**CS/CNS/EE 253: Advanced Topics in Machine Learning**
**Topic:** Dealing with Partial Feedback #1 **Lecturer:** Daniel Golovin
**Authors:** Daniel Golovin & Andreas Krause **Date:** Jan 27, 2009

## 7.1 Dealing with Partial Feedback

In previous lectures we talked about algorithms for sequential decision problems in the *full information* model. In that model, after we make a decision $x_t$ in round $t$, we observe a reward (or cost) for the decision $x_t$, as well as the rewards (or costs) for all the possible decisions we did not make.

This is a natural model for some applications, including many applications in which we use an online algorithm as a streaming algorithm to solve problems on massive data sets, as in online support vector machines. However in some applications you are unable to pose these counterfactual, "what would be the reward if" queries to the world. Instead you only get to see the reward of what you chose to do. We call this the *partial information* or *bandit feedback* models[1].

For now, we consider the case where there are $k$ choices (or "arms") and the rewards for each arm are stochastic, that is, the rewards from each arm $j$ are drawn from an unknown fixed distribution $\mathcal{R}_j$ that does not change over time. As usual, we make the simplifying assumption that the rewards are always in $[0, 1]$.

When faced with this situation, what should we do?

## 7.2 Some Greed Is Good, But Too Much Will Harm You.

One suggestion from the class was to try each arm some number $q$ times, and then play the arm with the highest average reward.

The problem for this algorithm is there might be arms with high expected reward and very high variance relative to the other arms that we simply won't "discover" with sufficiently high probability. That is, imagine there are arms $\{a_1, a_2\}$, with the reward for $a_1$ deterministically set to $\delta$ and the reward from $a_2$ distributed as Bernoulli($2\delta$) (i.e., its reward is 1 with probability $2\delta$ and 0 otherwise). If $\delta = (2q)^{-1}$, it can be shown that with probability $(1 - 2\delta)^q \geq 4^{-2\delta q} = 1/4$, our historical average for $a_2$ will be zero. Hence after the first $q$ times we'll never play $a_2$ again, and so in this case we'll suffer $\delta$ regret in each round. Since this occurs with probability at least $1/4$, our expected regret grows as linearly with $T$ after the first $2q$ rounds. Nor is setting $q$ very large appealing, since we pay a price to play every suboptimal arm $q$ times.

One way to address this problem is to define a decreasing function $\epsilon(t)$ and with probability $(1-\epsilon(t))$ play the arm with the highest observed average reward (i.e., exploit previous knowledge), and with probability $\epsilon(t)$ try a random arm. If $\sum_{t\geq 1} \epsilon(t)$ diverges, this ensures that we play each arm

[1] Historically, some of the problems posed here were introduced in the context of a gambler entering a rigged casino, encountering multiple slot-machines, and deciding adaptively what slot machines to play. "One armed bandit" is a slang term for slot machine. Hence came the terms"multiarmed bandit problem" and "bandit feedback."

infinitely often, and so the historically observed average reward for each arm $j$ converges to the mean of its reward distribution $\mathcal{R}_j$. If additionally $\limsup_{t\to\infty} \epsilon(t) = 0$, so that our exploration probability is dropping to zero, then eventually we will play only optimal arms with probability arbitrarily close to one. This is indeed the approach taken by the $\epsilon$-Greedy algorithm [1], and using the right function $\epsilon(t)$ does yield expected regret sublinear in $T$.

## 7.3 The Value of Being Optimistic

The $\epsilon$-Greedy algorithm has some disadvantages. When it explores, it explores *uniformly* over the arms. This despite the fact that after a while it will have very good estimates of the mean rewards, and including some arms that are very, very likely to have terrible mean rewards. A better algorithm would still play each arm infinitely often so that the historically observed average reward for each arm $j$ converges to the mean of its reward distribution $\mathcal{R}_j$, however it would do so in a way that explored on "more promising" arms more often than on "less promising" ones. In the rest of these notes we describe such an algorithm, the UCB1 algorithm of Auer et al. [1], and we prove logarithmic regret bounds (in the number of rounds $T$) for it. Incidently, the value of optimism shows up in other algorithms as well, for example, in reinforment learning it can be seen in the R-Max algorithm [4] and the UCRL and UCRL2 algorithms [3, 2].

