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

Lecture 10: Optimal methods for contextual bandits

Instructor: Frederic Koehler

Announcements

- No office hours this Wednesday (Wed Feb 11) due to a conflict, however we can have office hours by appointment on Friday or Monday to compensate.
- FYI: I will soon upload HW2 (bandits and online forecasting), which will be due in a couple of weeks.

References

Chapter 3.5 of Foster-Rakhlin notes.
My course notes from last year (DATA 37200, 2025).

Recap: Contextual Bandits

- **Setting:**
  - At each round $t = 1, \ldots, T$:
  - Nature reveals context $x_t \in \mathcal{X}$.
  - Learner chooses policy/action $i_t \in [K]$.
  - Nature reveals reward $r_t \in [0, 1]$.
- **Goal:** Minimize (pseudo)regret

$$\mathrm{Reg}_{CB}(T) = \sum_{t=1}^{T} (f^*(x_t, i^*) - f^*(x_t, i_t))$$

where $i^* = \arg\max_i f^*(x_t, i)$ is the optimal action and f^* is the true mean reward.

Previous Approach: ϵ-Greedy

- **Strategy:**
  - Use an oracle to estimate rewards: $\hat{f}_t(x_t, i)$.
  - With probability $1 - \epsilon$: Play greedy $\hat{i}_t = \arg\max_i \hat{f}_t(x_t, i)$.
  - With probability ϵ: Explore uniformly $i_t \sim \text{Unif}([K])$.
- **Limitation:**
  - Exploration is uniform.
  - Does not distinguish between "slightly suboptimal" and "very terrible" arms.
  - We want to explore arms with **small estimated gaps** more than those with large gaps.
- We will see later how to solve this in linear case with LinUCB, but first learn an approach which is closer in spirit to ϵ-greedy and Thompson sampling.

Inverse Gap Weighting (SquareCB)

- **Idea:** Probability of choosing an arm should decrease as the estimated suboptimality gap increases.
- Parameter γ is analogous to $1/\epsilon$ in ϵ-greedy. Optimal tuning will usually lead to $\gamma \to \infty$ as $T \to \infty$.
- **Algorithm SquareCB(γ):**
  1. Receive context x_t.
  2. Obtain forecast $\hat{f}_t(i)$ for all arms $i \in [K]$.
  3. Identify greedy action: $\hat{i}_t = \arg\max_i \hat{f}_t(i)$.
  4. Compute assignment probabilities $p_t(i)$:

$$p_t(i) = \frac{1}{\lambda_t + 2\gamma\left(\hat{f}_t(\hat{i}_t) - \hat{f}_t(i)\right)}$$

     where $\lambda_t \in [1, K]$ is a normalization constant s.t. $\sum p_t(i) = 1$.
  5. Play $i_t \sim p_t$.

Algorithm Intuition

- Define the estimated gap: $\hat{\Delta}_t(i) = \hat{f}_t(\hat{i}_t) - \hat{f}_t(i)$.
- **Selection Rule:**

$$p_t(i) = \frac{1}{\lambda_t + 2\gamma\hat{\Delta}_t(i)}$$

- **Small Gap ($\hat{\Delta} \approx 0$):** $p_t(i)$ is large. We explore arms that look almost as good as the best.
- **Large Gap ($\hat{\Delta}$ is large):** $p_t(i)$ is small. We rarely play arms that the forecaster thinks are bad.
- So if the forecaster thinks the gap is large, we mostly play greedily. Allows us to be more "aggressive" than ϵ-greedy; proof will need to worry about *what happen when the forecast is inaccurate*.

