SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

18AIC304J-REINFORCEMENT LEARNING TECHNIQUES

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

UNIT-1

Introduction to Reinforcement Learning and Examples, Elements of Reinforcement Learning - Limitations and Scope, Tic-Tac-Toe example and History of RL, Probability concepts and Axioms of probability, Concepts of a random variable: PMF, PDFs, CDFs, Expectation, Concepts of joint and multiple random variables, joint, conditional, marginal distributions, Correlation and independence, An-Armed Bandit Problem and Action-Value Methods

Introduction to Reinforcement Learning

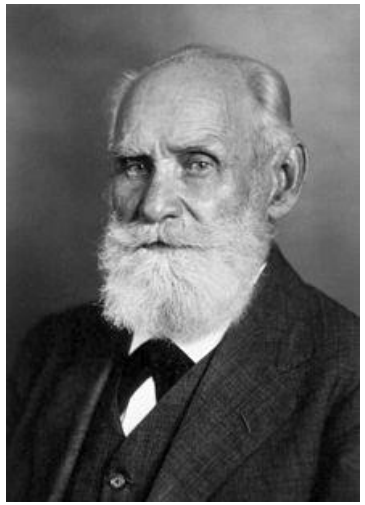
"**Goal-directed** learning from **interaction** in an **uncertain** environment"
– Reinforcement Learning: An Introduction, Sutton and Barto, Second Edition

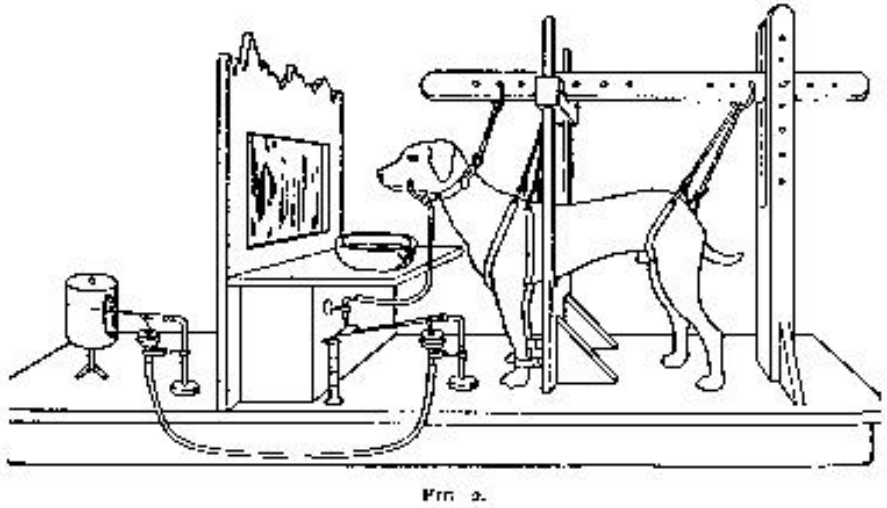
Introduction to Reinforcement Learning

Definition of Learning from Psychology: Learning is a relatively **permanent** change in behavior that occurs through **experience**. It's a **continuous** and **gradual** process.

Some more of Psychology

Classical Conditioning: Theory developed by Ivan Pavlov (Nobel prize in 1904).



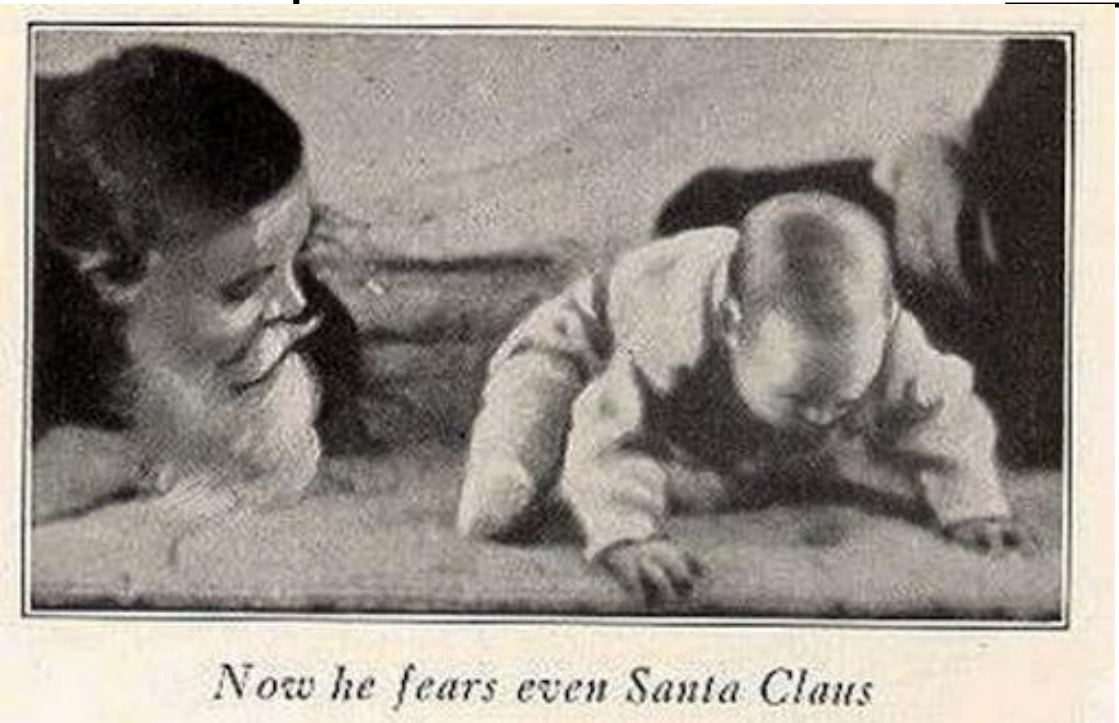


- Dog salivates seeing the food (Unconditioned response (salivating) to Unconditioned Stimulus (food))
- Dog salivates hearing the bell (Conditioned response (salivating) to Conditioned Stimulus (bell)).
- Learning to associate bell to salivation via food.

Some more of Psychology

Classical Conditioning: In an experiment with human babies, it was learnt that classical conditioning can occur in humans too.

John B. Watson in 1920 performed his famous "Little Albert experiment"



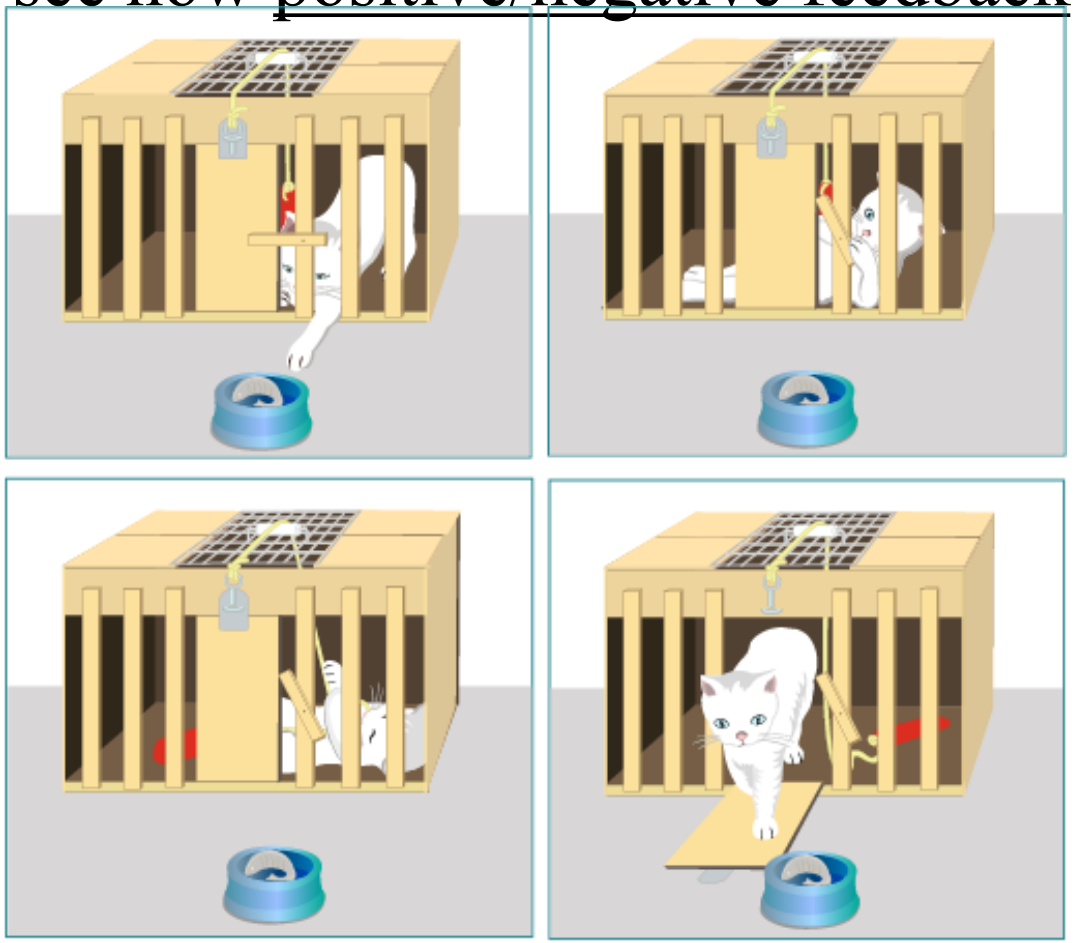
Now he fears even Santa Claus

Mary Cover Jones in 1924 performed her "Little Peter experiment" to show that "counter conditioning" is possible for human subjects.

Some more of Psychology

In Classical Conditioning: The reward (e.g., food) is present and that caused the response (e.g., salivation).

At a similar time, Edward Thorndike was conducting experiments on cats to see how positive/negative feedback affects goal directed learning.