### 7.3.1 Notation

- $j$: Index of slot machine arm (1 to $k$).
- $t$: a round number. At the start of round $t$, we have made $t-1$ plays.
- $T$: Total number of plays we did so far.
- $\mathcal{R}_j$: Reward distribution for arm $j$, with support in $[0,1]$ (i.e., they do not take any values outside $[0,1]$).
- $R_{j,c}$: Random variable for reward of arm $j$ when played for the $c^{\text{th}}$ time. For all values of $c$, $R_{j,c}$ is drawn from $\mathcal{R}_j$. Note the reward distribution $\mathcal{R}_j$ does not change over time. All $R_{j,c}$ are possibly continuous, but supported in the interval $[0,1]$. All $R_{j,c}$ are independent.
- $C(j,t)$: Number of times arm $j$ pulled during the first $t$ plays. Note that $C(j,t)$ is a random quantity.
- $\mu_j = \mathbb{E}[R_{j,c}]$, and $\mu^* = \max_j \mu_j$
- $\Delta_j = \mu^* - \mu_j$, and $\Delta = \min_j \Delta_j$
- regret$(t)$: Cumulative regret after $t$ plays.
- $\bar{R}_{j,c}$ is the sample average of all rewards obtained from arm $j$ during the first $c$ plays of it. (i.e., if we've observed rewards $r_1, \ldots, r_c$ then $\bar{R}_{j,c} = \frac{1}{c}(x_1 + \cdots + x_c)$).

### 7.3.2 The Upper Confidence Bound algorithm (UCB1)

UCB1 is based on the idea of maintaining a confidence interval for the mean payoff for an arm, and after playing each arm once to initialize the intervals, always playing the arm with the maximum upper bound on its interval.

---

**UCB1 Algorithm**

Initially, play each arm once (hence $C(j,t) \geq 1$ for all $t \geq k$).

Loop (for $t = k+1$ to $n$)

- For each arm $j$ compute a *score* (sometimes called an "index")

$$v_j = \bar{R}_{j,C(j,t-1)} + \sigma(t, C(j,t-1)),$$

  where $\sigma(t,c) := \sqrt{\frac{\ln t}{c}}$.
- Play the arm with maximum score, $j_{\max} = \text{argmax}_j\, v_j$.
- Observe payoff and update average observed reward for $j_{\max}$

---

Note that the confidence interval $[\bar{R}_{j,C(j,t-1)} \pm \sigma(t, C(j,t-1))]$ slowly grows for arms that are not played, because $\sigma(t,c)$ is an increasing function of $t$. This is how UCB1 ensures $\lim_{t\to\infty} C(j,t) = \infty$ and thus $\lim_{t\to\infty} \bar{R}_{j,C(j,t)} = \mu_j$. Also note that due to concentration of measure of the observed historical average rewards, the probability that $\mu_j > v_j$ is extremely small. This is made formal in the analysis by invoking a Chernoff-Hoeffding bound. To get some intuition on it, consider the case that $\mathcal{R}_j$ is Bernoulli($p$), so that $\bar{R}_{j,c}$ is distributed like $\frac{1}{c}\,\text{Bin}(c,p)$. The standard deviation of $\bar{R}_{j,c}$ is then $\sqrt{\frac{p(1-p)}{c}} \leq \frac{1}{2\sqrt{c}}$. The upper confidence bound used by the algorithm is thus at least $2\sqrt{\ln t}$ standard deviations above the historically observed mean.

