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

Lecture 7: ϵ-Greedy contextual bandits

Instructor: Frederic Koehler

Reference

Chapter 3.4 of Rakhlin-Foster notes.

Continuing from last time

- Last time: introduced CB (contextual bandit) model.
- Saw that naive ETC (Explore-Then-Commit) does not work in our model. "Distribution shift"
- We saw how to *successfully* reduce CB to MAB if the number of contexts is small.
- This means: we used "black box" the fact that we have provably good algorithms for MAB ("MAB oracle"), to come up with algorithms for CB.
- This algorithm is reasonable for small number of contexts. Now lets go beyond this.
- We continue with the *reduction-based* approach !

The Contextual Bandit Setting

- **Setup:**
  - Time $t = 1, \ldots, T$.
  - Context space $\mathcal{X}$, Arm set $\mathcal{A} = [K]$.
  - Environment generates context $x(t) \in \mathcal{X}$.
  - We observe $x(t)$, choose arm $i(t) \in \mathcal{A}$, and observe reward $r(t)$.

- **Realizability Assumption:**
  - There exists $f^* \in \mathcal{F}$ such that

$$\mathbb{E}[r(t) \mid x(t), i(t)] = f^*(x(t), i(t)).$$

  - Optimal arm: $i^*(t) = \arg\max_{i \in \mathcal{A}} f^*(x(t), i)$.

- **Goal:** Minimize Regret defined as:

$$\text{Regret}_T = \sum_{t=1}^{T} \left(f^*(x(t), i^*(t)) - r(t) \right)$$

Our high-level strategy

- Last time: use an oracle to "black box" *decision-making* (MAB).
- Not a promising approach to "black box" decision-making in more general CB.
- Instead, lets "white box" decision-making and *black-box* forecasting/learning !
- "Specialization of labor": one agent handles decision-making, one agent handles forecasting.

The Online Regression Oracle

We will need to an access which can learn to predict *online*.

Definition (Online Square Loss Oracle)

An algorithm that produces predictors $\hat{f}_t \in \mathcal{F}$ based on the history $(x(1), i(1), r(1)), \ldots, (x(t-1), i(t-1), r(t-1))$.

Oracle Regret Bound ($R_{\mathbf{Sq}}$):

$$\sum_{t=1}^{T} (\hat{f}_t(x(t), i(t)) - r(t))^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (f(x(t), i(t)) - r(t))^2 \leq R_{\mathsf{Sq}}(T)$$

- Note: The oracle only sees the loss on the arms $i(t)$ we actually played.
- **Key point:** For CB, we (essentially) need the oracle to be accurate on *all* arms to find the argmax, but oracle is only trained/guaranteed to be accurate on *played* arms.

White box exploration strategy

1. For MAB we saw that UCB1 was a pretty good exploration strategy.
2. UCB1 doesn't just build estimates of arms, but also tries to measure their *uncertainty*.
3. Makes it tricky to build UCB based on a black-box forecaster.
4. Instead: lets use another beloved exploration strategy — ϵ-greedy !

Algorithm: ϵ-Greedy with Regression

Algorithm 1 ϵ-Greedy with Regression Oracle

Input: Exploration parameter $\epsilon \in (0, 1)$, Regression Oracle.
for $t = 1, \ldots, T$ **do**
    Receive context $x(t)$.
    Get current predictor $\hat{f}_t$ from the Regression Oracle.
    Compute Greedy Arm: $\hat{i}(t) = \arg\max_{i \in \mathcal{A}} \hat{f}_t(x(t), i)$.
    Arm Selection ($i(t)$):
    if Bernoulli$(\epsilon) = 1$ **then**
        Play $i(t) \sim$ Unif$(\mathcal{A})$ ▷ **Explore**
    else
        Play $i(t) = \hat{i}(t)$ ▷ **Exploit**
    end if
    Observe reward $r(t)$.
    Update Oracle with tuple $(x(t), i(t), r(t))$.
end for

Theorem: Regret of ϵ-Greedy

Theorem

The expected regret of the ϵ-greedy algorithm is bounded by:

$$\mathbb{E}[Regret_T] \leq \epsilon T + \sqrt{\frac{KT}{\epsilon} \cdot R_{Sq}(T)}$$

Optimizing ϵ, we get $\mathbb{E}[Regret_T] \leq O(T^{2/3}(KR_{Sq}(T))^{1/3})$.

