ECE 1508: Reinforcement Learning

Chapter 1: Introduction

Ali Bereyhi

ali.bereyhi@utoronto.ca

Department of Electrical and Computer Engineering

University of Toronto

Fall 2025

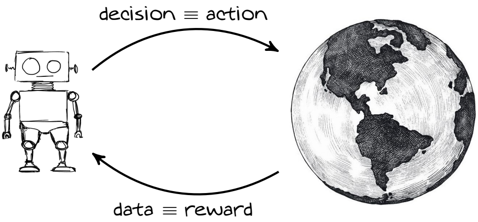
What is Reinforcement Learning?

David Silver from *Deep Mind* calls Reinforcement Learning

the art of decision making

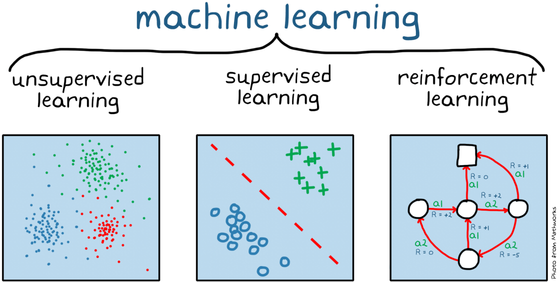
Of course *David Silver* can say it! But, for us as beginners it's better to say

it is all about learning how to interact with environment



We learn many things in our life by this approach

What is Reinforcement Learning?



Reinforcement learning is still a learning problem since

we still learn patterns or behaviors from data $\equiv$ observations

But, it has some fundamental differences to classical learning problems

What is Reinforcement Learning?

Reinforcement learning is *not supervised*

↳ *We don't collect data with labels*

↳ *We only see some rewards time to time*

  ↳ *We need to use them to adjust our behavior*

Reinforcement learning does *not have conventional dataset*

↳ *Each time we decide for an action we get a new feedback*

↳ *This feedback contributes to our learning ≡ a new data-point*

  ↳ *Our dataset is constructed through time*

  ↳ *Our decisions affect the data we collect*

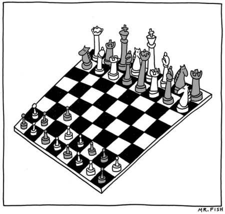
In reinforcement learning *time really matters*

↳ *We should act to see the next data*

↳ *It's also different from classical sequential data*

  ↳ *In classical form, we see the whole sequence before we start with training*

What is Reinforcement Learning?



Best example of a reinforcement setting is a game

- *In games, we start from noting*
  - ↳ *We know just a bunch of rules*
- *We decide for a move ≡ an action*
  - ↳ *We then wait for the other side to move*
- *We decide for next move based on our observation*

- *Reinforcement learning is not supervised*
  - ↳ *We can't label a move as good or bad!*
- *Reinforcement learning does not have conventional dataset*
  - ↳ *Data is what we see through this game and maybe our earlier plays*
  - ↳ *Our current move impacts future moves in the game*
- *In reinforcement learning time really matters*
  - ↳ *We can only decide for the next move once we've seen the opponent's move*

Reinforcement Learning: *Achievements*

Reinforcement Learning has been hugely *revisited* in recent years

Alpha-Zero could beat world champions

take a look at this video where you could laso meet David Silver ☺

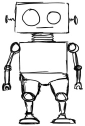
We are getting to there at some point, but first let's go back to 1952 and look at Herbert Robbins' multi-armed bandit problem which is pretty much

the most classic reinforcement learning problem

This helps us develop some intuition

Multi-armed Bandit

There are lots of variants! Let's start with our silly one: *you have developed a programming robot as your course project*

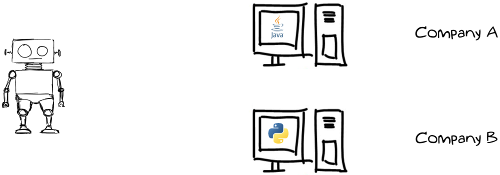


This robot can program in almost all programming languages. You now come up with a brilliant idea

You plan to rent it daily to make some money!