Theorem: Regret Bound

Theorem

Let $Reg_{Sq}(T) = \sum_{t=1}^{T}(\hat{f}_t(x_t, i_t) - f^*(x_t, i_t))^2$ *be the cumulative square loss of the predictor.*
The Inverse Gap Weighting algorithm with optimal choice of γ *guarantees:*

$$\mathbb{E}[Reg_{CB}(T)] \leq 2\sqrt{KT \cdot \mathbb{E}[Reg_{Sq}(T)]}$$

Key lemma: Define the expected instantaneous regret at time t

$$\text{InstRegret}_t = \mathbb{E}_{i \sim p_t}[f^*(i_t^*) - f^*(i_t)]$$

Then

$$\text{InstRegret}_t \leq \frac{K}{\gamma} + \gamma \mathbb{E}_{i \sim p_t}\left[(f^*(i) - \hat{f}_t(i))^2\right]$$

Proof of theorem given lemma

Since the lemma says

$$\text{InstRegret}_t \leq \frac{K}{\gamma} + \gamma \mathbb{E}_{i \sim p_t} \left[(f^*(i) - \hat{f}_t(i))^2 \right]$$

we can sum over all $t = 1, \ldots, T$ and take expectation to find

$$\mathbb{E}[\text{Reg}_{CB}(T)] \leq \frac{KT}{\gamma} + \gamma \mathbb{E}[\text{Reg}_{Sq}(T)].$$

Optimizing γ we find

$$\mathbb{E}[\text{Reg}_{CB}(T)] \leq 2\sqrt{KT\mathbb{E}[\text{Reg}_{Sq}(T)]}$$

proving the theorem.

Proof of lemma

Goal: want to prove

$$\text{InstRegret}_t = \mathbb{E}_{i\sim p_t}[f^*(i_t^*) - f^*(i_t)] \leq \frac{K}{\gamma} + \gamma \mathbb{E}_{i\sim p_t}\left[(f^*(i) - \hat{f}_t(i))^2\right]$$

Step 1: Decompose the term inside the expectation:

$$f^*(i^*) - f^*(i) = \underbrace{f^*(i^*) - \hat{f}_t(i^*)}_{\text{(I) Error at } i^*} + \underbrace{\hat{f}_t(i^*) - \hat{f}_t(\hat{i}_t)}_{\text{(II)}\leq 0 \text{ (Greedy)}}$$
$$+ \underbrace{\hat{f}_t(\hat{i}_t) - \hat{f}_t(i)}_{\text{(III) Estimated Gap } \hat{\Delta}(i)} + \underbrace{\hat{f}_t(i) - f^*(i)}_{\text{(IV) Error at } i}$$

Here (IV) is the error which the forecaster is trying to minimize (in its regret objective). Why are (I-III) controlled?

A key tool: The AM-GM Inequality

AM-GM Inequality (Young's Inequality): $ab \leq a^2/2 + b^2/2$, so for any $x \in \mathbb{R}$ and $\gamma > 0$:

$$x = \frac{\gamma^{1/2} x}{\gamma^{1/2}} \leq \frac{\gamma x^2}{2} + \frac{1}{2\gamma}$$

Apply this to the "easy" term (IV), where $x = \hat{f}_t(i) - f^*(i)$:

$$\hat{f}_t(i) - f^*(i) \leq \frac{\gamma}{2}(\hat{f}_t(i) - f^*(i))^2 + \frac{1}{2\gamma}$$

Substituting this back into our regret bound:

$$\text{InstRegret}_t \leq (I) + (II) + (III) + \frac{\gamma}{2} \underbrace{\sum_i p_t(i)(\hat{f}_t(i) - f^*(i))^2}_{\mathbb{E}[\text{SquareError}]} + \frac{1}{2\gamma}$$

Remains to control $(I) + (II)$ and (III) similarly.

Proof Step 3: Bounding the Gap Term (III)

We have:

$$\mathbb{E}(III) = \underbrace{\sum_{i=1}^{K} p_t(i)\hat{\Delta}(i)}_{\text{Gap Term}}$$

Recall $p_t(i) = \frac{1}{\lambda + 2\gamma\hat{\Delta}(i)}$ so

$$\sum_{i=1}^{K} p_t(i)\hat{\Delta}(i) = \sum_{i=1}^{K} \frac{\hat{\Delta}(i)}{\lambda + 2\gamma\hat{\Delta}(i)}$$

