CS-566 Deep Reinforcement Learning

Model-Free Learning - II
Action Selection

Nazar Khan
Department of Computer Science
University of the Punjab

Three Principles of Model-Free Learning

- There are three principles of model-free learning.
  1. *Reward Sampling*: How should rewards be sampled from the environment?
  2. *Action Selection*: How does the agent decide which action to take?
  3. *Learning from Rewards*: How to use reward to make the agent better?
- Different answers to these questions lead to different methods of reinforcement learning.
- In this lecture: Action Selection.

Exploration in Model-free RL

- In model-free settings, no transition model T is available.
- Agents must sample the environment directly.
- Sampling is often **expensive** (e.g., real-world robot actions).
- Hence, smart action selection is needed to:
  - Avoid wasting samples.
  - Find good policies faster.

Greedy Action Selection

- Idea: always select the action with the current highest Q-value.
- Pros: exploits current knowledge.
- Cons:
  - **Short-sighted**: may converge to local maxima.
  - High bias: based on few samples.
  - Risk of circular reinforcement: policy only reinforces what it already does.

Problem

A purely greedy agent may miss better long-term strategies.

Exploration vs. Exploitation

- To avoid local maxima, agents must sometimes try less-known actions.
- This introduces the **exploration–exploitation trade-off**:
  - **Exploitation**: use current best policy (max Q-values).
  - **Exploration**: try random actions to gather new information.
- Smart policies mix both to balance:
  - Learning speed.
  - Policy quality.

Preview

The ϵ-greedy strategy is one common way to achieve this balance.

Bandit Theory: The Exploration/Exploitation Trade-off

- Fundamental question:
  - How to obtain the most reliable information at the least cost?
- Studied extensively in literature for single-step decision problems
  - Known as the **multi-armed bandit problem**.
- A *bandit* $\Rightarrow$ casino slot machine with many arms
  - Each arm has an unknown payout probability
  - Each trial costs a coin
  - Goal: Find strategy to identify the best arm with minimal cost

Bandit Theory as Reinforcement Learning

- Multi-armed bandit is:
  - A single-state, single-decision RL problem
  - A one-step, non-sequential decision-making problem
- Actions $\Rightarrow$ arms of the bandit
- Simplified model $\Rightarrow$ allows in-depth study of exploration vs. exploitation

Bandit Applications: Clinical Trials

- Example: Testing new drugs in clinical trials
- Bandit $\Rightarrow$ the trial setup
- Arms $\Rightarrow$ choice of assigning subjects to:
  - Experimental drug
  - Placebo
- Serious implication: **human lives at stake**

Fixed vs. Adaptive Trials

Fixed Randomized Controlled Trial

- Group sizes fixed in advance
- Duration and confidence interval fixed
- Risk: More people exposed to harmful drug or deprived of beneficial drug

Adaptive Trial (Bandit Setup)

- Group sizes adapt during trial
- More subjects get promising drug
- Fewer subjects get ineffective/harmful drug

Adaptive Clinical Trial Illustration

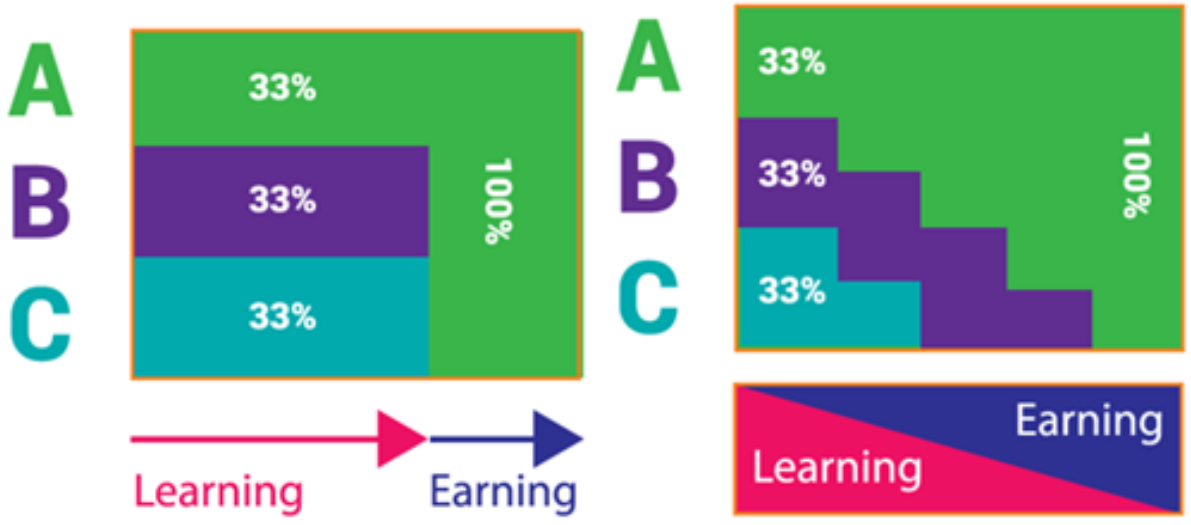

Figure: Adaptive trial: balancing exploration vs. exploitation[1]

[1] Abhishek. *Multi-Arm Bandits: a potential alternative to A/B tests* https://medium.com/brillio-data-science/multi-arm-bandits-a-potential-alternative-to-a-b-tests-a647d9bf2a7e. 2019.

ϵ-Greedy Exploration

- Simple pragmatic strategy:
  - Choose greedy action (highest estimated value) most of the time
  - With probability ϵ, explore another random action
- Example: $\epsilon = 0.1$
  - 90% exploit best-known action
  - 10% explore random actions
- ϵ-greedy is a **soft policy**: non-zero probability for all actions

Exploration/Exploitation Trade-off

- Central concept in reinforcement learning
- Determines:
  - How much confidence in outcomes
  - How quickly variance is reduced
- Variants:
  - **Adaptive** ϵ: decay over time or based on statistics
  - Add **Dirichlet noise**[2] to prior probabilities of actions for exploration
  - Use **Thompson sampling**[3] for Bayesian exploration

[2] Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous Multivariate Distributions, Volume 1: Models and Applications*. John Wiley & Sons, 2004.

[3] Daniel Russo et al. 'A tutorial on Thompson sampling'. In: *Found. Trends Mach. Learn.* 11.1 (2018), pp. 1–96.