Multi-armed Bandit

Your robot finds two options

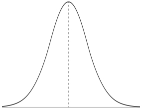


- *Company A which codes in Java*
- *Company B which codes in Python*

But none of these companies has a fixed daily payment

their payments is randomly changing every day

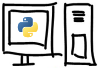
Multi-armed Bandit

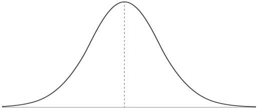


Company A pays rent R_A which is Gaussian random variable

↳ *It has mean* $\mu_A = 600$ ⇝ it pays in average \$600 per day

↳ *Its standard deviation is* $\sigma_A = 100$ ⇝ some days it may even pay \$300





Company B pays rent R_B which is Gaussian random variable

↳ *It has mean* $\mu_B = 400$ ⇝ it pays in average \$400 per day

↳ *Its standard deviation is* $\sigma_B = 200$ ⇝ some days it may even pay \$1000

Multi-armed Bandit: *Known Distributions*

Obviously, the money we make depends on our *renting policy*

Policy

$\pi(t)$ *is the renting policy which specifies the selected company on day* t

$$\pi(t) = \begin{cases} A & \text{if we select Company A on day } t \\ B & \text{if we select Company B on day } t \end{cases}$$

Now, the main question is

What is the optimal renting policy?

To answer this question *we need to define what we mean by optimal*

Multi-armed Bandit: *Goal*

+ *Companies have stochastic payments! How can we define optimality?*
– Well, let's try looking at expected return per day

Average Return

Say we rent out for a period of T days, the average return is

$$G_T = \mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} R_t\right\}$$

where R_t is the rent paid on day t

We now set our goal to maximize the average return

optimal policy $\equiv$ *maximum average return*

Review: *Expectation and Conditional Expectation*

We need to recall expectation both marginal and conditional

Expectation

Let's say X and Y are two random variables; we need to know,

- *How to compute marginal expectation of X*

$$\mathbb{E}\left\{X\right\}$$

- *How to compute expectation of X conditional to Y*

$$\mathbb{E}\left\{X|Y=y\right\}$$

If you don't remember clearly

Please look at the blackboard!

Multi-armed Bandit: *Known Distributions*

+ *But, isn't the answer obvious?!*
- If *we know the distributions yes!*

With known distributions, we could simply write

$$G_T = \mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} R_t\right\} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\{R_t\right\} = \frac{T_A\mu_A + T_B\mu_B}{T}$$

where T_A and T_B are defined to be

- *T_A is the number of days we rent to Company A*
- *T_B is the number of days we rent to Company B*

We can obviously say that $T_A, T_B \leqslant T$ and

$$T_A = T - T_B$$

Multi-armed Bandit: *Known Distributions*

The return is

$$G_T = \frac{T_A \mu_A + T_B \mu_B}{T}$$

We can now write

$$G_T = \frac{\left(T - T_B\right) \mu_A + T_B \mu_B}{T} = \left(1 - \frac{T_B}{T}\right) \mu_A + \frac{T_B}{T} \mu_B$$
$$= \mu_A - \frac{T_B}{T} \left(\mu_A - \mu_B\right)$$

Recall that $\mu_A = 600$ *and* $\mu_B = 400$, *so we have*

$$G_T = 600 - 200 \frac{T_B}{T}$$

Multi-armed Bandit: *Known Distributions*

The return is

$$G_T = 600 - 200\frac{T_B}{T}$$

We know that

$$\text{always rent to A} \leftsquigarrow 0 \leqslant \frac{T_B}{T} \leqslant 1 \rightsquigarrow \text{always rent to B}$$

Therefore, optimal return is when $T_B/T = 0$

$$G_T^\star = 600$$

Multi-armed Bandit: *Known Distributions*

Optimal return is $G_T^\star = 600$

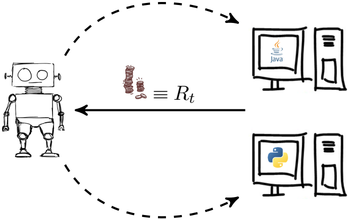
The optimal policy to achieve this return is to

always rent to Company A $\rightsquigarrow$ $\pi^\star(t) = A$ *for all* t

---

+ *This was pretty obvious! Then, what is special about this problem?*
– The problem is that *we don't know the distributions in practice!*