Multiply top and bottom by $1/2\gamma$:

$$= \frac{1}{2\gamma} \sum_{i=1}^{K} \frac{2\gamma\hat{\Delta}(i)}{\lambda + 2\gamma\hat{\Delta}(i)} \leq \frac{K-1}{2\gamma}$$

Summary so far

$$\text{InstRegret}_t \leq (I) + (II) + \frac{\gamma}{2} \underbrace{\sum_i p_t(i)(\hat{f}_t(i) - f^*(i))^2}_{\mathbb{E}[\text{SquareError}]} + \frac{K}{2\gamma}$$

and remains to control expected contribution of

$$\underbrace{f^*(i^*) - \hat{f}_t(i^*)}_{\text{(I) Error at } i^*} + \underbrace{\hat{f}_t(i^*) - \hat{f}_t(\hat{i}_t)}_{\text{(II)}\leq 0 \text{ (Greedy)}} = f^*(i^*) - \hat{f}_t(\hat{i}_t)$$

Obvious approach: use that $(II) \leq 0$ and try to handle (I) by itself. (How we handled ϵ-greedy.) Not good here.
Smart approach: take advantage of cancellations.

Key step in key lemma: AM-GM trick

Analogous to ϵ-greedy analysis, we insert $p_t(i^*)$. Also use AM-GM

$$\underbrace{f^*(i^*) - \hat{f}_t(i^*)}_{\text{(I) Error at } i^*} \leq \frac{\gamma p_t(i^*)}{2}(f^*(i^*) - \hat{f}_t(i^*))^2 + \frac{1}{2\gamma p_t(i^*)}$$

$$\leq \frac{\gamma}{2}\mathbb{E}_{i\sim p_t}\left[(f^*(i) - \hat{f}_t(i))^2\right] + \frac{1}{2\gamma p_t(i^*)}$$

Recall

$$\frac{1}{2\gamma p_t(i^*)} = \frac{\lambda_t + 2\gamma(\hat{f}_t(\hat{i}_t) - \hat{f}_t(i^*))}{2\gamma}$$

so

$$\frac{1}{2\gamma p_t(i^*)} + \underbrace{\hat{f}_t(i^*) - \hat{f}_t(\hat{i}_t)}_{\text{(II)}\leq 0 \text{ (Greedy)}} = \frac{\lambda_t}{2\gamma} \leq \frac{K}{2\gamma}$$

hence

$$(I) + (II) \leq \frac{\gamma}{2}\mathbb{E}_{i\sim p_t}\left[(f^*(i) - \hat{f}_t(i))^2\right] + \frac{K}{2\gamma},$$

Summary

Combining guarantees on $(I) + (II)$ with those for $(III) + (IV)$ we get

$$\text{InstRegret}_t \leq \gamma \underbrace{\sum_i p_t(i)(\hat{f}_t(i) - f^*(i))^2}_{\mathbb{E}[\text{SquareError}]} + \frac{K}{\gamma}$$

which is exactly the lemma statement. And we already proved theorem given lemma.

Summary so far

- ϵ-**Greedy:** Simple but inefficient exploration (uniform).
- **Inverse Gap Weighting:** Exploration probabilities scale inversely with the estimated gap.
- **Result:** Efficient reduction from Contextual Bandits to Online Regression.
- **Outcome:** Fast rates ($\sqrt{T}$) if we have a good predictor , without needing to know the exact reward structure.
  - $\sqrt{KT \log |F|}$ for finite classes
  - $\sqrt{KTd \log T}$ for linear models in $\mathbb{R}^d$
  - Essentially minimax optimal for linear models with $K = O(1)$.

Another example application

- The very first problem we studied was MAB for which UCB3 obtained regret $O(\sqrt{KT \log T})$.
- Obviously, MAB is a special case of contextual bandits.
- Function class is space of constant functions $\mathcal{F} = [0, 1]^K$.
- SquareCB + exponential weights can achieve $O(K\sqrt{T \log T})$ regret with a different algorithmic approach than ucb.

