RL seminar # 5: Multi-armed bandits

Eugene Golikov

MIPT, Deep learning lab & iPavlov.ai

December 2, 2017

Problem setup

Full RL:

We have:

- Episode length T (let it be finite)
- State space $\mathcal{S}$, action space $\mathcal{A}$
- Initial state distribution $p(s_0)$ (unknown)
- Transition distribution $p(s'|s, a)$ (unknown)
- Reward distribution $p(r|s, a)$ (unknown)

Problem setup

Full RL:

We have:

- Episode length T (let it be finite)
- State space $\mathcal{S}$, action space $\mathcal{A}$
- Initial state distribution $p(s_0)$ (unknown)
- Transition distribution $p(s'|s, a)$ (unknown)
- Reward distribution $p(r|s, a)$ (unknown)

Goal: derive policy $\pi^*(a|s)$, that maximizes the expected total reward over episode:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau \sim \pi} R_T, \quad \text{where } R_T = \sum_{t=1}^{T} r_t$$

Problem setup

Multi-armed bandit:

A special case: $|\mathcal{S}| = 1$.
So we have:

- Number of trials T (let it be finite)
- Action space $\mathcal{A}$ (let it be finite)
- Reward distribution $p(r|a)$ (unknown)

Goal: derive policy $\pi^*(a)$, that maximizes the expected total reward over all amount of trials:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau \sim \pi} R_T, \quad \text{where } R_T = \sum_{t=1}^{T} r_t$$

Goal: derive policy $\pi^*(a)$, that maximizes the expected total reward over all amount of trials:

$$\pi^* = \arg\max_\pi \mathbb{E}_{\tau\sim\pi} R_T, \quad \text{where } R_T = \sum_{t=1}^{T} r_t$$

However, since trials are independent,

$$\mathbb{E}_{\tau\sim\pi} R_T = \mathbb{E}_{\tau\sim\pi} \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} \mathbb{E}_{a_t\sim\pi(a_t)} \mathbb{E}_{r_t\sim p(r_t|a_t)} r_t = T\mathbb{E}_{a\sim\pi(a)} \mathbb{E}_{r\sim p(r|a)} r$$

Goal: derive policy $\pi^*(a)$, that maximizes the expected total reward over all amount of trials:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi} R_T, \quad \text{where } R_T = \sum_{t=1}^{T} r_t$$

However, since trials are independent,

$$\mathbb{E}_{\tau\sim\pi} R_T = \mathbb{E}_{\tau\sim\pi} \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} \mathbb{E}_{a_t\sim\pi(a_t)} \mathbb{E}_{r_t\sim p(r_t|a_t)} r_t = T\mathbb{E}_{a\sim\pi(a)} \mathbb{E}_{r\sim p(r|a)} r$$

Action-value function:

Let $Q^*(a) = \mathbb{E}_{r\sim p(r|a)} r$, then

$$\mathbb{E}_{\tau\sim\pi} R_T = T\mathbb{E}_{a\sim\pi(a)} Q^*(a)$$

Goal: derive policy $\pi^*(a)$, that maximizes the expected total reward over all amount of trials:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau \sim \pi} R_T, \quad \text{where } R_T = \sum_{t=1}^{T} r_t$$

However, since trials are independent,

$$\mathbb{E}_{\tau \sim \pi} R_T = \mathbb{E}_{\tau \sim \pi} \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} \mathbb{E}_{a_t \sim \pi(a_t)} \mathbb{E}_{r_t \sim p(r_t|a_t)} r_t = T \mathbb{E}_{a \sim \pi(a)} \mathbb{E}_{r \sim p(r|a)} r$$

Action-value function:

Let $Q^*(a) = \mathbb{E}_{r \sim p(r|a)} r$, then

$$\mathbb{E}_{\tau \sim \pi} R_T = T \mathbb{E}_{a \sim \pi(a)} Q^*(a)$$

So,

$$\pi^*(a) = \arg\max_{\pi} \mathbb{E}_{a \sim \pi(a)} Q^*(a)$$

We have:

$$\pi^*(a) = \arg\max_{\pi} \mathbb{E}_{a \sim \pi(a)} Q^*(a)$$

What is the optimal policy?

We have:

$$\pi^*(a) = \arg\max_{\pi} \mathbb{E}_{a\sim\pi(a)} Q^*(a)$$

What is the optimal policy?

It is greedy:

$$\pi^*(a) = [a = a^*], \quad a^* = \arg\max_{a\in\mathcal{A}} Q^*(a)$$

We can't derive it, since we don't know Q^*. Hence, we have to estimate it.

General training process:

Given: number of trials T

Goal: maximize the total reward R_T

1. Choose an initial estimate Q_0, set trial counter $t := 1$
2. Derive policy π_t using current estimate Q_{t-1}
3. Take an action a_t according to policy π_t; get a reward r_t
4. Correct estimate: $Q_{t-1} \to Q_t$
5. If $t < T$, $t := t + 1$ and return to point 2, else finish.

General training process:

Given: number of trials T
Goal: maximize the total reward R_T

1. Choose an initial estimate Q_0, set trial counter $t := 1$
2. Derive policy π_t using current estimate Q_{t-1}
3. Take an action a_t according to policy π_t; get a reward r_t
4. Correct estimate: $Q_{t-1} \to Q_t$
5. If $t < T$, $t := t + 1$ and return to point 2, else finish.