### 7.3.3 Formal Analysis

**Theorem 7.3.1** *The expected regret of UCB1 after $T$ rounds is at most*

$$\sum_{j=1}^{k} \frac{4\ln T}{\Delta_j} + \left(1 + \frac{\pi^2}{3}\right)\Delta_j$$

Here is some intuition for the proof. We bound the regret in terms of the number of times we play suboptimal arms. To bound the latter, we argue that after we sample a suboptimal arm $j$ "enough" times (which turns out be $(4\ln T)/\Delta_j^2$), its confidence interval has shrunk enough so that to pick it, either our estimate of its mean is far too high (i.e., its $\mu_j$ is actually below the minimum of its confidence interval $\bar{R}_{j,C(j,t-1)} \pm \sigma(t, C(j,t-1))$) or our estimate of the mean of the optimal arm $j^*$ is far too low (i.e., $\mu^*$ is above the maximum of $j^*$'s confidence interval). We then show that

the probability of either of these bad events happening is at most $2t^{-2}$, and put it all together to complete the proof.

**Proof:** Lemma 7.3.3 states that $\mathbb{E}\left[\text{regret}(T)\right] = \sum_j \mathbb{E}[C(j,T)]\Delta_j$. Thus to prove Theorem 7.3.1 it is sufficient to bound $\mathbb{E}[C(j,T)]$ for all arms $j$ and then apply Lemma 7.3.3. Let $j^*$ be the optimal arm. Suppose, at some time $t \leq T$, $UCB_1$ pulls a suboptimal arm $j$. That means, that

$$\bar{R}_{j,C(j,t-1)} + \sigma(t, C(j,t-1)) \geq \bar{R}_{j^*,C(j^*,t-1)} + \sigma(t, C(j^*,t-1)).$$

For brevity, we define $\bar{R}_j = \bar{R}_{j,C(j,t-1)}$ to be the historical average reward seen for $j$ by round $t$, $\sigma_j = \sigma(t, C(j,t-1))$, $\bar{R}_* = \bar{R}_{j^*,C(j^*,t-1)}$ to be the historical average reward seen for $j^*$ by round $t$, and $\sigma_* = \sigma(t, C(j^*,t-1))$.

Hence, in this case,

$$\begin{array}{rlcl} & \bar{R}_j + 2\sigma_j - \sigma_j + (\mu_j - \mu_j) & \geq & \bar{R}_* + \sigma_* + (\mu^* - \mu^*) \\ \Leftrightarrow & \underbrace{\bar{R}_j - (\mu_j + \sigma_j)}_{A} + \underbrace{(\mu_j - \mu^* + 2\sigma_j)}_{B} & \geq & \underbrace{\bar{R}_* - (\mu^* - \sigma_*)}_{-C} \end{array}$$

We can see that $A \geq 0$ or $B > 0$ or $C \geq 0$, (since otherwise we have $A + B < 0 < -C$, violating our inequality above). Thus, at least one of the following inequalities must hold:

$$\bar{R}_{j,C(j,t-1)} \geq \mu_j + \sigma(t, C(j,t-1)) \tag{7.3.1}$$

$$\bar{R}_{j^*,C(j^*,t-1)} \leq \mu^* - \sigma(t, C(j^*,t-1)) \tag{7.3.2}$$

$$\mu^* > \mu_j + 2\sigma(t, C(j,t-1)) \tag{7.3.3}$$

In order to bound the probability of (7.3.1) and (7.3.2), we use the Chernoff-Hoeffding inequality:

**Fact 7.3.2 (Chernoff-Hoeffding inequality)** *Let $X_1, \ldots, X_n$ be independent random variables supported on $[0,1]$, with $\mathbb{E}[X_i] = \mu$. Then, for every $a > 0$,*

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + a) \leq e^{-2a^2 n}$$

*and*

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i < \mu - a) \leq e^{-2a^2 n}$$

Hence, we can bound the probability of (7.3.1) as

$$P\left(\bar{R}_{j,C(j,t-1)} \geq \mu_j + \sigma(t, C(j,t-1))\right) \leq e^{-2\sigma(t,C(j,t-1))^2 C(j,t-1)} = e^{-2\frac{\ln t}{C(j,t-1)}C(j,t-1)} = e^{-2\ln t} = t^{-2}.$$

