DATA 37200: Learning, Decisions, and Limits (Winter 2026)

Lecture 3: UCB1 Algorithm

Instructor: Frederic Koehler

Azuma–Hoeffding inequality (bounded differences)

Let $(X_t)_{t=0}^n$ be a martingale w.r.t. $(\mathcal{F}_t)$.
Assume bounded increments: for constants $c_1, \dots, c_n$,

$$|X_t - X_{t-1}| \le c_t \qquad \text{almost surely for each } t = 1, \dots, n.$$

Then for all $u > 0$,

$$\mathbb{P}(X_n - X_0 \ge u) \;\le\; \exp\left(-\frac{u^2}{2\sum_{t=1}^n c_t^2}\right),$$

and similarly

$$\mathbb{P}(|X_n - X_0| \ge u) \;\le\; 2\exp\left(-\frac{u^2}{2\sum_{t=1}^n c_t^2}\right).$$

Application: concentration of regret

- Today we will see how to use Azuma-Hoeffding to analyze the UCB algorithm.
- RMK: design of the algorithm is *tightly connected* with the analysis!
- We will start with a bound on *expected regret* and then later show a high probability bound.

Regret decomposition (for later)

Fix a policy for the bandit. Define (random) regret

$$R_T := \sum_{t=1}^{T} \big(\mu^\star - r(t)\big), \qquad \mu^\star = \max_i \mu_i, \quad \mu_i := \mathbb{E}[r(t) \mid i(t) = i].$$

Let $\mathcal{F}_t$ be the history up to time t (arms and rewards up to time t).

Remark: we can decompose R_T as "predictable part + noise":

$$R_T = \sum_{t=1}^{T} \big(\mu^\star - \mu_{i(t)}\big) \; - \; \sum_{t=1}^{T} \big(r(t) - \mu_{i(t)}\big).$$

The first term is called the *pseudoregret*. Unlike regret, it is nonnegative (why?) In bandits textbooks, it is often nice to study the pseudoregret in place of the regret. *Note that they have the same expectation.*

UCB algorithm (Upper Confidence Bound / UCB1)

We maintain for each arm $i \in [K]$:

$$N_i(t) = \#\{s \le t-1 : \; i(s) = i\}$$

$$\hat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{s \le t-1: \; i(s)=i} r(s) \quad \text{(when } N_i(t) \ge 1\text{)}.$$

Algorithm (for $t = 1, \dots, T$):

1. **Initialization:** pull each arm once (for $t = 1, \dots, K$).
2. For $t > K$, compute indices

$$\mathrm{UCB}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{2 \log t}{N_i(t)}}.$$

3. Choose $i(t) \in \arg\max_i \mathrm{UCB}_i(t)$, observe reward $r(t) \in [0, 1]$, update.

Optimism-under-uncertainty: sample mean + confidence radius.

Azuma-Hoeffding confidence bounds for adaptively sampled means

Let $\mathcal{F}_t$ be the history up to time t. The choice $i(t)$ is $\mathcal{F}_{t-1}$-measurable, and rewards satisfy

$$\mathbb{E}[r(t) \mid i(t) = i] =: \mu_i, \qquad r(t) \in [0, 1].$$

For a fixed arm i, define martingale differences

$$X_t^{(i)} := \mathbf{1}\{i(t) = i\}\big(r(t) - \mu_i\big).$$

Then $\mathbb{E}[X_t^{(i)} \mid \mathcal{F}_{t-1}] = 0$ and $|X_t^{(i)}| \leq 1$ a.s. So $M_t^{(i)} := \sum_{s=1}^t X_s^{(i)}$ is a martingale with bounded increments.

Azuma-Hoeffding $\Rightarrow$ for any $\delta \in (0, 1)$ and any time t,

$$\Pr\left(|\hat{\mu}_i(t) - \mu_i| \geq \sqrt{\frac{2\log(t)}{N_i(t)}}\right) \;\leq\; 2/t,$$

interpreting $\hat{\mu}_i(t)$ as the mean of the $N_i(t)$ observed rewards from arm i.

Regret decomposition

Let $\mu^\star = \max_i \mu_i$ and define expected (pseudo-)regret

$$\bar{R}_T := T\mu^\star - \mathbb{E}\Big[\sum_{t=1}^{T} r(t)\Big] = \sum_{i=1}^{K} (\mu^\star - \mu_i)\,\mathbb{E}[N_i(T+1)].$$

Good and bad arms. Split arms into:

$$\mathcal{S} := \{i:\ \Delta_i := \mu^\star - \mu_i \le \varepsilon\}, \qquad \mathcal{L} := \{i:\ \Delta_i > \varepsilon\},$$

for a threshold $\varepsilon > 0$ to be chosen later.