Proof Strategy:

1. **Decompose** regret into "Exploration cost" and "Estimation cost".
2. **Relate** the "Estimation cost" to the oracle's prediction error on *all* arms.
3. **Link** the error on *all* arms to the error on *observed* arms (using the ϵ probability).

Step 1: Decomposition

Consider the *expected* instantaneous regret (conditioned on history). Since $\mathbb{E}[r(t)|x(t), i(t)] = f^*(x(t), i(t))$, we analyze the gap in expected rewards:

$$\text{Gap}_t = f^*(x(t), i^*(t)) - f^*(x(t), i(t))$$

Taking the expectation over the algorithm's randomness at step t:

$$\begin{aligned}\mathbb{E}_t[\text{Gap}_t] &= \epsilon \mathbb{E}_{i\sim U}[f^*(x(t), i^*(t)) - f^*(x(t), i)] \\ &\quad + (1-\epsilon)(f^*(x(t), i^*(t)) - f^*(x(t), \hat{i}(t))) \\ &\leq \epsilon + (f^*(x(t), i^*(t)) - f^*(x(t), \hat{i}(t)))\end{aligned}$$

- The first term is the cost of forced exploration (ϵ).
- The second term is the sub-optimality of our greedy choice $\hat{i}(t)$.

Step 2: Bounding the Greedy Gap

We need to bound $f^*(x(t), i^*(t)) - f^*(x(t), \hat{i}(t))$. Recall $\hat{i}(t)$ maximizes $\hat{f}_t$, so $\hat{f}_t(x(t), \hat{i}(t)) \geq \hat{f}_t(x(t), i^*(t))$.
We can rewrite the gap as:

$$\begin{aligned} f^*(x(t), i^*(t)) - f^*(x(t), \hat{i}(t)) = {} & f^*(x(t), i^*(t)) - \hat{f}_t(x(t), i^*(t)) \\ & + \underbrace{\hat{f}_t(x(t), i^*(t)) - \hat{f}_t(x(t), \hat{i}(t))}_{\leq 0} \\ & + \hat{f}_t(x(t), \hat{i}(t)) - f^*(x(t), \hat{i}(t)) \end{aligned}$$

Using triangle inequality:

$$\leq |f^*(x(t), i^*(t)) - \hat{f}_t(x(t), i^*(t))| + |\hat{f}_t(x(t), \hat{i}(t)) - f^*(x(t), \hat{i}(t))|$$

This is bounded by the sum of prediction errors on all arms:

$$\leq \sum_{i \in \mathcal{A}} |\hat{f}_t(x(t), i) - f^*(x(t), i)|$$

Step 3: From L_1 Error to Squared Error

We have the gap bounded by the L_1 error sum. Using Cauchy-Schwarz $(\sum_{j=1}^{K} z_j)^2 \leq K \sum_{j=1}^{K} z_j^2$:

$$(f^*(x(t), i^*(t)) - f^*(x(t), \hat{i}(t)))^2 \leq \left(\sum_{i \in \mathcal{A}} |\hat{f}_t(x(t), i) - f^*(x(t), i)| \right)^2$$
$$\leq K \sum_{i \in \mathcal{A}} (\hat{f}_t(x(t), i) - f^*(x(t), i))^2$$

Let's denote the total squared estimation error across all arms as:

$$L_t(\hat{f}_t) = \sum_{i \in \mathcal{A}} (\hat{f}_t(x(t), i) - f^*(x(t), i))^2$$

$$\implies f^*(x(t), i^*(t)) - f^*(x(t), \hat{i}(t)) \leq \sqrt{K \cdot L_t(\hat{f}_t)}$$

Step 4: Comparison to uniform exploration

The Oracle optimizes loss on the *observed* arm $i(t)$. How does this relate to the total error $L_t(\hat{f}_t)$?

The algorithm chooses $i(t)$ from a distribution π_t where $\pi_t(i) \geq \frac{\epsilon}{K}$ for all $i \in \mathcal{A}$.

Consider the expected squared error of the oracle at step t:

$$\begin{aligned}
&\mathbb{E}_{i(t)\sim\pi_t}[(\hat{f}_t(x(t), i(t)) - f^*(x(t), i(t)))^2] \\
&= \sum_{i\in\mathcal{A}} \pi_t(i)(\hat{f}_t(x(t), i) - f^*(x(t), i))^2 \\
&\geq \frac{\epsilon}{K} \sum_{i\in\mathcal{A}} (\hat{f}_t(x(t), i) - f^*(x(t), i))^2 = \frac{\epsilon}{K} L_t(\hat{f}_t)
\end{aligned}$$

since $\pi_t(i) \geq \frac{\epsilon}{K}$.

