Bandits

CISC 7404 - Decision Making

Steven Morad

University of Macau

Notation 2
Bandits 12
Multiarmed Bandits 30
Coding 43

Notation

Notation

Let us review some notation I will use in the course

Notation

Let us review some notation I will use in the course

If you ever get confused, come back to these slides

Notation

Let us review some notation I will use in the course

If you ever get confused, come back to these slides

Vectors

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

Notation

We will represent vectors or matrices of **tensors**

Vector of tensors

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}$$

Each $\boldsymbol{x}_i$ could be a vector, matrix, 3x3 tensor, etc

Notation

We will represent vectors or matrices of **tensors**

Vector of tensors

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}$$

Each $\boldsymbol{x}_i$ could be a vector, matrix, 3x3 tensor, etc

Notation

Same for matrices

Matrix of tensors

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{1,1} & \boldsymbol{x}_{1,2} & \dots & \boldsymbol{x}_{1,n} \\ \boldsymbol{x}_{2,1} & \boldsymbol{x}_{2,2} & \dots & \boldsymbol{x}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{m,1} & \boldsymbol{x}_{m,2} & \dots & \boldsymbol{x}_{m,n} \end{bmatrix}$$

Notation

Question: What is the difference between the following?

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{1,1} & \boldsymbol{x}_{1,2} & \cdots & \boldsymbol{x}_{1,n} \\ \boldsymbol{x}_{2,1} & \boldsymbol{x}_{2,2} & \cdots & \boldsymbol{x}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{m,1} & \boldsymbol{x}_{m,2} & \cdots & \boldsymbol{x}_{m,n} \end{bmatrix}$$

Notation

Capital letters will often refer to **sets**

Notation

Capital letters will often refer to **sets**

$$X = \{1, 2, 3, 4\}$$

Notation

Capital letters will often refer to **sets**

$$X = \{1, 2, 3, 4\}$$

We will represent important sets with blackboard font

$\mathbb{R}$	Set of all real numbers $\{1, 2.03, \pi, ...\}$
$\mathbb{Z}$	Set of all integers $\{-2, -1, 0, 1, 2, ...\}$
$\mathbb{Z}_+$	Set of all **positive** integers $\{1, 2, ...\}$