Motivation for LinUCB

- SquareCB + ridge forecaster is optimal for linear classes except in dependence on K.
- I.e. suboptimal in dependence on number of arms.
- The forecaster does not even know there are arms, so the suboptimality here is more to do with the selection of arms via inverse gap weighting. (Wastes too much time exploring different arms when K is large.)
- The fix is LinUCB (think: UCB3 + ridge forecaster) and relatives.

Intuition for the improvement

- Suppose K is large, e.g., $K = d^{100}$. Then $\sqrt{KTd \log T}$ is quite suboptimal.
- LinUCB attains regret $d\sqrt{T \log T}$ regardless of K.
- Sidenote: more complex versions of LinUCB can guarantee $\sqrt{dTpolylog(KT)}$ regret (see references in Ch. 22 of Lattimore-Svesespari, Ch. 21 regarding Kiefer-Wolfowitz design).
- Why should K-independent regret be possible for linear case, when $\sqrt{KT}$ is a lower bound for MAB ?
- Think of the case $d = 1$ or 2... not that many possible "truly different" arms.
- LinUCB understands how to share findings across different arms efficiently.

The Linear Contextual Bandit Setting

- **Assumption:** The true mean reward function f^* is linear in a known feature map $\phi : \mathcal{X} \times [K] \to \mathbb{R}^d$.

$$f^*(x_t, i) = \langle \phi(x_t, i), \theta^* \rangle$$

  where $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector.
- **Notation:** Let $\phi_t(i) := \phi(x_t, i)$.
- **Goal:** Learn θ^* while minimizing regret.
- Unlike the generic setting, we can extrapolate: learning about arm i tells us about arm j if their feature vectors $\phi_t(i)$ and $\phi_t(j)$ are aligned.

Ridge Regression & Covariance

- At round t, we estimate θ^* using Ridge Regression on the history of played actions $(\phi_\tau(i_t), r_t)_{\tau<t}$.
- **Covariance Matrix (The "Confidence" Metric):**

$$A_t = \lambda I + \sum_{\tau=1}^{t-1} \phi_\tau(i_t)\phi_\tau(i_t)^\top$$

- **Ridge Estimator:**

$$\hat{\theta}_t = A_t^{-1} \sum_{\tau=1}^{t-1} \phi_\tau(i_t) r_t$$

- A_t captures how much we have explored different directions in the feature space $\mathbb{R}^d$.

LinUCB Algorithm

- **Principle:** Optimism in the Face of Uncertainty.
- Construct a confidence interval for the reward of each arm i:

$$\text{UCB}_t(i) = \underbrace{\langle \hat{\theta}_t, \phi_t(i) \rangle}_{\text{Estimate}} + \underbrace{\beta_t \sqrt{\phi_t(i)^\top A_t^{-1} \phi_t(i)}}_{\text{Exploration Bonus}}$$

- **Algorithm:**
  1. Observe context x_t.
  2. Compute A_t and $\hat{\theta}_t$.
  3. Select arm $i_t = \arg\max_{i \in [K]} \text{UCB}_t(i)$.
  4. Observe reward r_t and update Σ_{t+1}.
- The term $\|\phi_t(i)\|_{A_t^{-1}} = \sqrt{\phi_t(i)^\top A_t^{-1} \phi_t(i)}$ measures uncertainty in the direction of arm i.

Theoretical Guarantee

Theorem (Regret of LinUCB)

With probability $1-\delta$*, if we choose the width parameter* $\beta_t \approx \sqrt{d\log(t)+\log(1/\delta)}$*, LinUCB satisfies:*

$$Reg_{CB}(T) \leq O\left(d\sqrt{T\log(T)}\right)$$

- **Key change:** The bound depends on dimension d, but not the number of arms K.
- SquareCB guarantees $\approx \sqrt{dKT}$. LinUCB yields $\approx d\sqrt{T}$.

Next time/conclusions

- Will finish LinUCB analysis.
- Move on to last third of the class: RL and multiplayer games.
- Interaction between players, interaction between player and environment.