Multi-armed Bandit



In practice, we can only *observe payments done each day by the companies*

we should decide the renting policy based on our observations

+ *OK! It seems hard to find an optimal policy!*
- Well! Let's take a look

Multi-armed Bandit: *Exploring Policy*

We start by a dumb policy: *we flip a uniform coin every day to select company*

$$\pi(t) = \begin{cases} A & \text{with Probability } 0.5 \\ B & \text{with Probability } 0.5 \end{cases}$$

Let's compute the return in this case

we denote it as $G_T^{(\mathrm{r})}$ *as it's a random policy*

Multi-armed Bandit: *Exploring Policy*

We have in this case

$$G_T^{(\mathrm{r})} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\{R_t\right\} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\pi(t)}\left\{\mathbb{E}_{R_t}\left\{R_t | \pi\left(t\right)\right\}\right\}$$

$$= \frac{1}{T}\sum_{t=1}^{T} 0.5\mathbb{E}_{R_t}\left\{R_t | \pi\left(t\right) = A\right\} + 0.5\mathbb{E}_{R_t}\left\{R_t | \pi\left(t\right) = B\right\}$$

$$= \frac{1}{T}\sum_{t=1}^{T} 0.5\mu_A + 0.5\mu_B = 0.5\mu_A + 0.5\mu_B = 500$$

Now comparing to the maximal return we could get, we achieve

$$\text{regret} \leftsquigarrow \rho^{(\mathrm{r})} = G_T^{\star} - G_T^{(\mathrm{r})} = 600 - 500 = 100$$

less return!

Multi-armed Bandit: *Exploring-Exploiting Policy*

+ *But, can we get closer to the maximal return?*
- Sure! We should try to *learn the behavior of the companies*

In the previous approach, *we were only exploring the companies*

↳ *Maybe, we should exploit our exploration*

  ↳ *We get a rough idea about companies, once we get paid by them*

    ↳ *We could try both companies*

  ↳ *If we conclude that one company pays more*

    ↳ *We could stick to that company for the rest of the days*

↳ *This sounds like a learning procedure*

  ↳ *We try to learn the distribution of payments*

Multi-armed Bandit: *Exploring-Exploiting Policy*

Let's make another policy: *we try each company once, and stick to the one with who pays higher, i.e.,* $\pi(1) = A$, $\pi(2) = B$, *and*

$$\pi(t) = \begin{cases} A & \text{if } R_1 > R_2 \\ B & \text{if } R_1 \leqslant R_2 \end{cases}$$

Let's compute the return in this case

we denote it as $G_T^{(1)}$ *as we explore only once*

Review: *Properties of Gaussian Variables*

We need to recall Gaussian distribution and its properties

Expectation

We need to know,

- *What* $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ *means*
- *What Q-function is*
- *What the distribution of sum of Gaussian variables is*

If you don't remember clearly

Please look at the blackboard!

Multi-armed Bandit: *Exploring-Exploiting Policy*

Let's start with finding the return

$$
\begin{aligned}
G_T^{(1)} &= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{R_t\right\} \\
&= \frac{1}{T} \left(\mathbb{E}\left\{R_1\right\} + \mathbb{E}\left\{R_2\right\} + \sum_{t=3}^{T} \mathbb{E}_{\pi(t)} \left\{ \mathbb{E}_{R_t} \left\{ R_t | \pi\left(t\right) \right\} \right\} \right) \\
&= \frac{1}{T} \left(\mu_A + \mu_B + \sum_{t=3}^{T} \Pr\left\{\pi\left(t\right) = A\right\} \mu_A + \left(1 - \Pr\left\{\pi\left(t\right) = A\right\}\right) \mu_B \right)
\end{aligned}
$$

Let's define $\Pr\left\{\pi\left(t\right) = A\right\} = p_1$*; then, we can write*

$$
G_T^{(1)} = \frac{1}{T} \left(\mu_A + \mu_B + \left(T-2\right) \left[p_1 \mu_A + \left(1 - p_1\right) \mu_B \right] \right)
$$

