AIL 722: Reinforcement Learning

Lecture 39: Multi-armed bandits

Raunak Bhattacharyya

ScAI YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Recap

- Actor-Critic methods
- Continuous actions
- Montezuma’s revenge

Montezuma’s Revenge, Source: Youtube

Pitfall, Source: Youtube

Exploration vs Exploitation

How can an agent decide whether to attempt new behaviors (to discover ones with higher reward) or continue to do the best thing it knows so far?

Exploitation: Do what you think will yield the highest reward

Exploration: Do the things that you haven’t done before hoping that it will yield even higher reward

Recap & Today's Outline

- Actor-Critic methods
- Continuous actions
- Montezuma's revenge
- Multi-armed bandits
- Regret
- Optimism under uncertainty

Single-armed Bandit

One-arm bandit, Source: Youtube

Multi-armed Bandit

Multi-armed Bandit, Source: Wikipedia

- Faced repeatedly with a choice among k diff options

- Make a choice

- Receive a numerical reward chosen from a stationary prob dist depending on the action you selected

- Goal is to maximise expected total reward over some time period: eg. 1000 action selections

Section 2.1, Reinforcement Learning: An Introduction, Sutton & Barto

Formulation

$$\mathcal{A} = \{\text{pull}_1, \text{pull}_2, \ldots, \text{pull}_n\}$$

$$p(r_i = 1) = \theta_i, \quad p(r_i = 0) = 1 - \theta_i$$

$$\text{estimate } \hat{Q}_t(a) \approx Q(a) = \mathbb{E}[R(a)]$$

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^{t-1} r_i \mathbb{1}(a_i = a)$$

Example: Treating a Broken Toe

- Surgery
- Buddy taping
- Do nothing

- Surgery: $Q(a^1) = \theta_1 = 0.95$
- Buddy taping: $Q(a^2) = \theta_2 = 0.9$
- Do nothing: $Q(a^3) = \theta_3 = 0.1$

- After 6 weeks: do an X-ray to check whether healed (1) or not (0)

Example from Emma Brunskill

Greedy Approach

Take action a^1 ($r \sim$Bernoulli(0.95)), get 0, $\hat{Q}(a^1) = 0$

Take action a^2 ($r \sim$Bernoulli(0.90)), get $+1$, $\hat{Q}(a^2) = 1$

Take action a^3 ($r \sim$Bernoulli(0.1)), get 0, $\hat{Q}(a^3) = 0$

$$a_t^* = \arg\max_{a \in \mathcal{A}} \hat{Q}_t(a)$$

Regret: Greedy Approach

$$Reg(T) = T\mathbb{E}[r(a^*)] - \sum_{t=1}^{T} r(a_t)$$

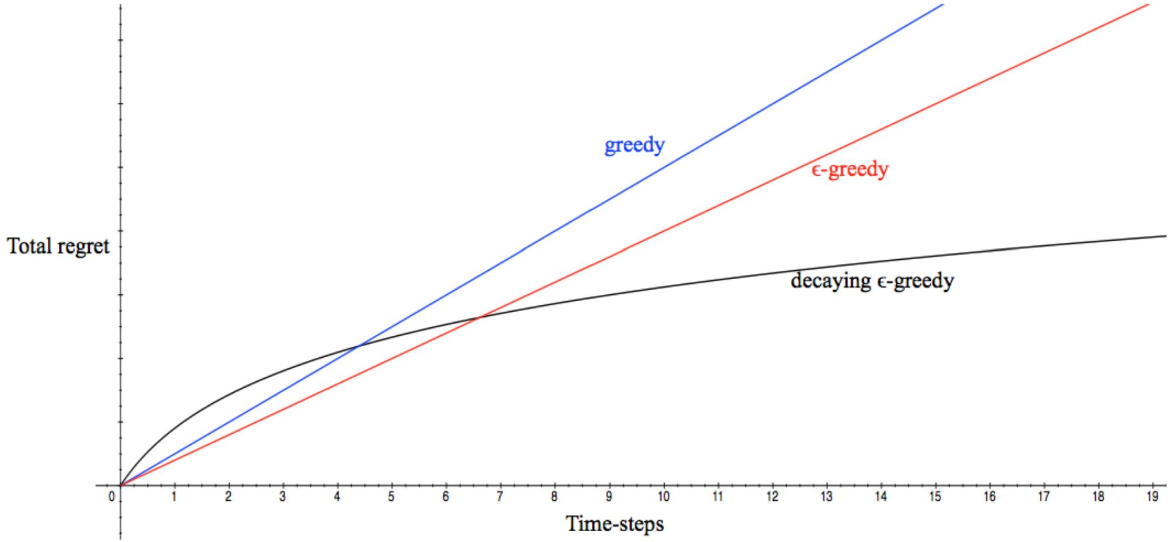
Action	Optimal Action	Observed Reward	Regret
a^1	a^1	0	0
a^2	a^1	1	0.05
a^3	a^1	0	0.85
a^2	a^1	1	0.05
a^2	a^1	0	0.05