Notation

The max operator returns the maximum of a function over its domain

Notation

The max operator returns the maximum of a function over its domain

$$\max_x f(x)$$

The arg max operator returns the input that maximizes a function

$$\arg\max_x f(x)$$

Notation

The max operator returns the maximum of a function over its domain

$$\max_x f(x)$$

The arg max operator returns the input that maximizes a function

$$\arg \max_x f(x)$$

Notation

Let's quiz you on some notation

Notation

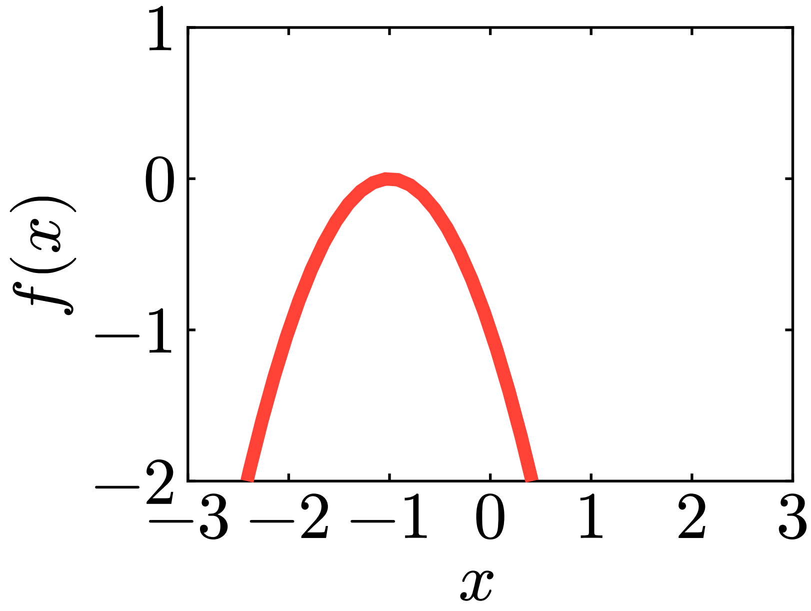
Let's quiz you on some notation

$$f(x) = -(x + 1)^2$$

Notation

Let's quiz you on some notation

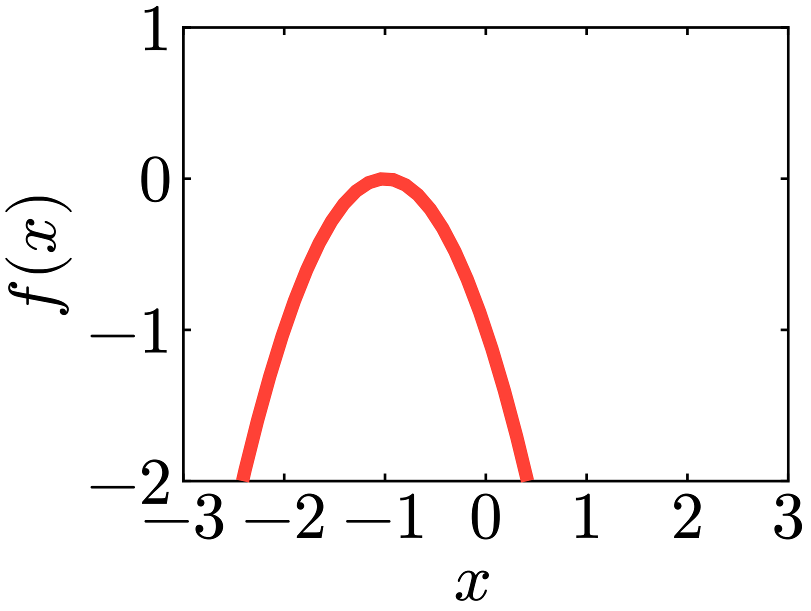
$$f(x) = -(x+1)^2$$



Notation

Let's quiz you on some notation

$$f(x) = -(x+1)^2$$

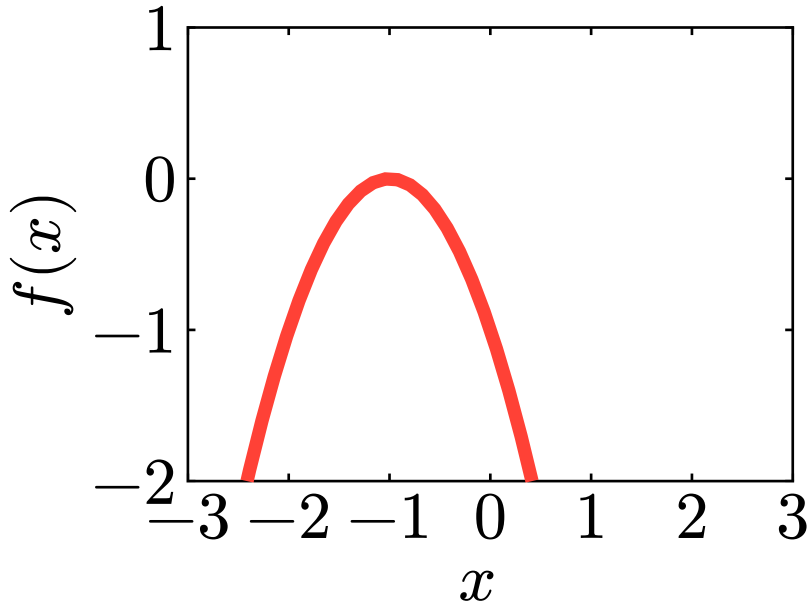


$$\max_x f(x)$$

Notation

Let's quiz you on some notation

$$f(x) = -(x+1)^2$$



$$\max_x f(x)$$

$$\arg\max_x f(x)$$

Notation

$$\mathbb{R}^n$$

Notation

$\mathbb{R}^n$	Set of all vectors containing n real numbers

Notation

$\mathbb{R}^n$	Set of all vectors containing n real numbers
$\mathbb{Z}_{3:6}$	

Notation

$\mathbb{R}^n$	Set of all vectors containing n real numbers
$\mathbb{Z}_{3:6}$	Set of all integers between 3 and 6

Notation

$\mathbb{R}^n$	Set of all vectors containing n real numbers
$\mathbb{Z}_{3:6}$	Set of all integers between 3 and 6
$[0, 1]^n$	

Notation

$\mathbb{R}^n$	Set of all vectors containing n real numbers
$\mathbb{Z}_{3:6}$	Set of all integers between 3 and 6
$[0, 1]^n$	Set of all vectors of length n with values between 0 and 1
$\{0, 1\}^n$	

Notation

$\mathbb{R}^n$	Set of all vectors containing n real numbers
$\mathbb{Z}_{3:6}$	Set of all integers between 3 and 6
$[0, 1]^n$	Set of all vectors of length n with values between 0 and 1
$\{0, 1\}^n$	Set of all boolean vectors of length n

Notation

We define **functions** or **maps** between sets

$$f : X \times \Theta \mapsto Y$$

Notation

We define **functions** or **maps** between sets

$$f : X \times \Theta \mapsto Y$$

The function f takes in elements from sets X and Θ and outputs elements of set Y

Notation

We define **functions** or **maps** between sets

$$f : X \times \Theta \mapsto Y$$

The function f takes in elements from sets X and Θ and outputs elements of set Y

Question: What is X and Y?

Notation

We define **functions** or **maps** between sets

$$f : X \times \Theta \mapsto Y$$

The function f takes in elements from sets X and Θ and outputs elements of set Y

Question: What is X and Y?

Answer: I did not say yet, let us define it

Notation

We define **functions** or **maps** between sets

$$f : X \times \Theta \mapsto Y$$

The function f takes in elements from sets X and Θ and outputs elements of set Y

Question: What is X and Y?

Answer: I did not say yet, let us define it

$$X \in \mathbb{R}^n, Y \in [0, 1]^{n \times m}$$

Notation

We define **functions** or **maps** between sets

$$f : X \times \Theta \mapsto Y$$

The function f takes in elements from sets X and Θ and outputs elements of set Y

Question: What is X and Y?

Answer: I did not say yet, let us define it

$$X \in \mathbb{R}^n, Y \in [0, 1]^{n \times m}$$

Question: What does this function do?

Bandits

Bandits

The Sutton and Barto textbook reviews bandits before introducing reinforcement learning

Bandits

The Sutton and Barto textbook reviews bandits before introducing reinforcement learning

Bandits are a simplified version of reinforcement learning

Bandits

The Sutton and Barto textbook reviews bandits before introducing reinforcement learning

Bandits are a simplified version of reinforcement learning

It provides a "taste" of reinforcement learning in a single lecture

Bandits

The Sutton and Barto textbook reviews bandits before introducing reinforcement learning

Bandits are a simplified version of reinforcement learning

It provides a "taste" of reinforcement learning in a single lecture

Today's lecture will be difficult

Bandits

The Sutton and Barto textbook reviews bandits before introducing reinforcement learning

Bandits are a simplified version of reinforcement learning

It provides a "taste" of reinforcement learning in a single lecture

Today's lecture will be difficult

But if you can understand it, then reinforcement learning will be easy for you

Bandits

The simplest form of decision making problem is the **bandit**

Bandits

The simplest form of decision making problem is the **bandit**

Question: What is a bandit?

Bandits

The simplest form of decision making problem is the **bandit**

Question: What is a bandit?

Bandits

The simplest form of decision making problem is the **bandit**

Question: What is a bandit?