Multi-armed Bandit: *Exploring-Exploiting Policy*

We have in this case

$$p_1 = \Pr\{\pi(t) = A\} = \Pr\{R_1 > R_2\} = \Pr\{R_1 - R_2 > 0\}$$

Let's look at the random variable $\Delta = R_1 - R_2$*:* R_1 *and* R_2 *are Gaussian*

↳ Δ *is also a Gaussian variable*

↳ *It's mean is*

$$\mu_\Delta = \mathbb{E}\{R_1 - R_2\} = \mu_A - \mu_B = 200$$

↳ *It's standard variation is*

$$\sigma_\Delta = \sqrt{\sigma_A^2 + \sigma_B^2} = \sqrt{50000} = 223.6$$

Well, p_1 is readily computed

$$p_1 = \Pr\{\Delta > 0\} = \mathrm{Q}\left(-\frac{200}{223.6}\right) = \mathrm{Q}(-0.89) \approx 0.81$$

Multi-armed Bandit: *Exploring-Exploiting Policy*

The average return is hence given by

$$
\begin{aligned}
G_T^{(1)} &= \frac{\mu_A + \mu_B}{T} + \left(1 - \frac{2}{T}\right)(0.81\mu_A + 0.19\mu_B) \\
&= (0.81\mu_A + 0.19\mu_B) - \frac{0.62}{T}(\mu_A - \mu_B) \\
&= 562 - \frac{124}{T}
\end{aligned}
$$

Comparing to maximal average return, we have

$$
\rho^{(1)} = G_T^{\star} - G_T^{(1)} = 38 + \frac{124}{T}
$$

which can be much less than $\rho^{(\mathrm{r})} = 100$ *if* T *is large enough!*

Multi-armed Bandit: *Exploring-Exploiting Policy*

There is however an obvious problem with this approach

Though better in average, it's not that reliable!

Reliability Issue

This would not be reliable in general: we could get for one random sample $R_1 \leqslant R_2$ *even though we have* $\mu_A > \mu_B$*!*

+ *Do you think we can make it more reliable?*
– Sure! Let's explore a bit more

Multi-armed Bandit: *Exploring-Exploiting Policy*

Let's make the policy more reliable: *we try each company d days and stick to the one with higher sum payments, i.e., we set $\pi(t) = A$ for $t = 1, \dots, d$ and then compute a parameter*

$$S_A = \sum_{t=1}^{d} R_t$$

We then set $\pi(t) = B$ for $t = d+1, \dots, 2d$ and compute

$$S_B = \sum_{t=d+1}^{2d} R_t$$

We finally set for $t > 2d$

$$\pi(t) = \begin{cases} A & \text{if } S_A > S_B \\ B & \text{if } S_A \leqslant S_B \end{cases}$$

Multi-armed Bandit: *Exploring-Exploiting Policy*

We can follow the same lines of calculation

$$\begin{aligned} G_T^{(d)} &= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{R_t\right\} \\ &= \frac{1}{T} \left(\sum_{t=1}^{d} \mathbb{E}\left\{R_t\right\} + \sum_{t=d+1}^{2d} \mathbb{E}\left\{R_t\right\} + \sum_{t=2d+1}^{T} \mathbb{E}_{\pi(t)}\left\{\mathbb{E}_{R_t}\left\{R_t | \pi\left(t\right)\right\}\right\} \right) \\ &= \frac{1}{T} \left(d\mu_A + d\mu_B + \sum_{t=2d+1}^{T} \Pr\left\{\pi\left(t\right) = A\right\}\mu_A + \left(1 - \Pr\left\{\pi\left(t\right) = A\right\}\right)\mu_B \right) \end{aligned}$$

We now define $\Pr\left\{\pi\left(t\right) = A\right\} = p_d$ *for* $t > 2d$ *and write*

$$G_T^{(d)} = \frac{1}{T}\left(d\left[\mu_A + \mu_B\right] + \left(T - 2d\right)\left[p_d\mu_A + \left(1 - p_d\right)\mu_B\right]\right)$$