Therefore:

$$L_t(\hat{f}_t) \leq \frac{K}{\epsilon} \mathbb{E}_{i(t)}[(\hat{f}_t(x(t), i(t)) - f^*(x(t), i(t)))^2]$$

Step 5: Summing Over Time

We found: $\mathbb{E}_t[\text{Regret}_t] \leq \epsilon + \sqrt{K \cdot L_t(\hat{f}_t)}$. Summing over T:

$$\mathbb{E}[\text{Regret}_T] \leq \epsilon T + \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{K L_t(\hat{f}_t)}\right]$$

By Jensen's Inequality (concavity of $\sqrt{\cdot}$) and then Cauchy-Schwarz ($\sum \sqrt{z_t} \leq \sqrt{T \sum z_t}$):

$$\mathbb{E}[\text{Regret}_T] \leq \epsilon T + \sqrt{T \cdot K \cdot \sum_{t=1}^{T} \mathbb{E}[L_t(\hat{f}_t)]}$$

Substitute our bound $L_t \leq \frac{K}{\epsilon}\text{ObservedError}_t$:

$$\leq \epsilon T + \sqrt{T \cdot K \cdot \frac{K}{\epsilon} \sum_{t=1}^{T} \mathbb{E}[(\hat{f}_t(x(t), i(t)) - f^*(x(t), i(t)))^2]}$$

Step 6: Using the Oracle Guarantee

The term inside the square root corresponds to the Oracle's performance.

Recall: The Oracle bounds $\sum(\hat{f}_t - r(t))^2$, not $(\hat{f}_t - f^*)^2$.

However, under the realizability assumption (where $r(t) = f^*(x(t), i(t)) + \text{noise}$):

$$\mathbb{E}[(\hat{y} - y)^2] - \mathbb{E}[(f^*(x) - y)^2] = \mathbb{E}[(\hat{y} - f^*(x))^2]$$

(The noise variance cancels out in the regret difference).

So the Oracle property implies:

$$\sum_{t=1}^{T} \mathbb{E}[(\hat{f}_t(x(t), i(t)) - f^*(x(t), i(t)))^2] \leq R_{\mathsf{Sq}}(T)$$

Substituting this back:

$$\mathbb{E}[\text{Regret}_T] \leq \epsilon T + \sqrt{\frac{K^2 T}{\epsilon} R_{\mathsf{Sq}}(T)}$$

Step 7: Balancing ϵ

We have the bound: $B(\epsilon) = \epsilon T + C\epsilon^{-1/2}$ where $C = K\sqrt{TR_{\mathsf{Sq}}}$.
To minimize this, set the derivative to 0:

$$T - \frac{1}{2}C\epsilon^{-3/2} = 0 \implies \epsilon^{3/2} = \frac{C}{2T} \implies \epsilon \approx \left(\frac{K\sqrt{R_{\mathsf{Sq}}}}{\sqrt{T}}\right)^{2/3}$$

This yields the final rate:

$$\mathrm{Regret}_T \leq O(T^{2/3}(KR_{\mathsf{Sq}}(T))^{1/3})$$

Conclusion:

- Contextual Bandits are solvable given an Online Regression Oracle.
- The rate is (at best) $T^{2/3}$, which is suboptimal compared to $T^{1/2}$.
- **Next:** how to do online regression?
- **Later:** SquareCB (achieving $T^{1/2}$ via optimal exploration).

A natural baseline learner

- Suppose that $|\mathcal{F}|$ is finite and not too large.
  - (It is usually safe to think of $\mathcal{F}$ as finite but possibly large relative to T, possibly superpolynomial size in T.)
- Natural strategy: forecaster = ERM =

$$\hat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{s<t} (r(s) - f(x(s), i(s)))^2$$

- Question: how good of a regret bound does this forecaster satisfy for online least squares?

Comment on baseline learner

- It can be made to suffer regret $\Omega(|\mathcal{F}|)$. (Why?)
- This is bad if $|\mathcal{F}| = T$.
- Question 2: is this avoidable with a better strategy?
- Question 3: for ERM, does it obtain $o(T)$ regret if $|\mathcal{F}| = O(1)$?