Cannot evaluate regret in real settings because requires knowledge of the true best action

Regret: Rate of Growth

- Explore forever
- Explore never

Linear Regret



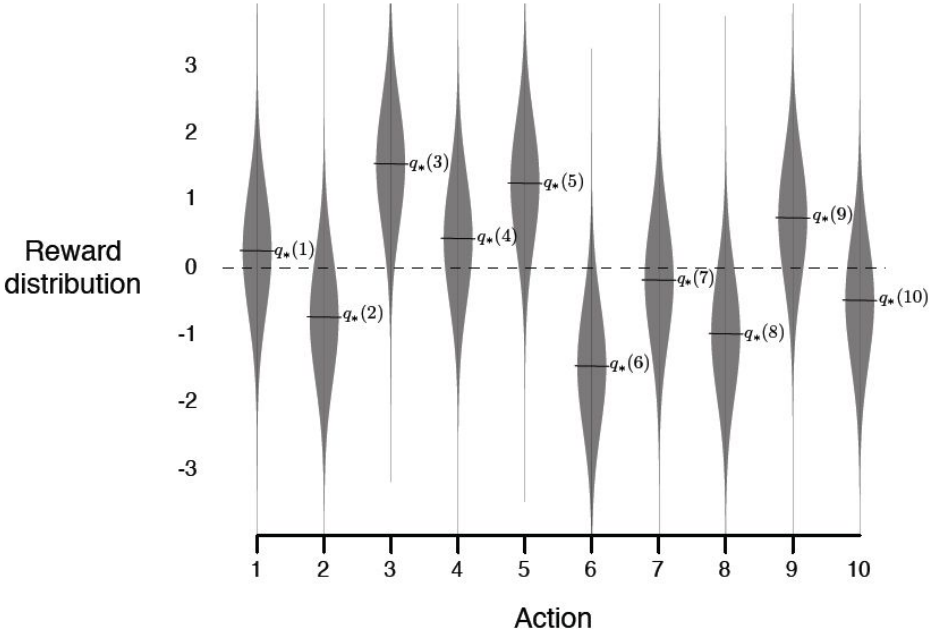
Adapted from Emma Brunskill

10-armed Testbed

- 2000 10-armed bandit problems

- For each bandit problem, set the true action values by sampling from a standard Normal distribution