A bandit steals your money

Bandits

Here is the bandit we will focus on in this course

Bandits

Here is the bandit we will focus on in this course

Bandits

Here is the bandit we will focus on in this course

This is a **one-armed** bandit

Bandits

Bandits

Question: How does a one-armed bandit steal your money?

Bandits

Question: How does a one-armed bandit steal your money?

Answer: You win less money than you put in

Bandits

Question: How does a one-armed bandit steal your money?

Answer: You win less money than you put in

Example: Costs 10 MOP to play, you can win 1000 MOP each spin

Bandits

Question: How does a one-armed bandit steal your money?

Answer: You win less money than you put in

Example: Costs 10 MOP to play, you can win 1000 MOP each spin

Your chance of winning is $\frac{1}{200}$

Bandits

Question: How does a one-armed bandit steal your money?

Answer: You win less money than you put in

Example: Costs 10 MOP to play, you can win 1000 MOP each spin

Your chance of winning is $\frac{1}{200}$

Let us see if we can make money playing this game

Bandits

We will use **probability** to understand how much money we will make

Bandits

We will use **probability** to understand how much money we will make

The world is based on random **outcomes**, down to the atomic level

Bandits

We will use **probability** to understand how much money we will make

The world is based on random **outcomes**, down to the atomic level

$$\Omega \in \{\text{win}, \text{lose}\}$$

Bandits

We will use **probability** to understand how much money we will make

The world is based on random **outcomes**, down to the atomic level

$$\Omega \in \{\text{win}, \text{lose}\}$$

We define the probability over the outcome space

Bandits

We will use **probability** to understand how much money we will make

The world is based on random **outcomes**, down to the atomic level

$$\Omega \in \{\text{win}, \text{lose}\}$$

We define the probability over the outcome space

$$\Pr(\text{win}) = \frac{1}{200}, \quad \Pr(\text{lose}) = \frac{199}{200}$$

Bandits

We will use **probability** to understand how much money we will make

The world is based on random **outcomes**, down to the atomic level

$$\Omega \in \{\text{win}, \text{lose}\}$$

We define the probability over the outcome space

$$\Pr(\text{win}) = \frac{1}{200}, \quad \Pr(\text{lose}) = \frac{199}{200}$$

Probabilities **must be positive** and **must sum to one**

Bandits

We will use **probability** to understand how much money we will make

The world is based on random **outcomes**, down to the atomic level

$$\Omega \in \{\text{win}, \text{lose}\}$$

We define the probability over the outcome space

$$\Pr(\text{win}) = \frac{1}{200}, \quad \Pr(\text{lose}) = \frac{199}{200}$$

Probabilities **must be positive** and **must sum to one**

$$\sum_{\omega \in \Omega} \Pr(\omega) = 1$$

Bandits

A **random variable** $\mathcal{X}$ maps an outcome to a real number

Bandits

A **random variable** $\mathcal{X}$ maps an outcome to a real number

$$\mathcal{X} : \Omega \mapsto \mathbb{R}$$

Bandits

A **random variable** $\mathcal{X}$ maps an outcome to a real number

$$\mathcal{X} : \Omega \mapsto \mathbb{R}$$

Our bandit has two outcomes, lose (-10) or win (1000)

Bandits

A **random variable** $\mathcal{X}$ maps an outcome to a real number

$$\mathcal{X} : \Omega \mapsto \mathbb{R}$$

Our bandit has two outcomes, lose (-10) or win (1000)

Question: What is the random variable for the bandit?

Bandits

A **random variable** $\mathcal{X}$ maps an outcome to a real number

$$\mathcal{X} : \Omega \mapsto \mathbb{R}$$

Our bandit has two outcomes, lose (-10) or win (1000)

Question: What is the random variable for the bandit?

$$\mathcal{X} : \{\text{lose}, \text{win}\} \mapsto \{-10, 1000\}$$

Bandits

A **random variable** $\mathcal{X}$ maps an outcome to a real number

$$\mathcal{X} : \Omega \mapsto \mathbb{R}$$

Our bandit has two outcomes, lose (-10) or win (1000)

Question: What is the random variable for the bandit?

$$\mathcal{X} : \{\text{lose}, \text{win}\} \mapsto \{-10, 1000\} \qquad \mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

Bandits

We can also compute the probability over random variables

Bandits

We can also compute the probability over random variables

$$\Pr(\mathcal{X} = x) = \left\{ \Pr\left(\underbrace{\mathcal{X}(\omega)}_{\text{Outcome to real}} = \underbrace{x}_{\text{Real}} \right) \middle| \underbrace{\omega}_{\text{Outcome}} \in \underbrace{\Omega}_{\text{Outcomes}} \right\}$$

Bandits

We can also compute the probability over random variables

$$\Pr(\mathcal{X} = x) = \left\{ \Pr\left(\underbrace{\mathcal{X}(\omega)}_{\text{Outcome to real}} = \underbrace{x}_{\text{Real}} \right) \middle| \underbrace{\omega}_{\text{Outcome}} \in \underbrace{\Omega}_{\text{Outcomes}} \right\}$$

$$\mathcal{X} : \{\text{lose}, \text{win}\} \mapsto \{-10, 1000\}$$

Bandits

We can also compute the probability over random variables

$$\Pr(\mathcal{X} = x) = \left\{ \Pr\left(\underbrace{\mathcal{X}(\omega)}_{\text{Outcome to real}} = \underbrace{x}_{\text{Real}} \right) \middle| \underbrace{\omega}_{\text{Outcome}} \in \underbrace{\Omega}_{\text{Outcomes}} \right\}$$

$$\mathcal{X} : \{\text{lose}, \text{win}\} \mapsto \{-10, 1000\} \qquad \mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

$$\Pr(\mathcal{X}) = \begin{bmatrix} \Pr(\mathcal{X} = -10) \\ \Pr(\mathcal{X} = 1000) \end{bmatrix} =$$