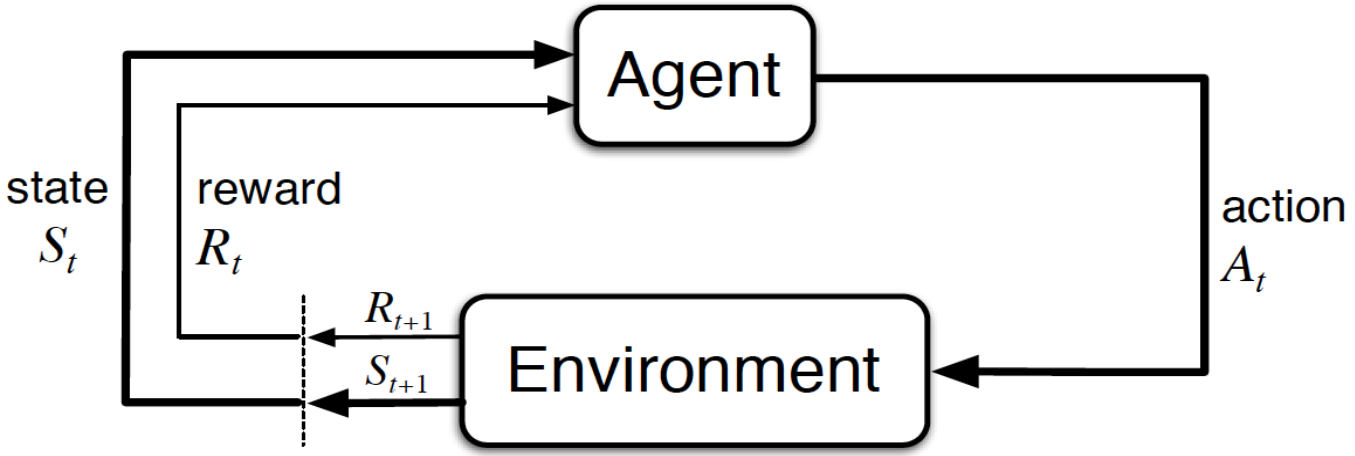
Some more of Psychology

- **Reinforcement/Punishment:-** In Psychological terms, **Reinforcement** is a consequence that causes a behavior to occur with greater frequency. **Punishment** is a response or consequence that causes a behavior to occur with less frequency.

- Reinforcement can also mean removal of aversive consequence (Negative Reinforcement).

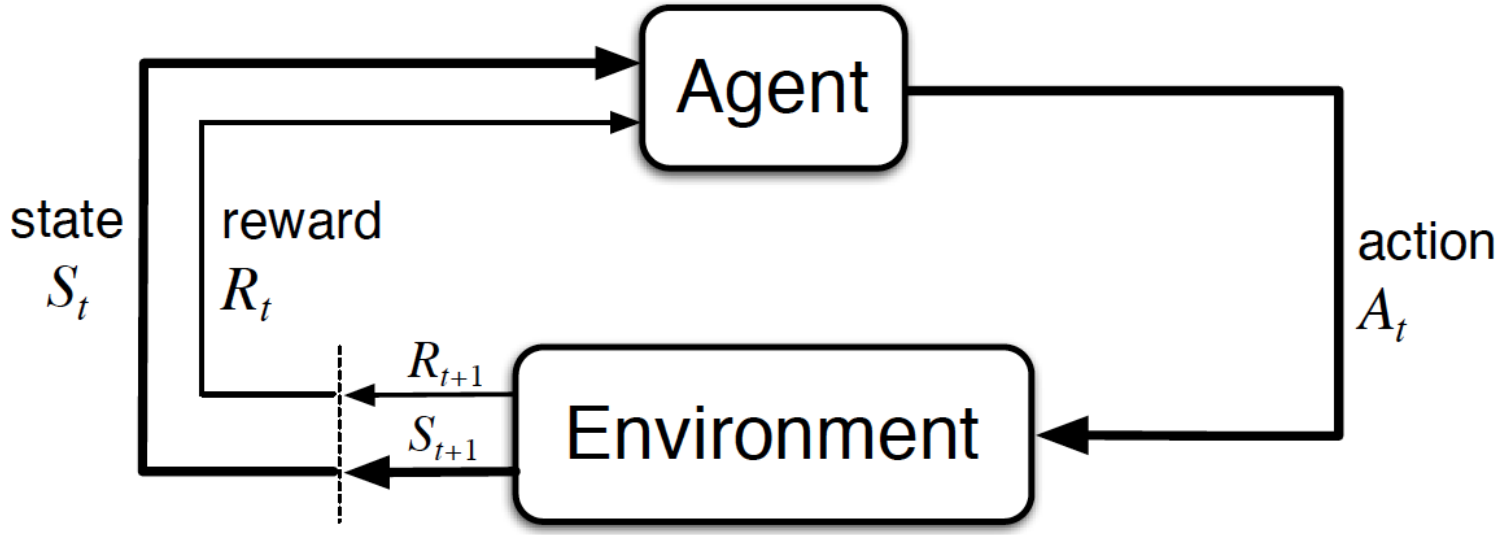
- Similarly, Punishment can also mean removal of rewarding consequence (Negative Punishment).

Reinforcement Learning Setting



- **Agent:-** The learner or the decision maker.
- **Environment:-** The thing the learner interacts with, comprising everything outside the agent.
- They interact continually. The agent selects actions. The environment responds to these actions by presenting new situation and giving rewards for the action.

Reinforcement Learning Setting



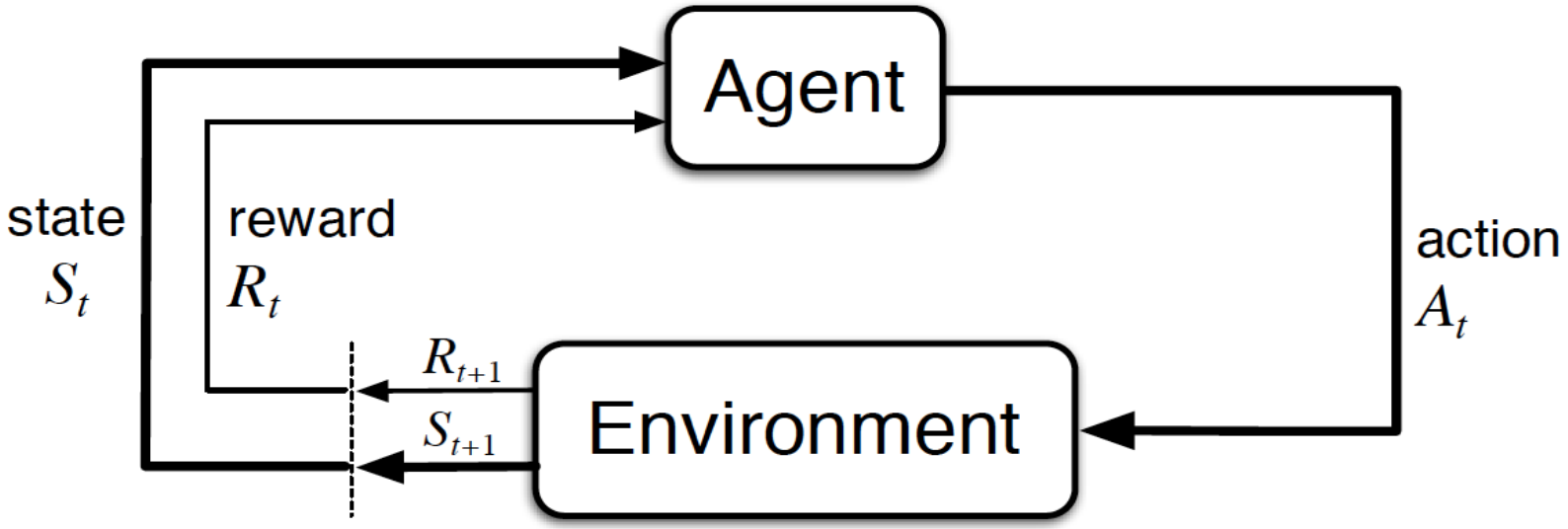
Reward:- Rewards are scalar measures defining what are the good and bad events for the agent.

Value:- Value of a state is the total amount of reward an agent is expected to get in future starting from that state.

To make a human analogy, rewards are somewhat like pleasure (if high) and pain (if low), whereas values correspond to a more refined and farsighted judgment of how pleased or displeased we are that our environment is in a particular state.

Sutton and

Reinforcement Learning Setting



Goal in RL Problem:- to maximize the total reward "in expectation" over the long run.

$$\tau \stackrel{\text{def}}{=} (s_1, a_1, s_2, a_2, \dots), p(\tau) = p(s_1) \prod_t p(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

$$\max \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_t R(s_t, a_t) \right]$$

Distinguishing Features of Reinforcement Learning

- **Trial and error:-** The learner is not told which actions to take, but instead must discover which actions are most rewarding by trying them.

- **Delayed rewards:-** actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards.

- **Exploration-vs-exploitation**:- To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But to discover such actions, it has to try actions that it has not selected before. - *Sutton and Barto*

When to use Reinforcement Learning

Trajectories:- Data comes in the form trajectories.

Decisions:- Need to take decisions that affect the trajectory data.

Feedback:- Need to get feedback about the choice of actions.

Supervised, Unsupervised and Reinforcement Learning

Supervised Learning:- Learn $y = f(x)$ – You are given a bunch of (x, y) pairs and your goal is to find the function f mapping x to y.

Unsupervised Learning:- Learn $f(x)$ – You don't have access to any y, you are given a bunch of $x's$ and your goal is to find some f that gives a "compact description" of these $x's$.

Reinforcement Learning:- Learn $y = f(x)$, given z – You are given a bunch of (x, z) pairs and your goal is to find the function f mapping x to y.

x is state, y is action, z is reward

Supervised, Unsupervised and Reinforcement Learning

Supervised Learning:- Does not involve the problem of temporal credit assignment and exploration.