Questions:

1. How to choose initial estimate Q_0?
2. How to derive policy π_t?
3. How to correct an estimate?

How to correct an estimate?

Trial-average:

$$Q_t(a) := \frac{\sum_{k=1}^{t} r_k [a_k = a]}{\sum_{k=1}^{t} [a_k = a]}, \text{ if } a = a_t, \text{ else } Q_t(a) := Q_{t-1}(a)$$

Let $k_t(a) = \sum_{k=1}^{t} [a_k = a]$

Motivation:

- If $k_t(a) \to \infty$ as $t \to \infty$ for every a, Q_t converges to Q^* according to the law of big numbers

Recurrent formula:

$$Q_t(a) := Q_{t-1}(a) + \frac{1}{t}(r_t - Q_{t-1}(a)), \text{ if } a = a_t, \text{ else } Q_t(a) := Q_{t-1}(a)$$

$1/t$ — value of an observation; decreases with time

How to correct an estimate?

Exponential running average:

Motivation:

- Sometimes Q^* can change with time — unstationary problems
- In this setup, recent observations should be more valueable

Idea:

- Make value of an observation constant instead of $1/t$

$$Q_t(a) := Q_{t-1}(a) + \alpha(r_t - Q_{t-1}(a)), \text{ if } a = a_t, \text{ else } Q_t(a) := Q_{t-1}(a)$$

How to derive policy?

The main tradeoff:

- We want to have high rewards, so we want to take actions with high Q^* $\Leftarrow$ EXPLOITATION
- We want to know, what actions have higher Q^*, so we want to have better estimate Q_t
- We want to have better estimate Q_t, so we want to explore in action space $\Leftarrow$ EXPLORARTION

How to derive policy?

Greedy:

Motivation:

- Take the "best" possible action

$$\pi_t(a) = \left[a = \arg\max_a Q_t(a) \right]$$

Why bad:

- The best according to Q_t is not necessarily the best according to Q^*
- No exploration at all!

How to derive policy?

ϵ-greedy:

Motivation:

- We should try to take different actions to know, which one is best

Idea:

- Take "suboptimal" actions with some probability ϵ

$$\pi_t(a) = (1-\epsilon)\left[a = \arg\max_a Q_t(a)\right] + \frac{\epsilon}{|\mathcal{A}|}$$

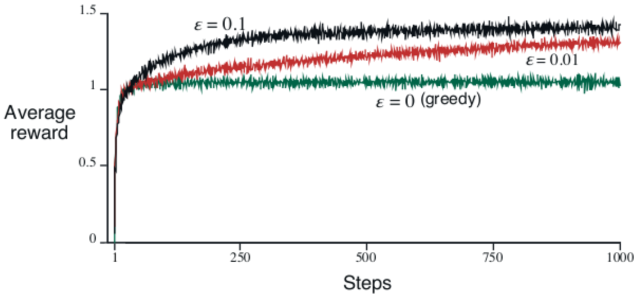
Usually we want to gradually decrease epsilon during training

Greedy – ϵ-greedy policies comparison

Experimental setup (from Suttons book):

- $|\mathcal{A}| = 10, \ T = 1000$
- $p(r|a) = N(Q^*(a), 1)$
- $Q^*(a) \sim N(0, 1)$

Plots are averaged over 2000 independent experiments



How to derive policy?

Boltzmann policy:

Motivation:

- We want to find the best action, so no need to explore actions, which is already known to be bad

Idea:

- Sample probability should be higher for actions for which the estimate Q_t is higher

$$\pi_t = \mathrm{Softmax}(Q_t / \tau)$$

$\tau \to 0$ — greedy policy, $\tau \to \infty$ — all actions are sampled uniformly.

How to derive policy?

Upper confidence bound:

Motivation:

- We want find the best action, so we want to explore actions that "has chance to be the best"

Idea:

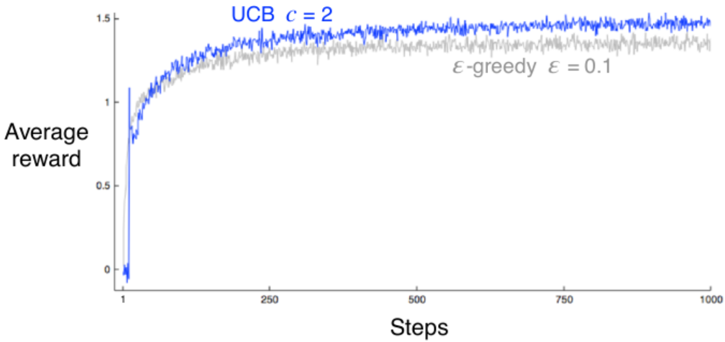
- Take an action with highest Q-value upper bound

$$\pi_t(a) = \left[a = \arg\max_a \left(Q_t(a) + c \frac{2 \ln t}{k_t(a)} \right) \right]$$

Interpretation:

- If $k_t(a)$ is small, we should explore the action a more
- $\ln t$ encourages us to re-explore actions periodically

UCB – ϵ-greedy policies comparison



How to choose initial estimate?

Realistic estimate:

Idea:

- Initialize $Q_0(a)$ with some estimate on action rewards

Optimistic estimate:

Motivation:

- We want to encourage exploration at the beginning of training

Idea:

- Initialize $Q_0(a)$ with unrealistically large constant

Then even simple greedy strategy will try all of the actions available

Realistic – optimistic estimates comparison