Bandits: Explore-Then-Commit and ε-greedy

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning
Fall 2022

Today

- Recap
- Explore-then-commit (ETC)
- ε-greedy

Recap

- Reinforcement learning is an *interactive* form of machine learning
  - Applicable whenever you want to learn to do something better
  - One component is learning while acting: exploration vs exploitation
  - Other component is optimization
- Multi-armed bandits (or MAB or just bandits)
  - Exemplify first component (exploration vs exploitation)
  - Pure greedy not much better than pure exploration (linear regret)
- Today: let's do better than linear regret!

Notes from yesterday

1. $$\text{Regret}_T = T\mu^\star - \sum_{t=0}^{T-1} \mu_{a_t} = \sum_{t=0}^{T-1} (\mu^\star - \mu_{a_t})$$

Expected regret at time t
given that you chose arm a_t

2. Recall $\text{Regret}_T = \Omega(T)$, i.e., linear regret

$\Rightarrow$ for some $c > 0$ and T_0, $\text{Regret}_T \geq cT \quad \forall T \geq T_0$

(and $\text{Regret}_T = O(T)$ means same except with $\leq cT$)

3. Why is linear regret bad? $\Rightarrow$ *average regret* $:= \dfrac{\text{Regret}_T}{T} \not\to 0$

Today

- Recap
- Explore-then-commit (ETC)
- ε-greedy

What we learned last lecture:

Lesson from pure greedy: exploring each arm once is not enough
Lesson from pure exploration: exploring each arm too much is bad too

Let’s allow both, and see how best to trade them off

Plan: (1) try each arm <u>multiple</u> times, (2) compute the empirical mean of each arm, (3) commit to the one that has the highest empirical mean

Explore-Then-Commit (ETC)

N_e = Number of explorations

Algorithm hyper parameter $N_e < T/K$ (we assume $T >> K$)

For $k = 1, \ldots, K$: (# Exploration phase)

Pull arm k N_e times to observe $\{r_i^{(k)}\}_{i=1}^{N_e} \sim \nu_k$

Calculate arm k's empirical mean: $\hat{\mu}_k = \frac{1}{N_e} \sum_{i=1}^{N_e} r_i^{(k)}$

For $t = N_e K, \ldots, (T-1)$: (# Exploitation phase)

Pull the best empirical arm $a_t = \arg\max_{i \in [K]} \hat{\mu}_i$

Q: how to set N_e?

Regret Analysis Strategy

1. Calculate regret during exploration stage
2. Quantify error of arm mean estimates at end of exploration stage
3. Using step 2, calculate regret during exploitation stage

   (Actually, will only be able to upper-bound total regret in steps 1-3)

4. Minimize our upper-bound over N_{e}

But First… An Important Inequality

<u>Hoeffding inequality</u>

Given N i.i.d samples $\{r_i\}_{i=1}^N \sim \nu \in \Delta([0,1])$ with mean μ, let $\hat{\mu} := \frac{1}{N}\sum_{i=1}^{N} r_i$.

Then with probability at least $1-\delta$,

$$\left|\hat{\mu}-\mu\right| \le \sqrt{\frac{\ln(2/\delta)}{2N}}$$

- Why is this useful? Quantify error of arm mean estimates at end of exploration stage (if all estimates are close, arm we commit to must be close to best)
- Why is this true? Full proof beyond course scope, but intuition easier…

Intuition Behind Hoeffding

Hoeffding inequality: sample mean of N i.i.d. samples on $[0,1]$ satisfies

$$\left|\hat{\mu} - \mu\right| \leq \sqrt{\frac{\ln(2/\delta)}{2N}} \text{ w/p } 1 - \delta$$

Think of as finite-sample (and conservative) version of Central Limit Theorem (CLT):

- CLT $\Rightarrow \hat{\mu} - \mu \approx$ Gaussian w/ mean 0 and standard deviation $\propto \sqrt{1/N}$
- CLT standard deviation explains the Hoeffding denominator
- Numerator is because Gaussian has double-exponential tails, i.e., probability of a deviation from the mean by x scales roughly like e^{-x^2}, which, when inverted (i.e., set $\delta = e^{-x^2}$ and solve for x) gives $x = \sqrt{\ln(1/\delta)}$
- Don't worry too much about the extra 2's… CLT is only approximate!