Unsupervised Learning:- Here, in addition, the correct labels are not available

Limitations of Reinforcement learning

- Too much of reinforcement may cause an overload which could weaken the results.
- Reinforcement learning is preferred for solving complex problems, not simple ones.
- It requires plenty of data and involves a lot of computation.
- Maintenance cost is high

CHATGPT_EXAMPLE OF RL

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

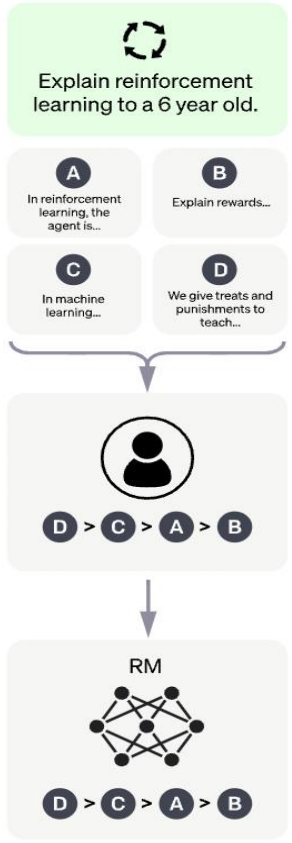
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

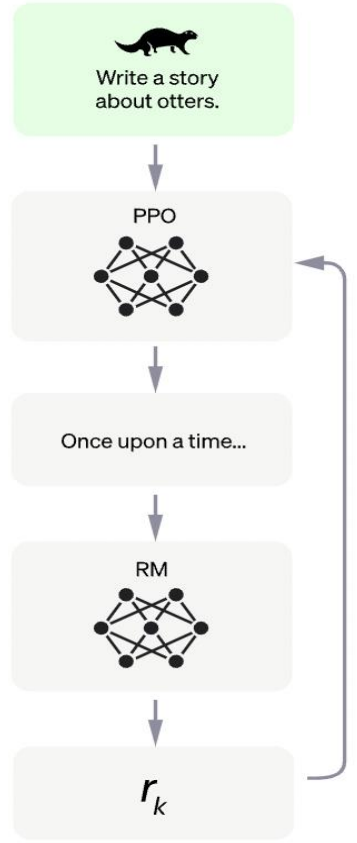
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

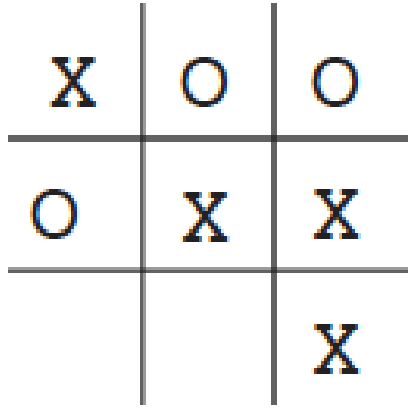
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



An Extended Example: Tic-Tac-Toe

The objective of Tic-Tac-Toe is to be the first to place their marks (either cross or naughts) in a horizontal, vertical, or diagonal arrangement. Now, let us define the game in terms of RL keywords mentioned earlier:



1.Agents involve 2 Tic-Tac-Toe players who attempt to outwit each other by taking a turn to place their mark,

2.Reward refers to an arbitrary value earned by the winning agent,

3.Actions dictate that each agent is allowed to place their corresponding mark only in an empty box,

4.The state is the configuration of the tic-tac-toe board after each turn until the game ends in either a win or a draw.

An Extended Example: Tic-Tac-Toe

The State

- The state of this game is the board state of both the agent and its opponent, so we will **initialise a 3x3 board with zeros indicating available positions and update positions with 1 if player 1 takes a move and -1 if player 2 takes a move.**
- Here, 4 possible break outcomes can be achieved after each playing turn: *column win, row win, diagonal win, and draw*.
- 1 point to the winning player, 0 to the losing player, and 0.1–0.5 to both in the case of a draw).

An Extended Example: Tic-Tac-Toe

The Agent

- To specify their behaviors (choosing an action that maximizes reward, calculating reward)
- Deals with hyperparmeters like learning rate, decay gamma and exploration rate.

1.learning rate: The speed by which learning is done by each agent,

2.decay gamma The factor by which reward decays per turn (ie. forces agent to win in the least possible moves, as winning with more steps result in a *decaying* reward,

3.Exploration rate which allows an agent to choose a random position available rather than greedily finding (exploiting) the best possible next move from the current collection of learned policies. Higher value permits more randomness in the agent's decision-making.

An Extended Example: Tic-Tac-Toe

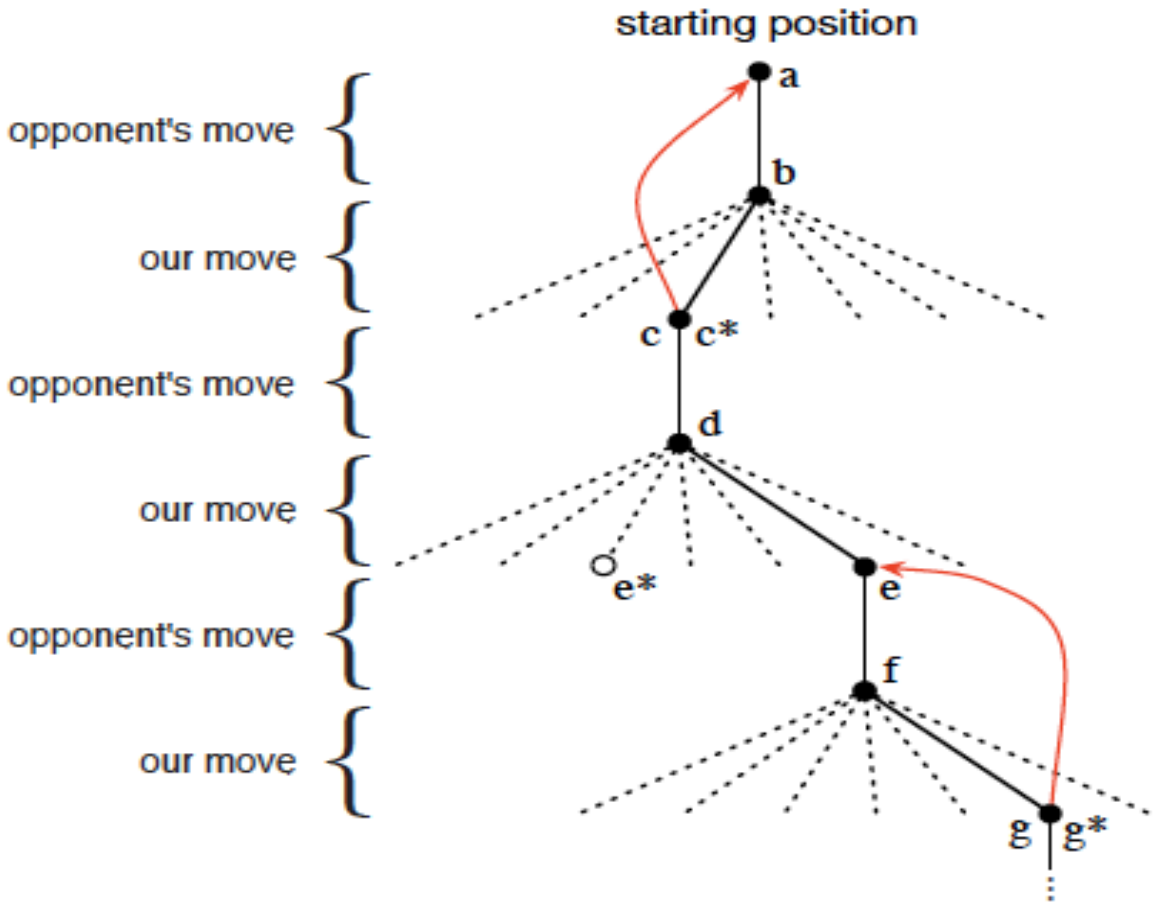


Figure 1.1: A sequence of tic-tac-toe moves. The solid black lines represent the moves taken during a game; the dashed lines represent moves that we (our reinforcement learning player) considered but did not make. Our second move was an exploratory move, meaning that it was taken even though another sibling move, the one leading to e^*, was ranked higher. Exploratory moves do not result in any learning, but each of our other moves does, causing updates as suggested by the red arrows in which estimated values are moved up the tree from later nodes to earlier nodes as detailed in the text.

Early History of Reinforcement Learning

Optimal Policy/Optimal Control [Pontryagin and Neustadt, 1962]

Dynamic Programming
[Bellman, 1959] and many more
Optimal solution to global problem is conjunction of solution to sub-problems
- Model knowledge is necessary
- Curse of dimensionality – Computation increases exponentially with number of states
- Has good convergence properties

Learning by Trial and Error
[Thorndike 1911, Skinner 1953]
[Turing 1948, Pleasure and Pain System, 1948]
[Minsky, PhD Thesis, 1954]
[Michie, 1961, 1968, 1974]
[Widrow, Gupta and Maitra, 1973] and many more
Continued to show importance of trial and error learning

Reinforcement Learning
[Sutton, Barto, Arthur Samuel, Bertsekas, Tsitsiklis and many more]

Advanced Topics

- **Inverse Reinforcement Learning:-** Learning the reward function.

  How can a learner *who does not know what there is to learn* manage to learn anyway?

  "The Development of Embodied Cognition: Six Lessons from Babies" - Linda Smith

- **Meta Reinforcement Learning:-** Learning to reinforcement learn.

  Babies explore – they move and act in highly variable and playful ways that are not goal-oriented and are seemingly random. In doing so, they discover new problems and new solutions. Exploration makes intelligence open-ended and inventive.

  "The Development of Embodied Cognition: Six Lessons from Babies" - Linda Smith

- Safe Reinforcement Learning:- When we do not have much opportunity to explore safely.

PROBABILITY CONCEPTS

x Probability theory is the study of uncertainty.

x The mathematical treatise of probability is very sophisticated, and delves into a branch of analysis known as measure theory.