Bandits

We can also compute the probability over random variables

$$\Pr(\mathcal{X} = x) = \left\{ \Pr\left(\underbrace{\mathcal{X}(\omega)}_{\text{Outcome to real}} = \underbrace{x}_{\text{Real}} \right) \middle| \underbrace{\omega}_{\text{Outcome}} \in \underbrace{\Omega}_{\text{Outcomes}} \right\}$$

$$\mathcal{X} : \{\text{lose}, \text{win}\} \mapsto \{-10, 1000\} \qquad \mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

$$\Pr(\mathcal{X}) = \begin{bmatrix} \Pr(\mathcal{X} = -10) \\ \Pr(\mathcal{X} = 1000) \end{bmatrix} = \begin{bmatrix} \frac{199}{200} \\ \frac{1}{200} \end{bmatrix} = \begin{bmatrix} 0.995 \\ 0.005 \end{bmatrix}$$

Bandits

We can also compute the probability over random variables

$$\Pr(\mathcal{X} = x) = \left\{ \Pr\left(\underbrace{\mathcal{X}(\omega)}_{\text{Outcome to real}} = \underbrace{x}_{\text{Real}} \right) \middle| \underbrace{\omega}_{\text{Outcome}} \in \underbrace{\Omega}_{\text{Outcomes}} \right\}$$

$$\mathcal{X} : \{\text{lose}, \text{win}\} \mapsto \{-10, 1000\} \qquad \mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

$$\Pr(\mathcal{X}) = \begin{bmatrix} \Pr(\mathcal{X} = -10) \\ \Pr(\mathcal{X} = 1000) \end{bmatrix} = \begin{bmatrix} \frac{199}{200} \\ \frac{1}{200} \end{bmatrix} = \begin{bmatrix} 0.995 \\ 0.005 \end{bmatrix}$$

Bandits

Like before, the probability over the random variable **must sum to one**

Bandits

Like before, the probability over the random variable **must sum to one**

$$\sum_{\omega \in \Omega} \Pr(X(\omega)) = 1$$

Bandits

Like before, the probability over the random variable **must sum to one**

$$\sum_{\omega \in \Omega} \Pr(X(\omega)) = 1$$

$$\Pr(\mathcal{X}(\text{lose}) = -10) + \Pr(\mathcal{X}(\text{win}) = 1000) = 1$$

Bandits

Like before, the probability over the random variable **must sum to one**

$$\sum_{\omega \in \Omega} \Pr(X(\omega)) = 1$$

$$\Pr(\mathcal{X}(\text{lose}) = -10) + \Pr(\mathcal{X}(\text{win}) = 1000) = 1$$

$$\frac{199}{200} + \frac{1}{200} = 1$$

Bandits

We defined our bandit's probabilities

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

Bandits

We defined our bandit's probabilities

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

And the expected values

$$\mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

Bandits

We defined our bandit's probabilities

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

And the expected values

$$\mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

But we still do not know how much money we will make!

Bandits

We defined our bandit's probabilities

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

And the expected values

$$\mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

But we still do not know how much money we will make!

But we can combine them to find out

Bandits

The **expectation** or **expected value** $\mathbb{E}$ tells us how much money we make on average

Bandits

The **expectation** or **expected value** $\mathbb{E}$ tells us how much money we make on average

$$\mathbb{E} : \underbrace{(\Omega \mapsto \mathbb{R})}_{\text{random variable}} \mapsto \mathbb{R}$$

Bandits

The **expectation** or **expected value** $\mathbb{E}$ tells us how much money we make on average

$$\mathbb{E} : \quad \underbrace{(\Omega \mapsto \mathbb{R})}_{\text{random variable}} \quad \mapsto \mathbb{R}$$

$$\mathbb{E}[\mathcal{X}] = \sum_{\omega \in \Omega} \mathcal{X}(\omega) \cdot \Pr(\omega)$$

Bandits

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

Bandits

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

$$\mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

Bandits

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

$$\mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

$$\mathbb{E}[\mathcal{X}] = \sum_{\omega \in \Omega} \Pr(\omega) \cdot \mathcal{X}(\omega)$$

Bandits

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

$$\mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

$$\mathbb{E}[\mathcal{X}] = \sum_{\omega \in \Omega} \Pr(\omega) \cdot \mathcal{X}(\omega)$$

Question: What is the expected value of the bandit?

Bandits

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

$$\mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

$$\mathbb{E}[\mathcal{X}] = \sum_{\omega \in \Omega} \Pr(\omega) \cdot \mathcal{X}(\omega)$$

Question: What is the expected value of the bandit?

$$\Pr(\text{lose}) \cdot \mathcal{X}(\text{lose}) + \Pr(\text{win}) \cdot \mathcal{X}(\text{win})$$

Bandits

$$\Pr(\text{lose}) = \frac{199}{200}; \quad \Pr(\text{win}) = \frac{1}{200}$$

$$\mathcal{X}(\text{lose}) = -10; \quad \mathcal{X}(\text{win}) = 1000$$

$$\mathbb{E}[\mathcal{X}] = \sum_{\omega \in \Omega} \Pr(\omega) \cdot \mathcal{X}(\omega)$$

Question: What is the expected value of the bandit?

$$\Pr(\text{lose}) \cdot \mathcal{X}(\text{lose}) + \Pr(\text{win}) \cdot \mathcal{X}(\text{win})$$

$$\frac{199}{200} \cdot -10 + \frac{1}{200} \cdot 1000 = -4.95$$

Bandits

Question: What does $\mathbb{E}[\mathcal{X}] = -4.95$ mean?

Bandits

Question: What does $\mathbb{E}[\mathcal{X}] = -4.95$ mean?

Expect to lose 4.95 MOP on average each time you spin the bandit

Bandits

Question: What does $\mathbb{E}[\mathcal{X}] = -4.95$ mean?

Expect to lose 4.95 MOP on average each time you spin the bandit

We call the value after each spin the **reward**

Bandits

Question: What does $\mathbb{E}[\mathcal{X}] = -4.95$ mean?