Multi-armed Bandit: *Exploring-Exploiting Policy*

We have in this case

$$p_d = \Pr\left\{\pi\left(t\right) = A\right\} = \Pr\left\{S_A > S_B\right\} = \Pr\left\{\Delta > 0\right\}$$

where we define

$$\Delta = S_A - S_B$$

- S_A *is sum of Gaussian variables* $\rightsquigarrow$ *it's Gaussian*
  - ↳ *It's mean is*

$$\mathbb{E}\left\{S_A\right\} = S_A = \sum_{t=1}^{d} \mathbb{E}\left\{R_t\right\} = d\mu_A$$

  - ↳ *It's standard variation is*

$$\sqrt{\mathbb{E}\left\{\left(S_A - d\mu_A\right)^2\right\}} = \sqrt{d\sigma_A^2} = \sigma_A\sqrt{d}$$

Multi-armed Bandit: *Exploring-Exploiting Policy*

We have in this case

$$p_d = \Pr\left\{\pi\left(t\right) = A\right\} = \Pr\left\{S_A > S_B\right\} = \Pr\left\{\Delta > 0\right\}$$

where we define

$$\Delta = S_A - S_B$$

- *S_B is sum of Gaussian variables $\rightsquigarrow$ it's Gaussian*
  - ↳ *It's mean is*

$$\mathbb{E}\left\{S_B\right\} = S_B = \sum_{t=d+1}^{2d} \mathbb{E}\left\{R_t\right\} = d\mu_B$$

  - ↳ *It's standard variation is*

$$\sqrt{\mathbb{E}\left\{\left(S_B - d\mu_B\right)^2\right\}} = \sqrt{d\sigma_B^2} = \sigma_B\sqrt{d}$$

Multi-armed Bandit: *Exploring-Exploiting Policy*

Let's look at the random variable $\Delta = S_A - S_B$: S_A and S_B are Gaussian

↳ Δ *is also a Gaussian variable*

↳ *It's mean is*

$$\mu_\Delta = \mathbb{E}\left\{S_A - S_B\right\} = d\left(\mu_A - \mu_B\right) = 200d$$

↳ *It's standard variation is*

$$\sigma_\Delta = \sqrt{\sigma_A^2 + \sigma_B^2}\sqrt{d} = 223.6\sqrt{d}$$

Well, p_d is readily computed

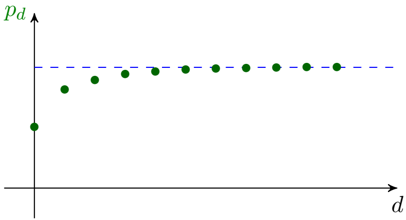
$$p_d = \Pr\left\{\Delta > 0\right\} = \mathrm{Q}\left(-\frac{200\sqrt{d}}{223.6}\right) = \mathrm{Q}\left(-0.89\sqrt{d}\right)$$

Multi-armed Bandit: *Exploring-Exploiting Policy*

The average return is hence given by

$$G_T^{(d)} = p_d \mu_A + (1 - p_d)\, \mu_B - \frac{2d}{T} (p_d - 0.5) (\mu_A - \mu_B)$$

where our p_d quickly converges to one



So we could say, for $d > 5$ we have

$$G_T^{(d)} \approx \mu_A - \frac{d}{T} (\mu_A - \mu_B)$$

Multi-armed Bandit: *Exploring-Exploiting Policy*

For slightly large d, we have

$$G_T^{(d)} \approx 600 - \frac{200d}{T}$$

So, we could conclude that

$$\rho^{(d)} \approx G_T^{\star} - G_T^{(d)} = \frac{200d}{T}$$

If we rent out for some long time, i.e., $T \gg d$; then, we eventually get to the optimal return

$$\lim_{T \to \infty} \rho^{(d)} = 0$$

Key Task: *Decision Making*

The key task in this problem was *decision making*

- *We train our robot to decide for best employer*
  - ↳ *This robot is referred to as an agent*
- *Everything would have been easy if we already had all data available*
  - ↳ *Before starting, we had lots of payment samples*
    - ↳ *We could then find out the payment distribution*
- *We are observing data while we are making decisions*
  - ↳ *We find out about payments after working there*
  - ↳ *We are interacting with an environment*
    - ↳ *The two companies in our example are the environment*

This is a simple example of a reinforcement learning problem

Let's introduce it and break its components down