Then

$$\begin{aligned} \bar{R}_T &= \sum_{i\in\mathcal{S}} \Delta_i\,\mathbb{E}[N_i(T+1)] + \sum_{i\in\mathcal{L}} \Delta_i\,\mathbb{E}[N_i(T+1)] && (1) \\ &\le\ \varepsilon T + \sum_{i\in\mathcal{L}} \Delta_i\,\mathbb{E}[N_i(T+1)]. && (2) \end{aligned}$$

So it remains to control $\mathbb{E}[N_i(T+1)]$ for arms with gap $> \varepsilon$.

Bounding pulls of a suboptimal arm

Fix a suboptimal arm i with gap $\Delta_i := \mu^\star - \mu_i > 0$.

On the *good event* that the UCB confidence bounds are valid for all arms at time t, if arm i is chosen at time t then

$$\hat{\mu}_i(t) + \sqrt{\frac{2\log t}{N_i(t)}} \;\geq\; \hat{\mu}_{i^\star}(t) + \sqrt{\frac{2\log t}{N_{i^\star}(t)}}.$$

Using the confidence bounds $\hat{\mu}_i(t) \leq \mu_i + \sqrt{\frac{2\log t}{N_i(t)}}$ and $\hat{\mu}_{i^\star}(t) \geq \mu^\star - \sqrt{\frac{2\log t}{N_{i^\star}(t)}}$, we get

$$\mu_i + 2\sqrt{\frac{2\log t}{N_i(t)}} \;\geq\; \mu^\star \quad\Rightarrow\quad 2\sqrt{\frac{2\log t}{N_i(t)}} \;\geq\; \Delta_i.$$

Hence, whenever i is pulled at time t on the good event,

$$N_i(t) \;\leq\; \frac{8\log t}{\Delta_i^2} \quad\Rightarrow\quad N_i(T+1) \;\leq\; 1 + \frac{8\log T}{\Delta_i^2}.$$

Bad event details

At step t we have a probability of K/t of having a bad event (one of the estimates of the arms is off). By linearity of expectation, the expected total contribution from bad events to regret is

$$\sum_{t=1}^{T} K/t \lesssim K \log T$$

so ignoring $O(K \log T)$ contribution to regret (which is lower-order compared to final regret bound), we can ignore the bad events.

Remark: even more precisely, the expected number of pulls of any particular arm i which occured due to "bad estimates" is $O(\log T)$. (Check yourself.)

Gap-independent regret bound via ε-splitting

Regret:

$$\bar{R}_T = \sum_{i:\Delta_i>0} \Delta_i \, \mathbb{E}[N_i(T+1)].$$

Fix $\varepsilon > 0$ and split suboptimal arms into

$$\mathcal{S} = \{i : \Delta_i \le \varepsilon\}, \qquad \mathcal{L} = \{i : \Delta_i > \varepsilon\}.$$

Small gaps: since $\sum_{i\in\mathcal{S}} N_i(T+1) \le T$,

$$\sum_{i\in\mathcal{S}} \Delta_i \, \mathbb{E}[N_i(T+1)] \;\le\; \varepsilon \, \mathbb{E}\Big[\sum_{i\in\mathcal{S}} N_i(T+1)\Big] \;\le\; \varepsilon T.$$

Large gaps: using the pull bound $\mathbb{E}[N_i(T+1)] \lesssim 1 + \frac{8\log T}{\Delta_i^2}$,

$$\sum_{i\in\mathcal{L}} \Delta_i \, \mathbb{E}[N_i(T+1)] \;\lesssim\; \sum_{i\in\mathcal{L}} \Delta_i + 8\log T \sum_{i\in\mathcal{L}} \frac{1}{\Delta_i} \;\le\; K + \frac{8K\log T}{\varepsilon},$$

since $\Delta_i > \varepsilon$ on $\mathcal{L}$. Therefore

$$\bar{R}_T \;\lesssim\; \varepsilon T \;+\; \frac{8K\log T}{\varepsilon} \;+\; K.$$

Optimize ϵ

We showed

$$\bar{R}_T \;\lesssim\; \varepsilon T \;+\; \frac{8K \log T}{\varepsilon} \;+\; K.$$

Optimize by $\varepsilon = \sqrt{\frac{8K \log T}{T}}$ to get

$$\bar{R}_T \;=\; O\big(\sqrt{K\ T\ \log T}\big).$$

In particular, for fixed K this is $O(\sqrt{T \log T})$.