Expect to lose 4.95 MOP on average each time you spin the bandit

We call the value after each spin the **reward**

$$
\begin{gathered}
r_1 = -10 \\
r_2 = -10 \\
\vdots \\
r_n = -10
\end{gathered}
$$

Bandits

Question: What does $\mathbb{E}[\mathcal{X}] = -4.95$ mean?

Expect to lose 4.95 MOP on average each time you spin the bandit

We call the value after each spin the **reward**

$$
\begin{aligned}
r_1 &= -10 \\
r_2 &= -10 \\
&\vdots \\
r_n &= -10
\end{aligned}
$$

Negative reward means we lose money

Bandits

$$r_1 = -10$$
$$r_2 = -10$$
$$\vdots$$
$$r_n = -10$$

Bandits

$$
\begin{aligned}
r_1 &= -10 \\
r_2 &= -10 \\
&\vdots \\
r_n &= -10
\end{aligned}
$$

As we play the game more and more, we converge to the expectation

$$\lim_{n \to \infty} \sum_{t=1}^{n} r_t = -4.95n = n\mathbb{E}[\mathcal{X}]$$

Bandits

$$\lim_{n \to \infty} \sum_{t=1}^{n} r_t = -4.95n = n\mathbb{E}[\mathcal{X}]$$

Bandits

$$\lim_{n \to \infty} \sum_{t=1}^{n} r_t = -4.95n = n\mathbb{E}[\mathcal{X}]$$

If you spin 1,000 times, you should expect to lose −4950 MOP

Bandits

$$\lim_{n \to \infty} \sum_{t=1}^{n} r_t = -4.95n = n\mathbb{E}[\mathcal{X}]$$

If you spin 1,000 times, you should expect to lose −4950 MOP

Question: What is the best way to make money with the bandit?

Bandits

$$\lim_{n \to \infty} \sum_{t=1}^{n} r_t = -4.95n = n\mathbb{E}[\mathcal{X}]$$

If you spin 1,000 times, you should expect to lose −4950 MOP

Question: What is the best way to make money with the bandit?

Answer: Do not play! If you must, play as little as possible

Bandits

If you know $\mathbb{E}[\mathcal{X}]$, you know the result of gambling

Bandits

If you know $\mathbb{E}[\mathcal{X}]$, you know the result of gambling

Question: Do gamblers know $\mathbb{E}[\mathcal{X}]$?

Bandits

If you know $\mathbb{E}[\mathcal{X}]$, you know the result of gambling

Question: Do gamblers know $\mathbb{E}[\mathcal{X}]$?

Answer: No! This is a secret of the casino

Bandits

If you know $\mathbb{E}[\mathcal{X}]$, you know the result of gambling

Question: Do gamblers know $\mathbb{E}[\mathcal{X}]$?

Answer: No! This is a secret of the casino

Question: How could a gambler find out $\mathbb{E}[\mathcal{X}]$?

Bandits

Question: How could a gambler find out $\mathbb{E}[\mathcal{X}]$?

Gambler only has access to the rewards

$$r_1, r_2, ..., r_n = -10, -10, ..., 1000$$

Bandits

Question: How could a gambler find out $\mathbb{E}[\mathcal{X}]$?

Gambler only has access to the rewards

$$r_1, r_2, ..., r_n = -10, -10, ..., 1000$$

We can sum the rewards

$$\sum_{t=1}^{n} r_t \approx n\mathbb{E}[\mathcal{X}]$$

Bandits

Question: How could a gambler find out $\mathbb{E}[\mathcal{X}]$?

Gambler only has access to the rewards

$$r_1, r_2, ..., r_n = -10, -10, ..., 1000$$

We can sum the rewards

$$\sum_{t=1}^{n} r_t \approx n\mathbb{E}[\mathcal{X}]$$

Divide by number of plays

$$\frac{1}{n} \sum_{t=1}^{n} r_t \approx \mathbb{E}[\mathcal{X}]$$

Bandits

Question: How could a gambler find out $\mathbb{E}[\mathcal{X}]$?

Gambler only has access to the rewards

$$r_1, r_2, ..., r_n = -10, -10, ..., 1000$$

We can sum the rewards

$$\sum_{t=1}^{n} r_t \approx n\mathbb{E}[\mathcal{X}]$$

Divide by number of plays

$$\frac{1}{n} \sum_{t=1}^{n} r_t \approx \mathbb{E}[\mathcal{X}]$$

After playing enough, the gambler can approximate the expectation!

Bandits

Exercise: You start a new casino in Macau.

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Write down:

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Write down:

- Probability for each outcome $P(\omega); \quad \forall \omega \in \Omega$

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Write down:

- Probability for each outcome $P(\omega); \quad \forall \omega \in \Omega$
- The random variable for each outcome $\mathcal{X}(\omega); \quad \forall \omega \in \Omega$

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Write down:

- Probability for each outcome $P(\omega); \quad \forall \omega \in \Omega$
- The random variable for each outcome $\mathcal{X}(\omega); \quad \forall \omega \in \Omega$
- The expected value $\mathbb{E}[\mathcal{X}]$

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Write down:

- Probability for each outcome $P(\omega); \quad \forall \omega \in \Omega$
- The random variable for each outcome $\mathcal{X}(\omega); \quad \forall \omega \in \Omega$
- The expected value $\mathbb{E}[\mathcal{X}]$
- How much money the gambler loses after 1000 plays

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Write down:

- Probability for each outcome $P(\omega); \quad \forall \omega \in \Omega$
- The random variable for each outcome $\mathcal{X}(\omega); \quad \forall \omega \in \Omega$
- The expected value $\mathbb{E}[\mathcal{X}]$
- How much money the gambler loses after 1000 plays

Make sure the expected value is **negative but near zero**:

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Write down:

- Probability for each outcome $P(\omega); \quad \forall \omega \in \Omega$
- The random variable for each outcome $\mathcal{X}(\omega); \quad \forall \omega \in \Omega$
- The expected value $\mathbb{E}[\mathcal{X}]$
- How much money the gambler loses after 1000 plays

Make sure the expected value is **negative but near zero**:

- Negative: The gambler loses money and you win money

Bandits

Exercise: You start a new casino in Macau. Create a bandit with the following outcomes $\Omega \in \{\text{Win Lemon}, \text{Win Cherry}, \text{Win BAR}, \text{Lose}\}$

Write down:

- Probability for each outcome $P(\omega); \quad \forall \omega \in \Omega$
- The random variable for each outcome $\mathcal{X}(\omega); \quad \forall \omega \in \Omega$
- The expected value $\mathbb{E}[\mathcal{X}]$
- How much money the gambler loses after 1000 plays

Make sure the expected value is **negative but near zero**:

- Negative: The gambler loses money and you win money
- Near zero: The gambler wins sometimes and will continue to play

Multiarmed Bandits

Multiarmed Bandits

The bandit problem is useful for casino owners and gamblers

Multiarmed Bandits

The bandit problem is useful for casino owners and gamblers

But it is a trivial decision making problem

Multiarmed Bandits

The bandit problem is useful for casino owners and gamblers

But it is a trivial decision making problem

If $\mathbb{E}[\mathcal{X}] > 0$ you should gamble

Multiarmed Bandits

The bandit problem is useful for casino owners and gamblers

But it is a trivial decision making problem

If $\mathbb{E}[\mathcal{X}] > 0$ you should gamble

If $\mathbb{E}[\mathcal{X}] < 0$ you should not gamble

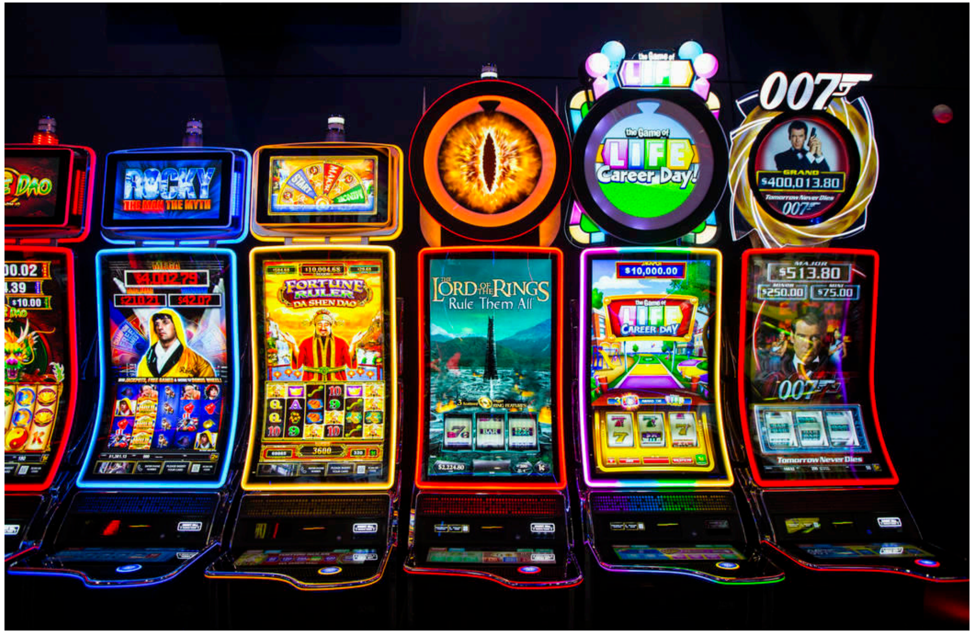
We will consider a more interesting problem

Multiarmed Bandits

You arrive at the Londoner with 1000 MOP and want to win money

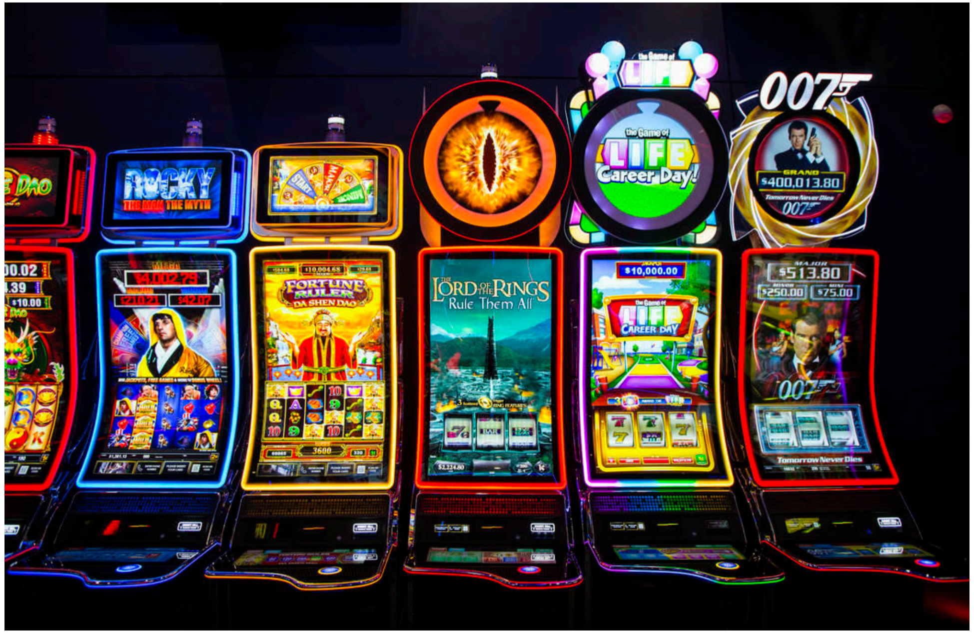
Multiarmed Bandits

You arrive at the Londoner with 1000 MOP and want to win money



Multiarmed Bandits

You arrive at the Londoner with 1000 MOP and want to win money



Question: Which machine do you play?