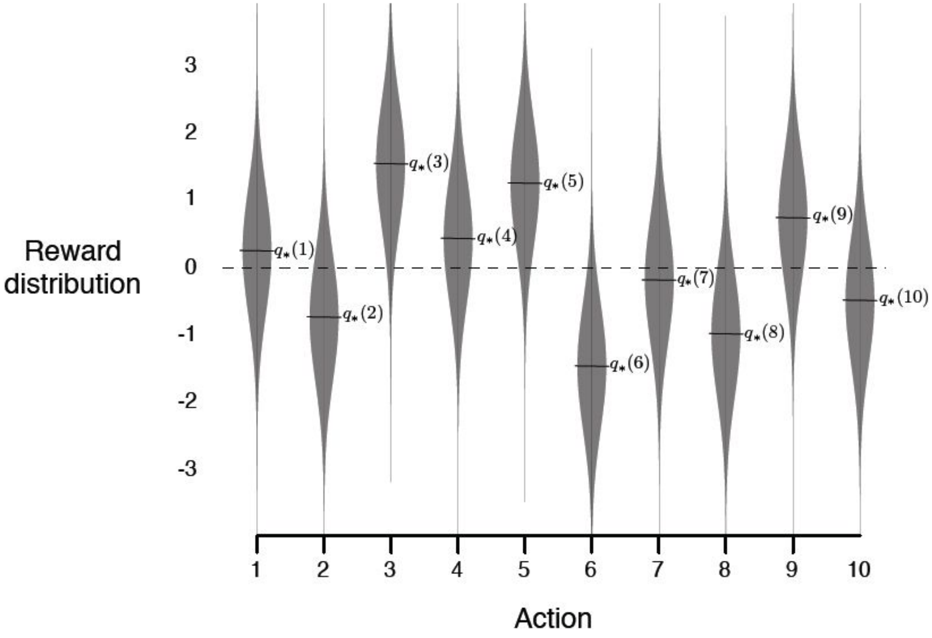
- Then, when we test our proposed learning algo, the obtained reward at each timestep is sampled from a Normal(Q(ai),1)



Section 2.3, Reinforcement Learning: An Introduction, Sutton & Barto

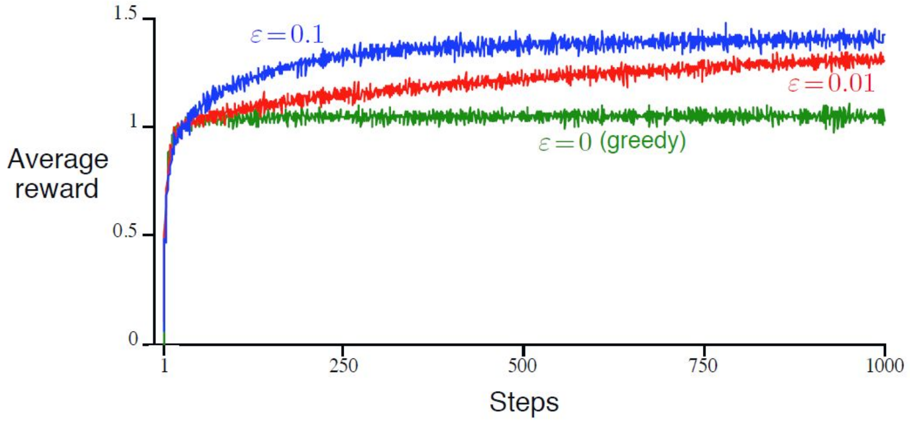
10-armed Testbed

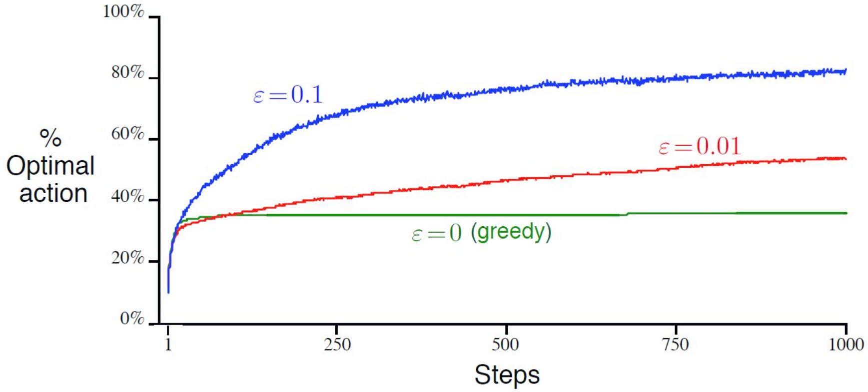
- 2000 10-armed bandit problems
- One run: 1000 timesteps on a bandit problem
- Do 2000 runs, each run for a different bandit problem. This establishes the learning algo performance



Section 2.3, Reinforcement Learning: An Introduction, Sutton & Barto

Performance on 10-armed Testbed: Eps-Greedy





Section 2.3, [Reinforcement Learning: An Introduction](), Sutton & Barto