NOTE: we saved because we did not waste time pulling bad arms ! Makes a lot of sense. We also assumed $K \log T \ll T$ to simplify the bound (if number of arms is similar to T, it becomes hopeless to find the best one.)

Extra slides

Some additional aspects of this problem related to the *pseudoregret* are commonly disussed in the literature.

Gap-dependent bound on expected (pseudo)regret

Suppose all gaps $\Delta_i > 0$. Then we can take $\varepsilon \to 0$ in our argument before and find that for expected regret

$$\bar{R}_T = \mathbb{E} R_T = \sum_{i:\Delta_i>0} \Delta_i \, \mathbb{E}[N_i(T+1)].$$

Large gaps bound: using the pull bound

$$\mathbb{E}[N_i(T+1)] \lesssim 1 + \frac{8 \log T}{\Delta_i^2},$$

$$\sum_i \Delta_i \, \mathbb{E}[N_i(T+1)] \;\lesssim\; \sum_i \Delta_i \;+\; 8 \log T \sum_i \frac{1}{\Delta_i}$$

we can write

$$\bar{R}_T \;\lesssim\; 8 \log(T) \sum_i \frac{1}{\Delta_i} + \; K.$$

So, for fixed Δ_i, expected regret grows *logarithmically* as $T \to \infty$.

Regret vs. pseudo-regret (Azuma-Hoeffding)

Recall the (random) regret and pseudo-regret:

$$R_T := \sum_{t=1}^{T} \big(\mu^\star - r(t)\big), \qquad \tilde{R}_T := \sum_{t=1}^{T} \big(\mu^\star - \mu_{i(t)}\big),$$

where $\mu_i := \mathbb{E}[r(t) \mid i(t) = i]$ and $\mu^\star = \max_i \mu_i$.

Their difference is the martingale noise term:

$$R_T - \tilde{R}_T = \sum_{t=1}^{T} \big(\mu_{i(t)} - r(t)\big) = -\sum_{t=1}^{T} \big(r(t) - \mu_{i(t)}\big) =: -M_T.$$

Let $\mathcal{F}_t$ be the history up to time t. Then

$$\mathbb{E}\big[r(t) - \mu_{i(t)} \mid \mathcal{F}_{t-1}\big] = 0, \qquad |r(t) - \mu_{i(t)}| \le 1,$$

so $(M_t)_{t \le T}$ is a martingale with bounded increments.

regret vs pseudo-regret

$$R_T - \tilde{R}_T = \sum_{t=1}^{T} \left(\mu_{i(t)} - r(t)\right) = -\sum_{t=1}^{T} \left(r(t) - \mu_{i(t)}\right) =: -M_T.$$

is a martingale with bounded increments.
Azuma-Hoeffding: for any $\delta \in (0,1)$,

$$\Pr\big(|R_T - \tilde{R}_T| \geq x\big) = \Pr\big(|M_T| \geq x\big) \leq 2\exp\Big(-\frac{x^2}{2T}\Big).$$

Setting $x = \sqrt{2T\log(2/\delta)}$ gives the high-probability bound

$$|R_T - \tilde{R}_T| \;\leq\; \sqrt{2T\log(2/\delta)} \qquad \text{with prob. } \;\geq 1-\delta.$$

In particular, $\mathbb{E}|R_T - \tilde{R}_T| \leq \sqrt{2T\log 2} = O(\sqrt{T})$.

Summary of gap-dependent theory

- Gap-dependent bound: fix gaps $\Delta_i > 0$ and consider large T behavior.
- Expected regret = expected pseudoregret = $O(\sum_i \log(T)/\Delta_i)$.
- So by Markov, with 99% probability pseudoregret is $O(\sum_i \log(T)/\Delta_i)$. Azuma-Hoeffding: true within $\pm\sqrt{T}$ for regret.
- Gap-independent bound: regret and pseudoregret whp is $O(\sqrt{KT\log(T)})$.
- High-probability $O(\log T)$ statement is *not* possible for realized regret. (Why?) $\sqrt{T}$ is a fundamental limit.

Next time

- Gap-independent bound: regret and pseudoregret whp is $O(\sqrt{KT\log(T)})$.
- How close is this to *optimal*?