§ Probability is the Mathematical language for quantifying *uncertainty*. The starting point is to specify random experiments, sample space and set of outcomes.

§ A **random experiment** is an experiment in which the outcome varies in an unpredictable fashion when the experiment is repeated under the same conditions.

§ An **outcome** is a result of the random experiment and it can not be decomposed in terms of other results. The **sample space** of a random experiment is defined as the se of all possible outcomes. An outcome and the sample space of a random experiment will be denoted as ζ and S respectively.

- Examples of random experiment
  - Flipping a coin
  - Rolling a die
  - Flipping a coin twice
  - Pick a number X at random between zero and one, then pick a number Y at random between zero and X.
- The corresponding sample spaces will be
  - $S_1 = \{H, T\}$
  - $S_2 = \{1, 2, 3, 4, 5, 6\}$
  - $S_3 = \{HH, HT, TH, TT\}$
  - $S_4 = \{(x, y) : 0 \leq y \leq x \leq 1\}$.

§ Any subset E of the sample space S is known as an **event**. We, sometimes, are no interested in the occurrence of specific outcomes but rather in the occurrence of combination of a few outcomes.

This requires that we consider subsets of S

- Getting even number when rolling a die, $E_2 = \{2, 4, 6\}$
- Number of heads equal to number of tails when flipping a coin twice, $E_3 = \{HT, TH\}$
- Two numbers differ by less than 1/10, $E_4 = \{(x, y) : 0 \leq y \leq x \leq 1 \text{ and } |x - y| < 1/10\}$.

§ We say that an event E occurs if the outcome ζ is in E

§ Three events are of special importance.

- **Simple event** are the outcomes of random experiments.
- **Sure event** is the sample space S which consists of all outcomes and hence alway occurs.
- **Impossible** or **null event** φ which contains no outcomes and hence never occurs.

- **Set of events** (or **event space**) F: A set whose elements are subsets of the sample space (*i.e.*, events). F = $\{A : A \subseteq S\}$. F is really a "set of sets".
- F should satisfy the following three properties.
  - $\varphi \in$ F
  - $A \in$ F $\Rightarrow$ A^c(, $S \setminus A$) $\in$ F
  - $A_1, A_2, \cdots \in$ F $\Rightarrow$ $\cup_i A_i \in$ F

- Probabilities are numbers assigned to events of $\mathcal{F}$ that indicate how "likely" it is that the events will occur when a random experiment is performed.
- Let a random experiment has sample space S and event space $\mathcal{F}$. Probability of an event A is a function $P : \mathcal{F} \rightarrow \mathbb{R}$ that satisfies the following properties
  - $P(A) \geq 0, \ \forall A \in \mathcal{F}$
  - $P(S) = 1$
  - If $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint events (*i.e.*, $A_i \cap A_j = \phi$ for $i \neq j$) then, $P(\cup_i A_i) = \sum_i P(A_i)$
- These three properties are called the **Axioms of Probability**.

§ Properties

- $P(A^c) = 1 - P(A)$
- $P(A) \leq 1$
- $P(\varphi) = 0$
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$
- $P(A \cup B) \leq P(A) + P(B)$

Events

- The ***probability of an event*** is the **sum** of the probabilities of the simple events that constitute the event.
- E.g. (assuming a fair die) S = {1, 2, 3, 4, 5, 6} and
- P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6
- Then:
- P(EVEN) = P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2

Conditional Probability

- ***Conditional probability*** is used to determine how two events are related; that is, we can determine the probability of one event ***given*** the occurrence of another related event.
- Experiment: random select one student in class.
- P(randomly selected student is male) =
- P(randomly selected student is male/student is on 3rd row) =
- Conditional probabilities are written as **P(A | B)** and read as "the probability of A *given* B" and is calculated as

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

- P(A and B) = P(A)*P(B/A) = P(B)*P(A/B) both are true
- Keep this in mind!

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

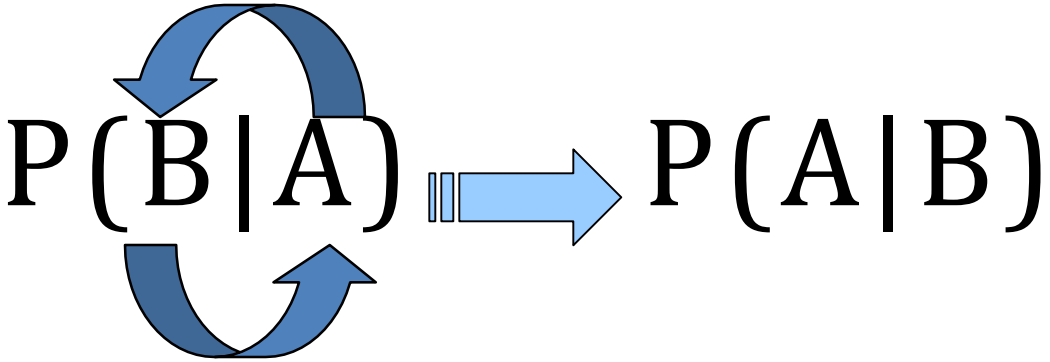
$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

Bayes' Law

- Bayes' Law is named for Thomas Bayes, an eighteenth century mathematician.

- In its most basic form, **if we know** P(B | A),

- we can apply Bayes' Law to determine P(A | B)



- The probabilities P(A) and P(A^C) are called ***prior probabilities*** because they are determined ***prior*** to the decision about taking the preparatory course.
- The conditional probability P(A | B) is called a ***posterior probability*** (or revised probability), because the prior probability is revised ***after*** the decision about taking the preparatory course.

Total probability theorem

- Take events A_i for I = 1 to k to be:
  - Mutually exclusive: $A_i \cap A_j = 0$ for all i,j
  - Exhaustive: $A_1 \cup \cdots \cup A_k = S$

For any event B on S

$$p(B) = p(B|A_1)\,p(A_1) + \cdots + p(B|A_k)\,p(A_k)$$

$$p(B) = \sum_{i=1}^{k} p(B|A_i)\,p(A_i)$$

Bayes theorem follows

$$p(A_j|B) = \frac{p(A_j \cap B)}{p(B)} = \frac{p(B|A_j)\cdot p(A)}{\sum_{i=1}^{k} p(B|A_i)\,p(A_i)}$$

Independence of events

- Do A and B depend on one another?
  - Yes! B more likely to be true if A.
  - A should be more likely if B.
- If Independent

$$p(A \cap B) = p(A) \cdot p(B)$$

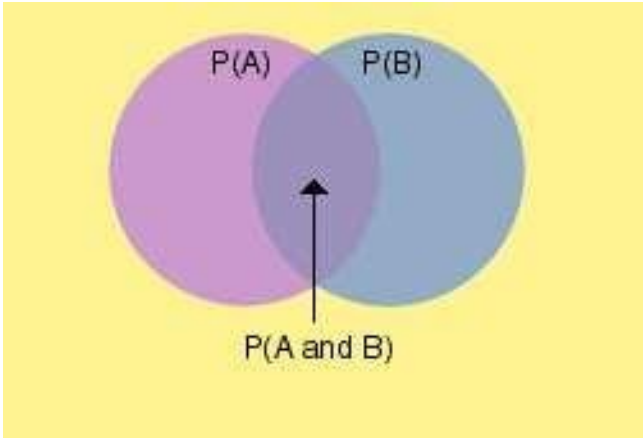
$$p(A|B) = p(A) \quad p(B|A) = p(B)$$

- If Dependent

$$p(A \cap B) \neq p(A) \cdot p(B)$$

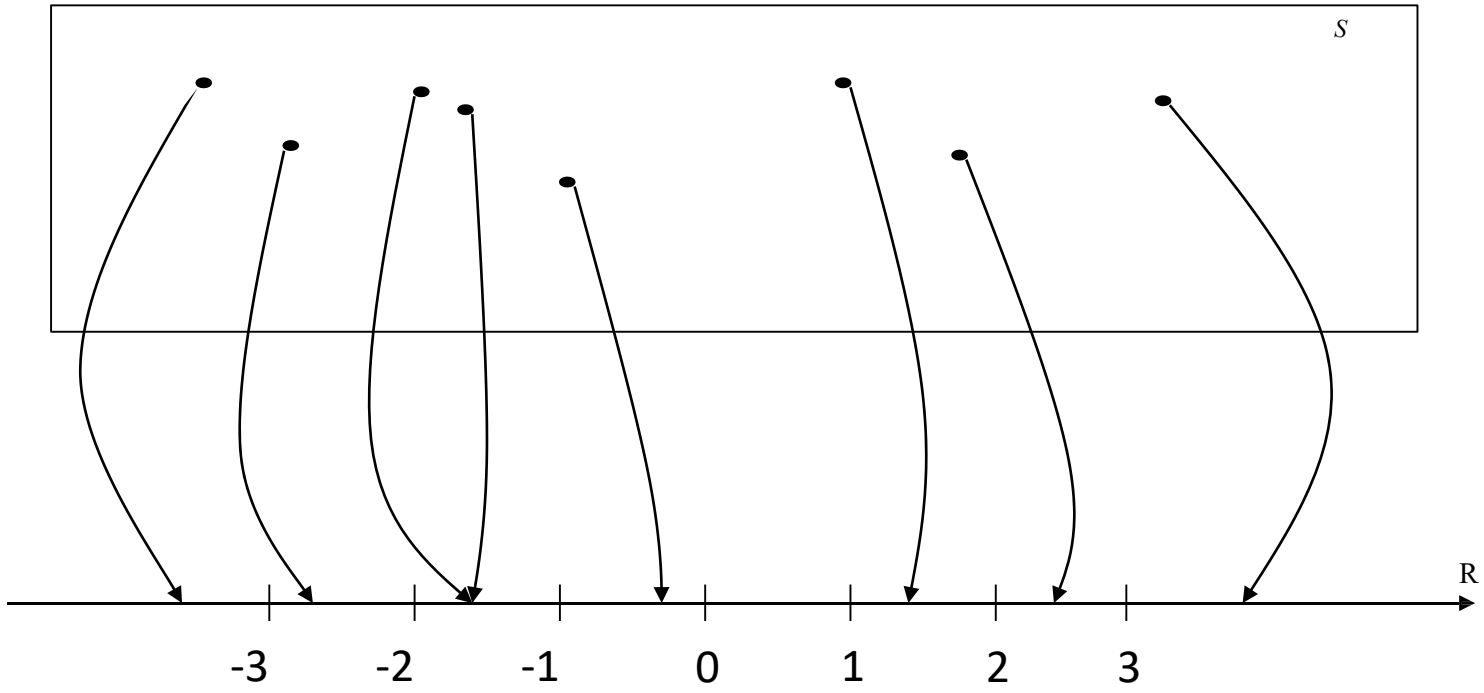
$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

$$p(A \cap B) = p(B|A) \cdot p(A)$$



Random variable

- Random variable
  - A numerical value to each outcome of a particular experiment