Similarly,

$$P(\bar{R}_{j^*,C(j^*,t-1)} \leq \mu^* - \sigma(t, C(j^*,t-1))) \leq t^{-2}.$$

Hence, (7.3.1) and (7.3.1) are very unlikely events. Now, note that whenever $C(j,t-1) \geq \ell = \lceil (4\ln T)/\Delta_j^2 \rceil$, (7.3.3) must be false, since

$$\mu_j + 2\sigma(t, C(j,t-1)) = \mu_j + 2\sqrt{\frac{\ln t}{C(j,t-1)}} \leq \mu_j + 2\sqrt{\frac{\ln t}{\frac{4\ln T}{\Delta_j^2}}} \leq \mu_j + \Delta_j = \mu^*$$

Hence, if arm $j$ has already been played at least $\ell \geq \lceil (4\ln T)/\Delta_j^2 \rceil$ times, then inequality (7.3.3) must be false, and hence arm $j$ is pulled with probability at most $2t^{-2}$ in round $t$.

Now we bound $\mathbb{E}[C(j,T)]$. Note that $C(j,T) \leq \ell + \max(0, C(j,T) - \ell)$ and so this holds in expectation as well. We thus bound $\mathbb{E}[\max(0, C(j,T) - \ell)]$. Let $I(j,t)$ be the indicator random variable for the event "we played $j$ on round $t$". Then

$$\mathbb{E}[\max(0, C(j,T) - \ell)] = \sum_{t=1}^{T} \mathbf{Pr}[I(j,t) \text{ and } C(j,t-1) \geq \ell]$$

Our arguments above show the above sum to at most $\sum_{t=1}^{T} 2t^{-2}$ whenever $\ell \geq \lceil (4\ln T)/\Delta_j^2 \rceil$. Since $\sum_{t=1}^{\infty} 2t^{-2} = \pi^2/3$, we get $\mathbb{E}[C(j,T)] \leq \lceil (4\ln T)/\Delta_j^2 \rceil + \pi^2/3 \leq (4\ln T)/\Delta_j^2 + 1 + \pi^2/3$. Applying Lemma 7.3.3 concludes the proof. ∎

**Lemma 7.3.3** *The expected regret of UCB1 after $T$ rounds is*

$$\mathbb{E}[\text{regret}(T)] = \sum_j \mathbb{E}[C(j,T)]\Delta_j$$

**Proof:** Let $j^*$ be the optimal arm, let $R_{j,t}$ be random variables sampled from $\mathcal{R}_j$, indicating the reward of arm $j$ if played in round $t$, and let $R_t^*$ be shorthand for $R_{j^*,t}$. Let $I(j,t)$ be the indicator random variable for the event "we played $j$ on round $t$". Then

$$\begin{aligned}
\mathbb{E}[\text{regret}(T)] &= \mathbb{E}\left[\textstyle\sum_{t=1}^{T}\sum_{j=1}^{k}(R_t^* - R_{j,t})\, I(j,t)\right] && \text{[definition of regret]} \\
&= \textstyle\sum_{t=1}^{T}\sum_{j=1}^{k}\mathbb{E}[(R_t^* - R_{j,t})\, I(j,t)] && \text{[linearity of expectation]} \\
&= \textstyle\sum_{t=1}^{T}\sum_{j=1}^{k}\mathbb{E}[R_t^* - R_{j,t}] \cdot \mathbb{E}[I(j,t)] && [(R_t^* - R_{j,t}) \text{ and } I(j,t) \text{ independent }] \\
&= \textstyle\sum_{t=1}^{T}\sum_{j=1}^{k}\Delta_j \mathbb{E}[I(j,t)] && \text{[definition of } \Delta_j] \\
&= \textstyle\sum_{j=1}^{k}\sum_{t=1}^{T}\Delta_j \mathbb{E}[I(j,t)] \\
&= \textstyle\sum_j \mathbb{E}[C(j,T)]\Delta_j
\end{aligned}$$

For the third line, recall that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ if $X$ and $Y$ independent. ∎

## References


[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.

[2] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 89–96. 2009.

[3] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *NIPS*, pages 49–56, 2006.

[4] Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *JMLR*, 3:213–231, 2003.