Multiarmed Bandits

We call this the **multi-armed bandit** problem

Multiarmed Bandits

We call this the **multi-armed bandit** problem

Multiarmed Bandits

We call this the **multi-armed bandit** problem

You don't know the expected value of each arm. Which should you pull?

Multiarmed Bandits

We can model many real problems as multiarmed bandits

Multiarmed Bandits

We can model many real problems as multiarmed bandits

For example, we can model hospital treatment as multiarmed bandits

Multiarmed Bandits

We can model many real problems as multiarmed bandits

For example, we can model hospital treatment as multiarmed bandits

We have new medicines, but do not know their effectiveness

Multiarmed Bandits

We can model many real problems as multiarmed bandits

For example, we can model hospital treatment as multiarmed bandits

We have new medicines, but do not know their effectiveness

Medicine A

Medicine B

Medicine C

Multiarmed Bandits

We can model many real problems as multiarmed bandits

For example, we can model hospital treatment as multiarmed bandits

We have new medicines, but do not know their effectiveness

Medicine A

Medicine B

Medicine C

We can find the best medicine while healing the most people

Multiarmed Bandits

YouTube, Youku, BiliBili, TikTok, Netflix use bandits to suggest videos

Multiarmed Bandits

YouTube, Youku, BiliBili, TikTok, Netflix use bandits to suggest videos

Dog videos

Gaming videos

Study videos

Multiarmed Bandits

YouTube, Youku, BiliBili, TikTok, Netflix use bandits to suggest videos

Dog videos

Gaming videos

Study videos

You are the bandit! The "money" is your ❤️

Multiarmed Bandits

YouTube, Youku, BiliBili, TikTok, Netflix use bandits to suggest videos

Dog videos

Gaming videos

Study videos

You are the bandit! The “money” is your ❤️

You like a specific type of video, but TikTok does not know what it is

Multiarmed Bandits

YouTube, Youku, BiliBili, TikTok, Netflix use bandits to suggest videos

Dog videos

Gaming videos

Study videos

You are the bandit! The “money” is your ❤️

You like a specific type of video, but TikTok does not know what it is

TikTok tries to find your favorite video category

Multiarmed Bandits

Problem: We have k bandits, and each bandit is a random variable

$$\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_k$$

Multiarmed Bandits

Problem: We have k bandits, and each bandit is a random variable

$$\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_k$$

We do not know $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

Multiarmed Bandits

Problem: We have k bandits, and each bandit is a random variable

$$\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_k$$

We do not know $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

You can take an **action** by pulling the arm of each bandit

$$a \in \{1, 2, ..., k\}$$

Multiarmed Bandits

Problem: We have k bandits, and each bandit is a random variable

$$\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_k$$

We do not know $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

You can take an **action** by pulling the arm of each bandit

$$a \in \{1, 2, ..., k\}$$

Which actions should you take to make the most money?

Multiarmed Bandits

Problem: We have k bandits, and each bandit is a random variable

$$\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_k$$

We do not know $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

You can take an **action** by pulling the arm of each bandit

$$a \in \{1, 2, ..., k\}$$

Which actions should you take to make the most money?

Question: How should we approach this problem?

Multiarmed Bandits

This is a hard problem!

Multiarmed Bandits

This is a hard problem!

We need to estimate $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

Multiarmed Bandits

This is a hard problem!

We need to estimate $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

But we do not have enough money to perfectly estimate all k bandits

Multiarmed Bandits

This is a hard problem!

We need to estimate $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

But we do not have enough money to perfectly estimate all k bandits

We must be careful in how we choose $a \in 1...k$

Multiarmed Bandits

This is a hard problem!

We need to estimate $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

But we do not have enough money to perfectly estimate all k bandits

We must be careful in how we choose $a \in 1...k$

We want to:

Multiarmed Bandits

This is a hard problem!

We need to estimate $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

But we do not have enough money to perfectly estimate all k bandits

We must be careful in how we choose $a \in 1...k$

We want to:

- Pick a to approximate bandits $\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$

Multiarmed Bandits

This is a hard problem!

We need to estimate $\mathbb{E}[\mathcal{X}_1], \mathbb{E}[\mathcal{X}_2], ..., \mathbb{E}[\mathcal{X}_k]$

But we do not have enough money to perfectly estimate all k bandits

We must be careful in how we choose $a \in 1...k$

We want to:

- Pick a to approximate bandits $\quad \mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$
- Pick a to make the most money $\quad \arg\max_{a \in 1...k} \mathbb{E}[\mathcal{X}_a]$

Multiarmed Bandits

We have names for each goal

Multiarmed Bandits

We have names for each goal

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Multiarmed Bandits

We have names for each goal

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Explore our options to improve our estimate of each expectation

Multiarmed Bandits

We have names for each goal

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1 ... k]$$

Explore our options to improve our estimate of each expectation

Exploitation:

$$\arg \max_{a \in 1 ... k} \mathbb{E}[\mathcal{X}_a]$$

Multiarmed Bandits

We have names for each goal

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Explore our options to improve our estimate of each expectation

Exploitation:

$$\arg\max_{a \in 1...k} \mathbb{E}[\mathcal{X}_a]$$

Use our estimates to make money

Multiarmed Bandits

We have names for each goal

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Explore our options to improve our estimate of each expectation

Exploitation:

$$\arg\max_{a \in 1...k} \mathbb{E}[\mathcal{X}_a]$$

Use our estimates to make money

It is important you understand this! Any questions?

Multiarmed Bandits

Question: How can we choose a to achieve each goal?

Multiarmed Bandits

Question: How can we choose a to achieve each goal?

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Explore our options to improve our estimate of each expectation

Exploitation:

$$\underset{a \in 1...k}{\arg\max}\, \mathbb{E}[\mathcal{X}_a]$$

Use our estimates to make money

Multiarmed Bandits

Question: How can we choose a to achieve each goal?