Back to Regret Analysis of ETC

1. Calculate regret during exploration stage

$$\text{Regret}_{N_\text{e}K} \leq N_\text{e}K \text{ with probability } 1$$

2. Quantify error of arm mean estimates at end of exploration stage

   a) Hoeffding $\Rightarrow \mathbb{P}\left(|\hat{\mu}_k - \mu_k| \leq \sqrt{\ln(2/\delta)/2N_\text{e}}\right) \geq 1 - \delta$

   $\mathbb{P}(\forall k, A_1^c, \ldots, A_K^c) \geq 1 - \sum_{k=1}^{K} \mathbb{P}(A_k)$

   $\Updownarrow$

   b) Recall Union/Boole/Bonferroni bound: $\mathbb{P}(\text{any of } A_1, \ldots, A_K) \leq \sum_{k=1}^{K} \mathbb{P}(A_k)$

   c) $\delta \to \delta/K$, Union bound with $A_k = \left\{ |\hat{\mu}_k - \mu_k| > \sqrt{\ln(2K/\delta)/2N_\text{e}} \right\}$, and Hoeffding:

$$\Rightarrow \mathbb{P}\left(\forall k, |\hat{\mu}_k - \mu_k| \leq \sqrt{\ln(2K/\delta)/2N_\text{e}}\right) \geq 1 - \delta$$

Regret Analysis of ETC (cont'd)

2. Quantify error of arm mean estimates at end of exploration stage

$$\mathbb{P}\left(\forall k,\ |\hat{\mu}_k - \mu_k| \leq \sqrt{\ln(2K/\delta)/2N_{\mathrm{e}}}\right) \geq 1 - \delta$$

3. Using step 2, calculate regret during exploitation stage:

Denote (apparent) best arm after exploration stage by $\hat{k}$ and actual best arm by $k^\star$

$$\text{regret at each step of exploitation phase} = \mu_{k^\star} - \mu_{\hat{k}}$$

Regret Analysis of ETC (cont'd)

4. From steps 1-3: with probability $1 - \delta$,

$$\text{Regret}_T \leq N_\text{e}K + T\sqrt{2\ln(2K/\delta)/N_\text{e}}$$

Take any N_e so that $N_\text{e} \to \infty$ and $N_\text{e}/T \to 0$ (e.g., $N_\text{e} = \sqrt{T}$): sublinear regret!

Minimize over N_e: (won't bore you with algebra)

$$\text{optimal } N_\text{e} = \left(\frac{T\sqrt{\ln(2K/\delta)/2}}{K}\right)^{2/3}$$

(A bit more algebra to plug optimal N_e into Regret_T equation above)

$$\Rightarrow \text{Regret}_T \leq 3T^{2/3}(K\ln(2K/\delta)/2)^{1/3}$$

Today

- Recap
- Explore-then-commit (ETC)
- ε-greedy

ε-greedy

ETC very abrupt (huge difference between exploration and exploitation stages)

ε-greedy like a smoother version of ETC:
at *every* step, do pure greedy w/p $1-\varepsilon$, and do pure exploration w/p ε

Initialize $\hat{\mu}_0 = \cdots = \hat{\mu}_K = 1$
For $t = 0, \ldots, T-1$:
Sample $E_t \sim$ Bernoulli(ε)
If $E_t = 1$, choose $a_t \sim$ Uniform$(1, \ldots, K)$ (pure explore)
If $E_t = 0$, choose $a_t = \arg\max_{k \in \{1,\ldots,K\}} \hat{\mu}_k$ (pure exploit)
Update $\hat{\mu}_{a_t}$

ε-greedy (cont'd)

Can also allow ε to depend on t, usually so that it decreases: the more learned by time t, the less exploration needed at/after time t

It turns out that ε-greedy with $\varepsilon_t = \left(\frac{K \ln(t)}{t}\right)^{1/3}$ also achieves

$$\text{Regret}_t = \tilde{O}(t^{2/3}K^{1/3}),$$

where $\tilde{O}(\,\cdot\,)$ hides logarithmic factors

- Regret rate (ignoring log factors) is the same as ETC, but holds for <u>all</u> t, not just the full time horizon T
- Nothing in ε-greedy (including ε_t above) depends on T, so don't need to know horizon!

Today

- Recap
- Explore-then-commit (ETC)
- ε-greedy

Today's summary:

Explore-then-commit and ε-greedy:

- balance exploration with exploitation
- Achieve sublinear regret of $\tilde{O}(T^{2/3}K^{1/3})$
- Exploration is non-adaptive (bad)

Next time:

- Upper Confidence Bound (UCB) explores adaptively
- Achieves regret $\tilde{O}(\sqrt{TK})$

1-minute feedback form: https://forms.gle/2mKHGRMCpFTRMQqd8