- Example 1 : Machine Breakdowns
  - —Sample space : $S = \{electrical, mechanical, misuse\}$
  - – Each of these failures may be associated with a repair cost
  - – State space : $\{50, 200, 350\}$
  - – Cost is a random variable : 50, 200, and 350
- Probability Mass Function (p.m.f.)
  - – A set of probability value assigned to each of the values taken by the discrete random variable x_i
  - – $0 \le p_i \le 1$ and $\sum_i p_i = 1$
  - – Probability : $P(X = x_i) = p_i$

Continuous and Discrete random variables

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

- **Distribution function**:

$$F_X(x) = P(X \leq x),\ -\infty < x < \infty$$

- If $F_X(x)$ is a continuous function of x, then X is a continuous random variable.
  - $F_X(x)$: discrete in x → Discrete rv's
  - $F_X(x)$: piecewise continuous → Mixed rv's
  - **PROPERTIES:**
- . $0 \leq F_X(x) \leq 1,\ -\infty < x < \infty$
  . $F_X(x)$: monotonically increasing func. of x
  . $x \overset{lim}{\to} -\infty\ F_X(x) = 0$ and $x \overset{lim}{\to} \infty\ F_X(x) = 1$

Probability Density Function (pdf)

- *X* : continuous rv, then, $f(x) = \frac{dF(x)}{dx}$ is the pdf of X.

$$CDF \longleftrightarrow pdf$$

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)du, \quad -\infty < x < \infty$$

$$P(X \in (a, b]) = P(a < X \leq b) = \int_{a}^{b} f_X(u)du.$$

- ***pdf* properties**:

1. $f(x) \geq 0$ for all x.

2. $\int_{-\infty}^{\infty} f(x)dx = 1.$

$$F(t) = \int_{-\infty}^{t} f(x)dx$$

$$= \int_{0}^{t} f(x)dx \quad ,$$

Binomial

- Suppose that the probability of success is p
- What is the probability of failure?

  $q = 1 - p$
- Examples
  - Toss of a coin (S = head): $p = 0.5 \Rightarrow q = 0.5$
  - Roll of a die (S = 1): $p = 0.1667 \Rightarrow q = 0.8333$
  - Fertility of a chicken egg (S = fertile): $p = 0.8 \Rightarrow q = 0.2$

Binomial

- Imagine that a trial is repeated *n* times

- Examples
  - A coin is tossed 5 times
  - A die is rolled 25 times
  - 50 chicken eggs are examined

- Assume *p* remains constant from trial to trial and that the trials are statistically independent of each other
- Example
  - What is the probability of obtaining 2 heads from a coin that was tossed 5 times?

$P(HHTTT) = (1/2)^5 = 1/32$

Poisson

- When there is a large number of trials, but a small probability of success, binomial calculation becomes impractical
  - Example: Number of deaths from horse kicks in the Army in different years

- The mean number of successes from n trials is $\mu = np$
- Example: 64 deaths in 20 years from thousands of soldiers If

we substitute μ/n for p, and let n tend to infinity, the binomial distribution becomes the Poisson distribution:

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

poisson

- Poisson distribution is applied where random events in space or time are expected to occur
- Deviation from Poisson distribution may indicate some degree of non-randomness in the events under study
- Investigation of cause may be of interest

Exponential Distribution

The random variable X that equals the distance between successive counts of a Poisson process with mean $\lambda > 0$ is an **exponential random variable** with parameter λ. The probability density function of X is

$$f(x) = \lambda e^{-\lambda x} \quad \text{for} \quad 0 \leq x < \infty \tag{4-14}$$

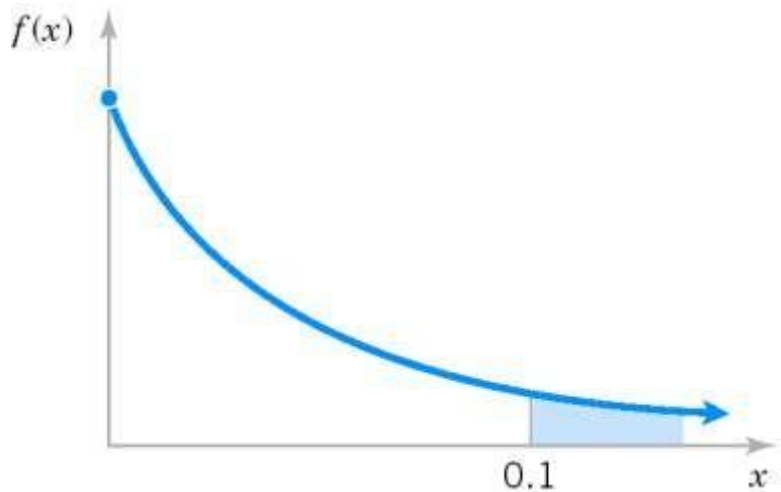
If the random variable X has an exponential distribution with parameter λ,

$$\mu = E(X) = \frac{1}{\lambda} \quad \text{and} \quad \sigma^2 = V(X) = \frac{1}{\lambda^2} \tag{4-15}$$

In a large corporate computer network, user log-ons to the system can be modeled as a Poisson process with a mean of 25 log-ons per hour. What is the probability that there are no log-ons in an interval of 6 minutes?

Let X denote the time in hours from the start of the interval until the first log-on. Then, X has an exponential distribution with $\lambda = 25$ log-ons per hour. We are interested in the probability that X exceeds 6 minutes. Because λ is given in log-ons per hour, we express all time units in hours. That is, 6 minutes $= 0.1$ hour. The probability requested is shown as the shaded area under the probability density function in Fig. 4-23. Therefore,

$$P(X > 0.1) = \int_{0.1}^{\infty} 25e^{-25x}\,dx = e^{-25(0.1)} = 0.082$$



Also, the cumulative distribution function can be used to obtain the same result as follows:

$$P(X > 0.1) = 1 - F(0.1) = e^{-25(0.1)}$$

An identical answer is obtained by expressing the mean number of log-ons as 0.417 log-ons per minute and computing the probability that the time until the next log-on exceeds 6 minutes. Try it.

What is the probability that the time until the next log-on is between 2 and 3 minutes? Upon converting all units to hours,

$$P(0.033 < X < 0.05) = \int_{0.033}^{0.05} 25e^{-25x}\,dx = -e^{-25x}\Big|_{0.033}^{0.05} = 0.152$$

An alternative solution is

$$P(0.033 < X < 0.05) = F(0.05) - F(0.033) = 0.152$$

Determine the interval of time such that the probability that no log-on occurs in the i
val is 0.90. The question asks for the length of time x such that $P(X > x) = 0.90$. Now,

$$P(X > x) = e^{-25x} = 0.90$$

Take the (natural) log of both sides to obtain $-25x = \ln(0.90) = -0.1054$. Therefore,

$$x = 0.00421 \text{ hour} = 0.25 \text{ minute}$$

Furthermore, the mean time until the next log-on is

$$\mu = 1/25 = 0.04 \text{ hour} = 2.4 \text{ minutes}$$

The standard deviation of the time until the next log-on is

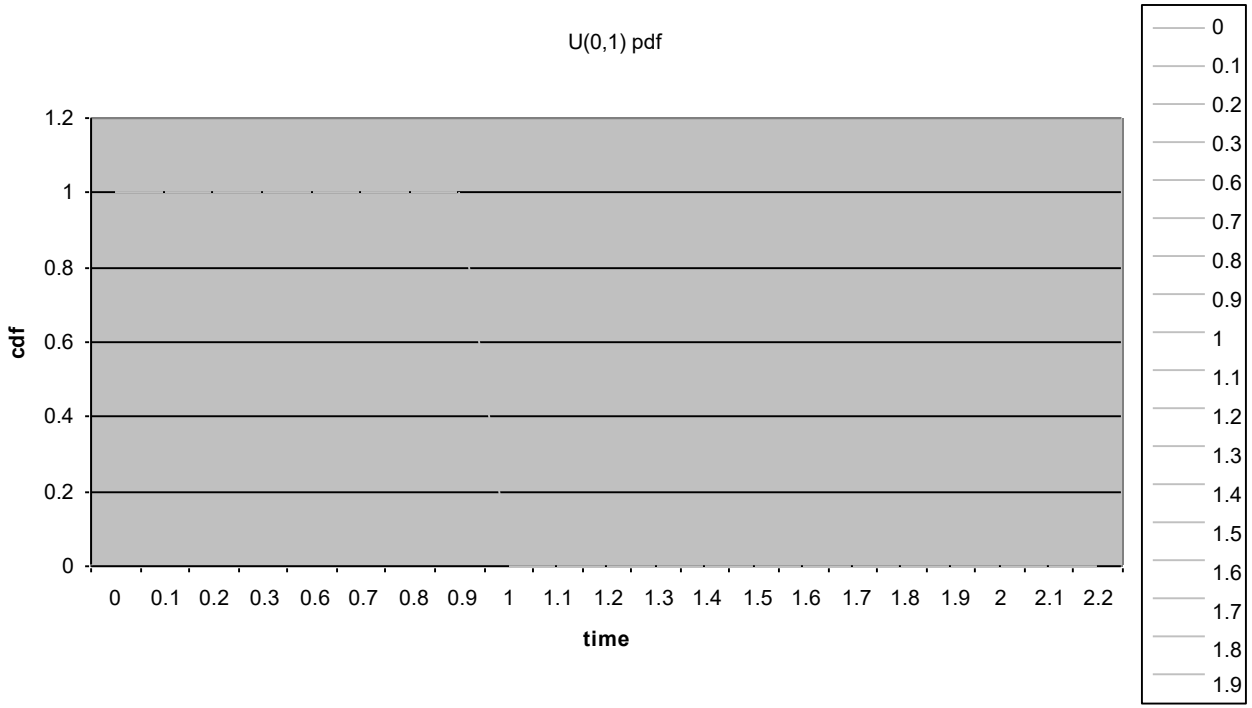
$$\sigma = 1/25 \text{ hours} = 2.4 \text{ minutes}$$

Uniform

All (pseudo) random generators generate random deviates of U(0,1) distribution; that is, if you generate a large number of random variables and plot their empirical distribution function, it will approach this distribution in the limit.

