{A}} U_{t,a} = M_{t,a} + \sqrt{\frac{\log(1/\delta)}{2N_{t,a}}}$. If $\Delta_a > 0$ we have

$$\mathbf{P}\left(A_{t+1} = a | N_{t,a} \geq 2\frac{\log(1/\delta)}{\Delta_a^2} \right) \leq 2\delta$$

Proposition 15.7

$$\mathbf{E}N_{n,a} \leq 2\frac{\varepsilon \log n}{\Delta_a^2} + \frac{2}{\varepsilon - 1}$$

Proof of Lemma 15.6 (Part I)

- By the definition of UCB, we have

$$\{A_{t+1} = a\} \subseteq \{U_{t,a^\star} \leq U_{t,a}\} \subseteq \big(\{U_{t,a^\star} \leq \mu_{a^\star}\} \cup \{\mu_{a^\star} \leq U_{t,a}\}\big)$$

- Let $s \in \mathbb{R}_+$. By the union bound,

$$\mathbf{P}(A_{t+1} = a | N_{t,a} \geq s) \leq \mathbf{P}(U_{t,a^\star} \leq \mu_{a^\star} | N_{t,a} \geq s) + \mathbf{P}(\mu_{a^\star} \leq U_{t,a} | N_{t,a} \geq s)$$

- If $X_1, \ldots, X_n$ are i.i.d. samples from $[0,1]$ with mean μ, then for any $x > 0$:

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq x\right) \leq e^{-2nx^2} \qquad \mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \leq -x\right) \geq e^{-2nx^2}$$

and with the choice $x = \sqrt{\frac{\log(1/\delta)}{2n}}$ we get, respectively,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq \delta \tag{1}$$

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \leq -\sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq \delta \tag{2}$$

Proof of Lemma 15.6 (Part II)

- By the independence between the rewards and the arms' pulls, using (2),

$$\mathbf{P}(U_{t,a^\star} \le \mu_{a^\star} | N_{t,a} \ge s) = \mathbf{P}\left(M_{t,a^\star} - \mu_{a^\star} \le -\sqrt{\frac{\log(1/\delta)}{2N_{t,a^\star}}} \middle| N_{t,a} \ge s \right)$$

$$= \mathbf{E}\left[\mathbf{P}\left(M_{t,a^\star} - \mu_{a^\star} \le -\sqrt{\frac{\log(1/\delta)}{2N_{t,a^\star}}} \middle| N_{t,a} \ge s, N_{t,a^\star} \right) \middle| N_{t,a} \ge s \right]$$

$$= \mathbf{E}\left[\mathbf{P}\left(M_{t,a^\star} - \mu_{a^\star} \le -\sqrt{\frac{\log(1/\delta)}{2N_{t,a^\star}}} \middle| N_{t,a^\star} \right) \middle| N_{t,a} \ge s \right]$$

$$\le \mathbf{E}[\delta | N_{t,a} \ge s] = \delta.$$

Distribution-Independent Bounds for UCB

NOTE: If $\Delta_a = \log n/n$ then previous result yields $R_n \lesssim n$

This can be improved:

Square-root pseudo-regret for UCB — distribution-independent (Thm 15.8)

$$\mathbf{E}R_n \leq 2\sqrt{2\varepsilon}\sqrt{kn\log n} + \frac{2k}{\varepsilon - 1}$$

- UCB achieves **optimal** **distribution-dependent** pseudo-regret
- UCB achieves **quasi-optimal** **distribution-independent** pseudo-regret (Next Lecture...)

NOTE: Playing Super Mario Bross involves a **state**... (Reinforcement Learning)
(also in that case there is the exploration/exploitation trade-off...)