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Explore our options to improve our estimate of each expectation

$$a \sim \text{uniform}(\{1...k\})$$

Exploitation:

$$\underset{a \in 1...k}{\arg\max}\, \mathbb{E}[\mathcal{X}_a]$$

Use our estimates to make money

Multiarmed Bandits

Question: How can we choose a to achieve each goal?

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Explore our options to improve our estimate of each expectation

$$a \sim \text{uniform}(\{1...k\})$$

Exploitation:

$$\underset{a \in 1...k}{\arg\max}\, \mathbb{E}[\mathcal{X}_a]$$

Use our estimates to make money

$$a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Multiarmed Bandits

Question: How can we choose a to achieve each goal?

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Explore our options to improve our estimate of each expectation

$$a \sim \text{uniform}(\{1...k\})$$

Exploitation:

$$\underset{a \in 1...k}{\arg\max}\, \mathbb{E}[\mathcal{X}_a]$$

Use our estimates to make money

$$a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Question: How can we achieve both goals at once?

Multiarmed Bandits

Question: How can we choose a to achieve each goal?

Exploration:

$$\mathbb{E}[\mathcal{X}_a \mid a \in 1...k]$$

Explore our options to improve our estimate of each expectation

$$a \sim \text{uniform}(\{1...k\})$$

Exploitation:

$$\underset{a \in 1...k}{\arg\max}\, \mathbb{E}[\mathcal{X}_a]$$

Use our estimates to make money

$$a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Question: How can we achieve both goals at once?

Answer: Sometimes choose a to explore, sometimes choose a to exploit

Multiarmed Bandits

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < 0.5 \text{ then } a \sim \text{uniform}(\{1 ... k\})$$

$$\text{if } u \geq 0.5 \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Multiarmed Bandits

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < 0.5 \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq 0.5 \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Half the time we explore, half the time we exploit

Multiarmed Bandits

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < 0.5 \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq 0.5 \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Half the time we explore, half the time we exploit

We can change the explore/exploit ratio using a parameter ε

Multiarmed Bandits

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < 0.5 \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq 0.5 \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Half the time we explore, half the time we exploit

We can change the explore/exploit ratio using a parameter ε

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < \varepsilon \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq \varepsilon \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Multiarmed Bandits

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < 0.5 \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq 0.5 \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Half the time we explore, half the time we exploit

We can change the explore/exploit ratio using a parameter ε

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < \varepsilon \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq \varepsilon \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Multiarmed Bandits

$$\varepsilon \in [0, 1]$$

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < \varepsilon \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq \varepsilon \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

Multiarmed Bandits

$$\varepsilon \in [0, 1]$$

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < \varepsilon \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq \varepsilon \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

We call this **epsilon greedy** because we are greedy with proportion ε

Multiarmed Bandits

$$\varepsilon \in [0, 1]$$

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < \varepsilon \text{ then } a \sim \text{uniform}(\{1 ... k\})$$

$$\text{if } u \geq \varepsilon \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

We call this **epsilon greedy** because we are greedy with proportion ε

Question: When should $\varepsilon \approx 1$? When should $\varepsilon \approx 0$?

Multiarmed Bandits

$$\varepsilon \in [0, 1]$$

$$u \sim \text{uniform}([0, 1])$$

$$\text{if } u < \varepsilon \text{ then } a \sim \text{uniform}(\{1...k\})$$

$$\text{if } u \geq \varepsilon \text{ then } a = \arg\max(\mathbb{E}[\mathcal{X}_a])$$

We call this **epsilon greedy** because we are greedy with proportion ε

Question: When should $\varepsilon \approx 1$? When should $\varepsilon \approx 0$?

Answer:

- $\varepsilon \approx 1$ when we trust our estimates $\mathbb{E}[\mathcal{X}]$
- $\varepsilon \approx 0$ when we do not trust our estimates

Multiarmed Bandits

Question: Do we use epsilon greedy in medicine?

Multiarmed Bandits

Question: Do we use epsilon greedy in medicine?

Answer: Yes!

Multiarmed Bandits

Question: Do we use epsilon greedy in medicine?

Answer: Yes!

- Usually give patients drug A that we know works (exploit)

Multiarmed Bandits

Question: Do we use epsilon greedy in medicine?

Answer: Yes!

- Usually give patients drug A that we know works (exploit)
- Sometimes test new drug B on patients (explore)

Multiarmed Bandits

Question: Do we use epsilon greedy in medicine?

Answer: Yes!

- Usually give patients drug A that we know works (exploit)
- Sometimes test new drug B on patients (explore)

Question: Does TikTok or BiliBili use epsilon greedy?

Multiarmed Bandits

Question: Do we use epsilon greedy in medicine?

Answer: Yes!

- Usually give patients drug A that we know works (exploit)
- Sometimes test new drug B on patients (explore)

Question: Does TikTok or BiliBili use epsilon greedy?

Answer: Yes!

Multiarmed Bandits

Question: Do we use epsilon greedy in medicine?

Answer: Yes!

- Usually give patients drug A that we know works (exploit)
- Sometimes test new drug B on patients (explore)

Question: Does TikTok or BiliBili use epsilon greedy?

Answer: Yes!

- If you watch dog videos, it usually suggests more dog videos

Multiarmed Bandits

Question: Do we use epsilon greedy in medicine?

Answer: Yes!

- Usually give patients drug A that we know works (exploit)
- Sometimes test new drug B on patients (explore)

Question: Does TikTok or BiliBili use epsilon greedy?

Answer: Yes!

- If you watch dog videos, it usually suggests more dog videos
- Sometimes it suggests study videos

Coding

Coding

Let us code some multiarmed bandits!

Coding

Let us code some multiarmed bandits!

https://colab.research.google.com/drive/1cyNLRa-J8oe7pgy_gs2mcypZPqqaquoa