U(a,b) → pdf constant over the (a,b) interval and CDF is the ramp function

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$


U(0,1) pdf
cdf
time
1.2
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2
0
0.1
0.2
0.3
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9

Uniform distribution

$$F(x)=\begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x > b. \end{cases}$$

Gaussian (Normal) Distribution

- Bell shaped pdf – intuitively pleasing!
- Central Limit Theorem: *mean of a large number of mutually independent rv's (having arbitrary distributions) starts following Normal distribution as n* →

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

- μ: mean, σ: std. deviation, σ^2: variance (N(μ, σ^2))
- μ and σ completely describe the statistics. This is significant in statistical estimation/signal processing/communication theory etc.

- *N(0,1)* is called normalized Guassian.
- *N(0,1)* is symmetric i.e.
  - *f(x)=f(-x)*
  - *F(z) = 1-F(z).*
- Failure rate *h(t)* follows IFR behavior.
  - Hence, *N( )* is suitable for modeling long-term wear or aging related failure phenomena

Exponential Distribution

$$f(t) = \sum_{i=1}^{k} \alpha_i \lambda_i \mathrm{e}^{-\lambda_i t}, \ t > 0, \ \lambda_i > 0, \ \alpha_i > 0, \ \sum_{i=1}^{k} \alpha_i$$

$$F(t) = \sum_{i} \alpha_i (1 - \mathrm{e}^{-\lambda_i t}), \ t \geq 0$$

$$h(t) = \frac{\sum_i \alpha_i \lambda \mathrm{e}^{-\lambda_i t}}{\sum_i \alpha_i \lambda \mathrm{e}^{-\lambda_i t}}, \ t \geq 0$$

Conditional Distributions

- The conditional distribution of *Y given X=1* is:
- While marginal distributions are obtained from the bivariate by summing, conditional distributions are obtained by "making a cut" through the bivariate distribution

The Expectation of a Random Variable

Expectation of a discrete random variable with p.m.f

$$P(X = x_i) = p_i$$

$$E(X) = \sum_i p_i x_i$$

Expectation of a continuous random variable with p.d.f *f(x)*

$$E(X) = \int_{\text{state space}} x f(x)\, dx$$

$$\text{expectation of } X = \text{mean of } X = \text{average of } X$$

$$E[X] = \bar{X} = \int_{-\infty}^{\infty} x f_X(x) dx \qquad \text{continuous r.v.}$$

$$E[X] = \bar{X} = \sum_{i=1}^{N} x_i P(x_i) \qquad \text{discrete r.v.}$$

$$f_X(x+a) = f_X(-x+a), \forall x \quad \Rightarrow \quad E[X] = a$$

X r.v. $\Rightarrow$ $Y = g(X)$ r.v. Ex: $Y = g(X) = X^2$

$$P(X=0) = P(X=-1) = P(X=1) = \frac{1}{3} \qquad P(Y=0) = \frac{1}{3} \qquad P(Y=1) = \frac{2}{3}$$

Expectation

expectation of a function of a r.v. X

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \qquad \text{continuous r.v.}$$

$$E[g(X)] = \sum_{i=1}^{N} g(x_i) P(x_i) \qquad \text{discrete r.v.}$$

conditional expectation of a r.v. X

$$E[X|B] = \int_{-\infty}^{\infty} x f_X(x|B) dx \qquad \text{continuous r.v.}$$

$$E[X|B] = \sum_{i=1}^{N} x_i P(x_i|B) \qquad \text{discrete r.v.}$$

4.4 CONDITIONAL PROBABILITY AND CONDITIONAL EXPECTATION

Conditional Probability

In Section 2.4, we know

$$P[Y \text{ in } A \mid X = x] = \frac{P[Y \text{ in } A, X = x]}{P[X = x]}. \qquad (4.22)$$

If X is discrete, then Eq. (4.22) can be used to obtain the conditional cdf of Y given $X = x_k$:

$$F_Y(y \mid x_k) = \frac{P[Y \le y, X = x_k]}{P[X = x_k]}, \quad \text{for } P[X = x_k] > 0. \qquad (4.23)$$

The conditional pdf of Y given $X = x_k$, if the derivative exists, is given by

$$f_Y(y \mid x_k) = \frac{d}{dy} F_Y(y \mid x_k). \qquad (4.24)$$

MULTIPLE RANDOM VARIABLES

Joint Distributions

The joint cumulative distribution function of $X_1, X_2, \ldots, X_n$ is defined as the probability of an n-dimensional semi-infinite rectangle associate with the point $(x_1, \ldots, x_n)$:

$$F_{X_1, X_2, \ldots X_n}(x_1, x_2, \ldots x_n) = P[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n]. \qquad (4.38)$$

The joint cdf is defined for discrete, continuous, and random variables of mixed type

FUNCTIONS OF SEVERAL RANDOM VARIABLES

One Function of Several Random Variables

Let the random variable Z be defined as a function of several random variables:

$$Z = g(X_1, X_2, ..., X_n). \tag{4.51}$$

The cdf of Z is found by first finding the equivalent event of that is, the set $R_Z = \{x = (x_1, ..., x_n)$ such that $g(x) \le z\}$, then

$$F_Z(z) = P[X \text{ in } R_z]$$
$$= \int \cdots \int_{x \text{ in } R_z} f_{X_1,...,X_n}(x_1', ..., x_n') dx_1' ... dx_n'. \tag{4.52}$$

EXAMPLE 4.31 Sum of Two Random Variables

Let $Z = X + Y$. Find $F_Z(z)$ and $f_Z(z)$ in terms of the joint pdf of X and Y.

The cdf of Z is

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x'} f_{X,Y}(x', y')dy'\,dx'.$$

The pdf of Z is

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x',\ z - x')dx' \ . \qquad (4.53)$$

Thus the pdf for the sum of two random variables is given by a superposition integral. If X and Y are independent random variables, then by Eq. (4.21) the pdf is given by the convolution integral of the margial pdf's of X and Y :

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x') f_Y(z - x')dx'. \qquad (4.54)$$

EXPECTED VALUE OF FUNCTIONS OF RANDOM VARIABLES

The expected value of $Z = g(X, Y)$ can be found using the following expressions

$$E[Z] = \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) & X, Y \text{ jointly continuous} \\ \sum_{i}\sum_{n} g(x_i, y_n) p_{X,Y}(x_i, y_n) & X, Y \text{ discrete.} \end{cases} \quad (4.64)$$

*Joint Characteristic Function

The joint characteristic function of n random variables is defined as

$$\Phi_{X_1,X_2,\ldots X_n}(w_1,w_2,\ldots w_n) = E\left[e^{j(w_1X_1+w_2X_2+\cdots+w_nX_n)}\right] \qquad (4.73a)$$

$$\Phi_{X,Y}(w_1,w_2) = E\left[e^{j(w_1X+w_2Y)}\right] \qquad (4.73b)$$

If X and Y are jointly continuous random variables, then

$$\Phi_{X,Y}(w_1,w_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)e^{j(w_1x+w_2y)}dxdy. \qquad (4.73c)$$

The inversion formula for the Fourier transform implies that the joint pdf is given by

$$f_{X,Y}(x,y) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi_{X,Y}(w_1,w_2)e^{j(w_1x+w_2y)}dw_1dw_2\,. \qquad (4.74)$$

JOINTLY GAUSSIAN RANDOM VARIABLES

The random variables X and Y are said to be jointly Gaussian if their joint pdf has the form

$$f_{X,Y}(x,y) = \frac{\exp\left\{\frac{-1}{2(1-\rho_{X,Y}^2)}\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 - 2\rho_{X,Y}\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + \left(\frac{y-m_2}{\sigma_2}\right)^2\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{X,Y}^2}} \tag{4.79}$$

$-\infty < x < \infty$ and $-\infty < y < \infty$

The pdf is constant for values x and y for which the argument of the exponent is constant

$$\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 - 2\rho_{X,Y}\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + \left(\frac{y-m_2}{\sigma_2}\right)^2\right] = \text{constant}$$

When $\rho_{XY} = 0$, X and Y are independent ; when $\rho_{XY} \neq 0$, the major axis of the ellipse is oriented along the angle

$$\theta = \frac{1}{2}\arctan\left(\frac{2\rho_{X,Y}\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\right). \qquad (4.80)$$

Note that the angle is 45º when the variance are equal.
The marginal pdf of X is found by integrating $f_{X,Y}(x, y)$ over all y

$$f_X(x) = \frac{e^{-(x-m_1)^2/2\sigma_1^2}}{\sqrt{2\pi}\sigma_1}, \qquad (4.81)$$

that is, X is a Gaussian random variable with mean m_1 and variance

$$\sigma_1^2$$

n Jointly Gaussian Random Variables

The random variables $X_1, X_2, \ldots, X_n$ are said to be jointly Gaussian if their joint pdf is given by

$$f_{\mathbf{X}}(\mathbf{x}) \equiv f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots x_n) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T K^{-1}(\mathbf{x}-\mathbf{m})\right\}}{(2\pi)^{n/2}|k|^{1/}}, \quad \textbf{(4.83)}$$

where **x** and **m** are column vectors defined by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ E[X_3] \\ E[X_4] \end{bmatrix}$$

and K is the covariance matrix that is defined by

$$K = \begin{bmatrix} \text{VAR}(X_1) & \text{COV}(X_2, X_1) & \cdot & \text{COV}(X_1, X_n) \\ \text{COV}(X_2, X_1) & \text{VAR}(X_2) & \cdot & \text{COV}(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ \text{COV}(X_n, X_1) & \cdot & & \text{VAR}(X_n) \end{bmatrix} \quad (4.84)$$

Transformations of Random Vectors

Let $X_1,\ldots, X_n$ be random variables associate with some experiment, and let the random variables $Z_1,\ldots, Z_n$ be defined by n functions of **X** = $(X_1,\ldots, X_n)$:

$$Z_1 = g_1(\mathrm{X}) \qquad Z_2 = g_2(\mathrm{X}) \quad \ldots \quad Z_n = g_n(\mathrm{X}).$$

The joint cdf of $Z_1,\ldots, Z_n$ at the point **z** = ($z1,\ldots, zn$) is equal to the probability of the region of **x** where

$$F_{Z_1,\ldots,Z_n}(z_1,\ldots, z_n) = P\left[g_1(\mathrm{X}) \le z_1,\ldots, g_n(\mathrm{X}) \le z_n\right]. \qquad (4.55a)$$

$$F_{Z_1,\ldots,Z_n}(z_1,\ldots, z_n) = \int_{\mathrm{x}':g_k(\mathrm{x}')\le z_k} \cdots \int f_{X_1,\ldots,X_n}(x_1',\ldots, x_n')dx_1' \cdots dx_n'. \qquad (4.55b)$$

DEMO

Total Reward	Total Plays
14	24

1

	Arm 1	Arm 2	Arm 3	Arm 4	Arm 5
Rewards:	6	2	2	2	2
Pulls:	8	4	4	4	4
Estimated Probs:	0.750	0.500	0.500	0.500	0.500
UCBs:	1.641	1.761	1.761	1.761	1.761

https://perso.crans.org/besson/phd/MAB_interactive_demo/

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

Multi-Arm Bandits

Sutton and Barto, Chapter 2

The simplest reinforcement learning problem

The Exploration/Exploitation Dilemma

- Online decision-making involves a fundamental choice:
  - Exploitation Make the best decision given current information
  - Exploration Gather more information
- The best long-term strategy may involve short-term sacrifices Gather enough information to make the best overall decisions

Examples

Restaurant Selection

Exploitation Go to your favourite restaurant

Exploration Try a new restaurant

Online Banner Advertisements

Exploitation Show the most successful advert Exploration Show a different advert

Oil Drilling

Exploitation Drill at the best known location

Exploration Drill at a new location

Game Playing

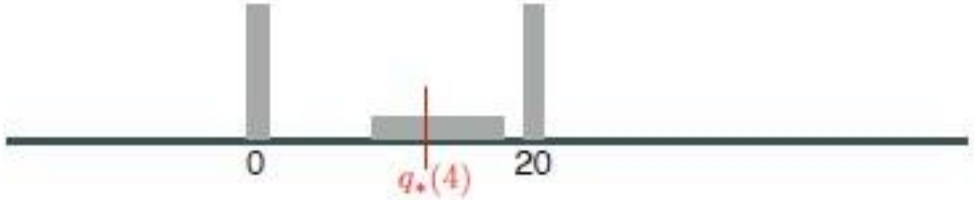
Exploitation Play the move you believe is best Exploration Play an experimental move

You are the algorithm! (bandit1)

- Action 1 — Reward is always 8
  - value of action 1 is $q_*(1) =$
- Action 2 — 88% chance of 0, 12% chance of 100!
  - value of action 2 is $q_*(2) = .88 \times 0 + .12 \times 100 =$
- Action 3 — Randomly between -10 and 35, equiprobable

$q_*(3) =$

- Action 4 — a third 0, a third 20, and a third from {8,9,…, 18}



The k-armed Bandit Problem

- On each of a sequence of *time steps*, t=1, 2, 3, ... ,
  you choose an action A_t from k possibilities, and receive a real-valued *reward* R_t

- The reward depends only on the action taken;
  it is indentically, independently distributed (i.i.d.):

$$q_*(a) \doteq \mathbb{E}[R_t | A_t = a], \quad \forall a \in \{1, \ldots, k\}$$

true values

- These true values are *unknown*. The distribution is unknown

- Nevertheless, you must maximize your total reward

- You must both try actions to learn their values (explore), and prefer those that appear best (exploit)

The Exploration/Exploitation Dilemma

- Suppose you form estimates

$$Q_t(a) \approx q_*(a), \quad \forall a$$ *action-value estimates*

- Define the *greedy action* at time *t* as

$$A_t^* \doteq \arg\max_a Q_t(a)$$

- If $A_t = A_t^*$ then you are *exploiting*
  If $A_t \neq A_t^*$ then you are *exploring*

- You can't do both, but you need to do both

- You can never stop exploring, but maybe you should explore less with time. Or maybe not.

Regret

The *action-value* is the mean reward for action a,

- $q^*(a) = \mathrm{E}[r|a]$

The *optimal value* V^* is

- $V^* = Q(a^*) = \max_{a \in A} q^*(a)$

The *regret* is the opportunity loss for one step

- $l_t = \mathrm{E}[V^* - Q(a_t)]$

The *total regret* is the total opportunity loss

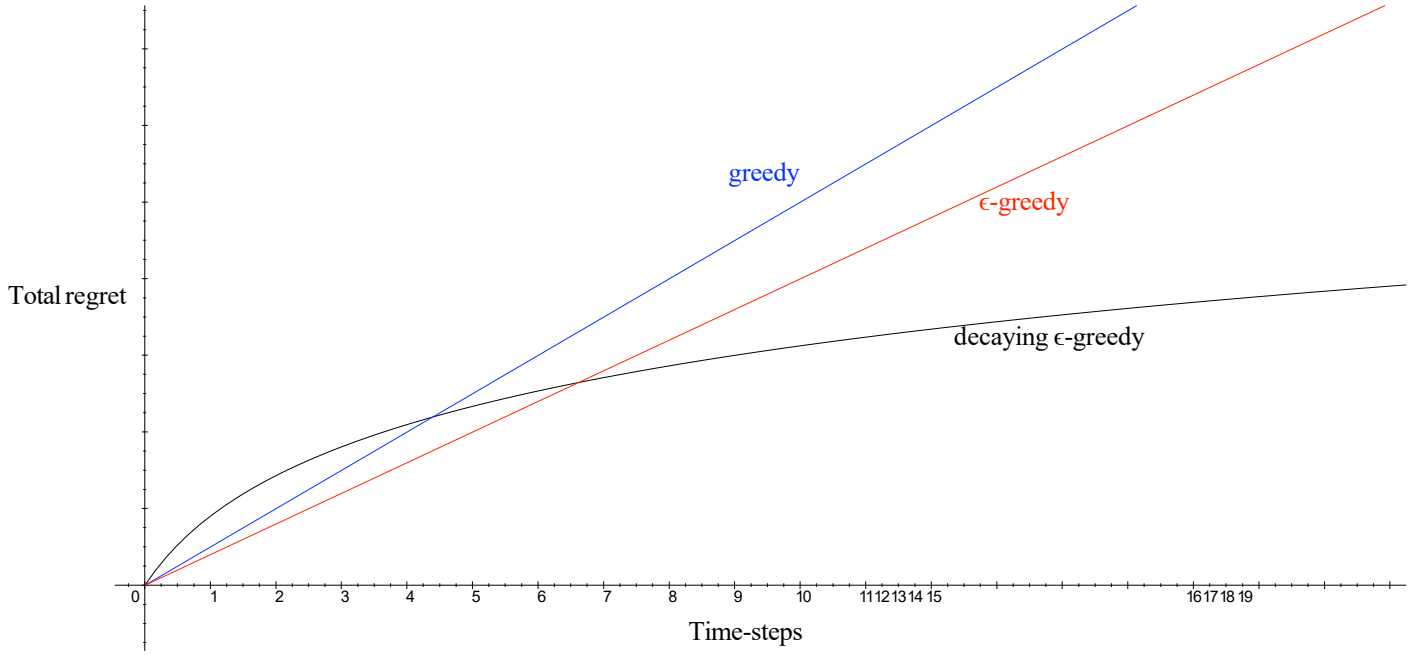
$$L_t = \mathbb{E}\left[\sum_{\tau=1}^{t} V^* - Q(a_\tau)\right]$$

- Maximise cumulative reward $\equiv$ minimise total regret

- The *count* $N_t(a)$ is expected number of selections for action a
- The *gap* Δ_a is the difference in value between action a and optimal action a^*, $\Delta_a = V^* - Q(a)$
- Regret is a function of gaps and the counts

$$\begin{aligned} L_t &= \mathbb{E}\left[\sum_{\tau=1}^{t} V^* - Q(a_\tau)\right] \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}\left[N_t(a)\right](V^* - Q(a)) \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}\left[N_t(a)\right] \Delta_a \end{aligned}$$

- A good algorithm ensures small counts for large gaps
- Problem: gaps are not known!



- If an algorithm forever explores it will have linear total regret
- If an algorithm never explores it will have linear total regret Is it possible to achieve sublinear total regret?

Complexity of regret

- The performance of any algorithm is determined by similarity between optimal arm and other arms
- Hard problems have similar-looking arms with different means
- This is described formally by the gap Δ_a and the similarity in distributions $KL(\mathcal{R}^a||\mathcal{R}^a*)$

Theorem (Lai and Robbins)

Asymptotic total regret is at least logarithmic in number of steps

$$\lim_{t\to\infty} L_t \geq \log t \sum_{a|\Delta_a>0} \frac{\Delta_a}{KL(\mathcal{R}^a||\mathcal{R}^{a^*})}$$

Overview

- Action-value methods
  - Epsilon-greedy strategy
  - Incremental implementation
  - Stationary vs. non-stationary environment
  - Optimistic initial values
- UCB action selection
- Gradient bandit algorithms
- Associative search (contextual bandits)

Basics

- Maximize total reward collected
  - vs learn (optimal) policy (RL)
- Episode is one step
- Complex function of
  - True value
  - Uncertainty
  - Number of time steps
  - Stationary vs non-stationary?

Action-Value Methods

- Methods that learn action-value estimates and nothing else
- For example, estimate action values as *sample averages*:

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbf{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbf{1}_{A_i=a}}$$

- The sample-average estimates converge to the true values
  If the action is taken an infinite number of times

$$\lim_{N_t(a)\to\infty} Q_t(a) = q_*(a)$$

The number of times action a has been taken by time t

ε-Greedy Action Selection

- In greedy action selection, you always exploit
- In ε-greedy, you are usually greedy, but with probability ε you instead pick an action at random (possibly the greedy action again)
- This is perhaps the simplest way to balance exploration and exploitation

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Repeat forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

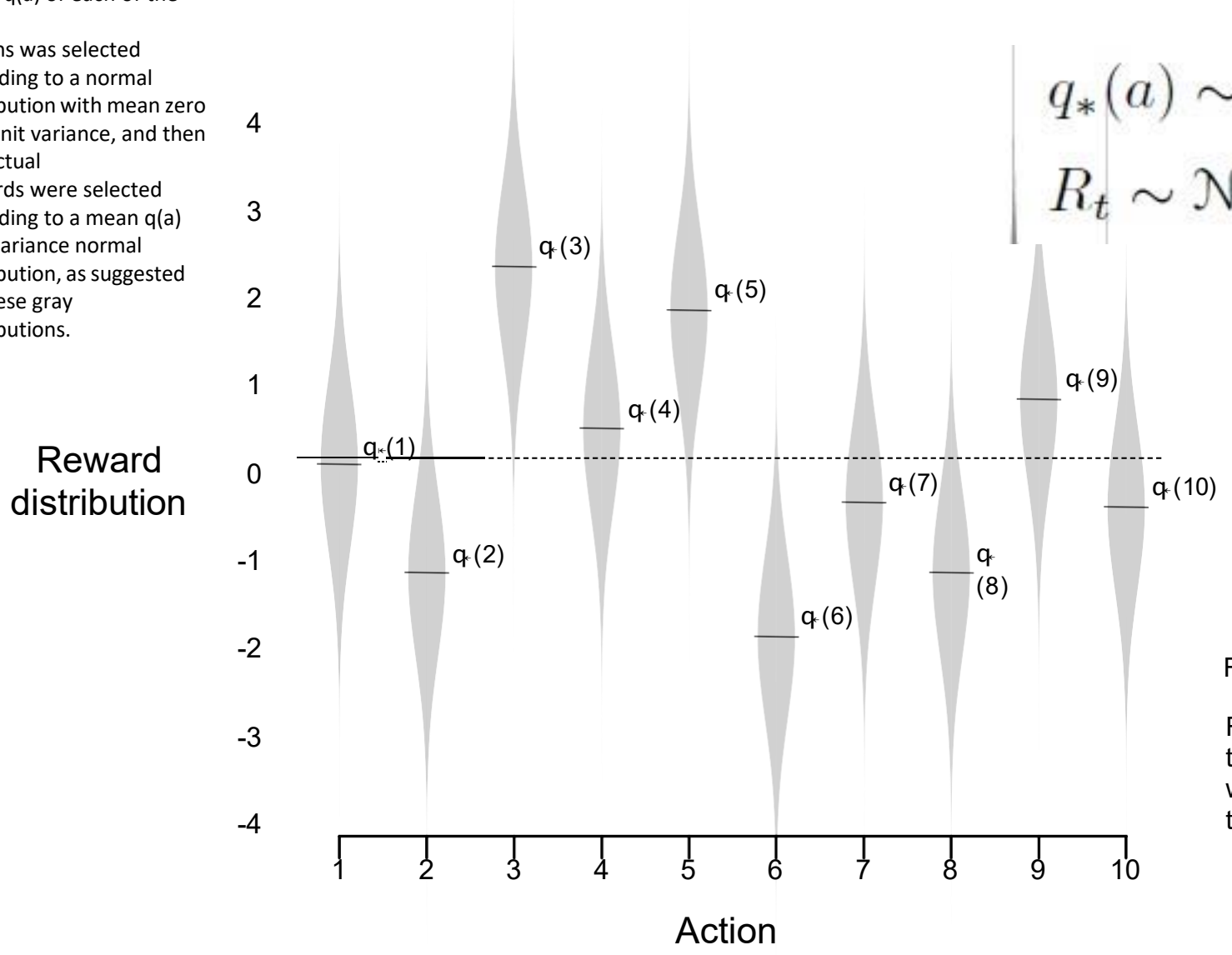
One Bandit Task from

The 10-armed Testbed

Figure 2.1: An example bandit problem from the 10-armed testbed. The true value q(a) of each of the ten
actions was selected according to a normal distribution with mean zero and unit variance, and then the actual
rewards were selected according to a mean q(a) unit variance normal distribution, as suggested by these gray distributions.

$$q_*(a) \sim \mathcal{N}(0, 1)$$

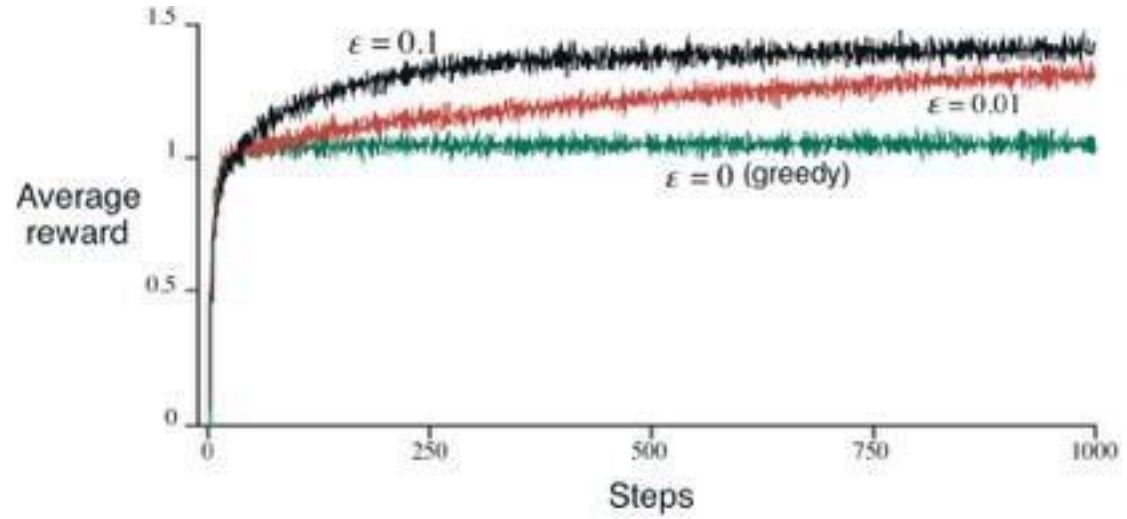
$$R_t \sim \mathcal{N}(q_*(a), 1)$$

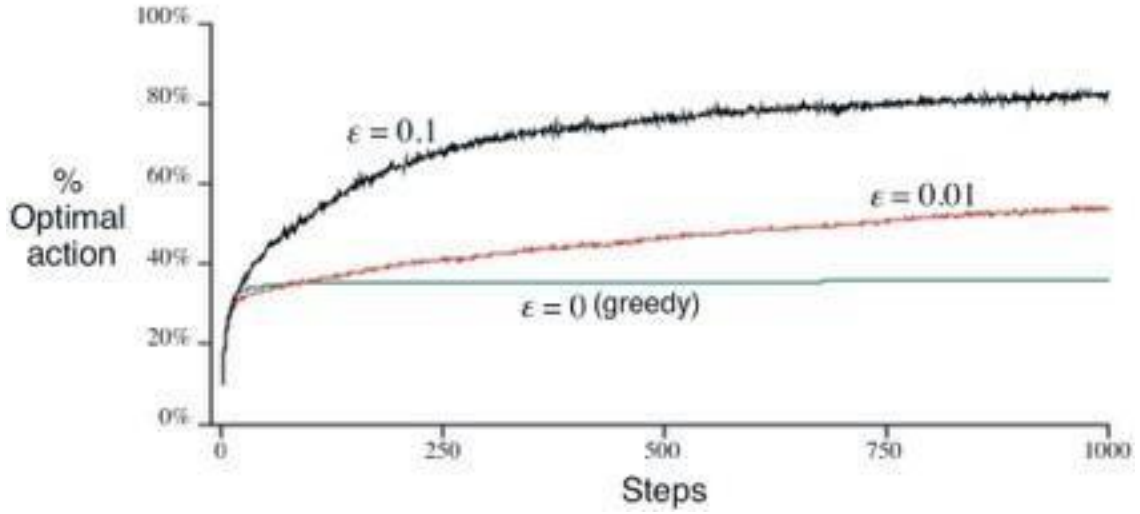


Run for 1000 steps

Repeat the whole thing 2000 times with different bandit tasks

ε-Greedy Methods on the 10-Armed Testbed





Averaging ⟶ learning rule

- To simplify notation, let us focus on one action
  - We consider only its rewards, and its estimate after $n+1$ rewards:

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

- How can we do this incrementally (without storing all the rewards)?
- Could store a running sum and count (and divide), or equivalently:

$$Q_{n+1} = Q_n + \frac{1}{n}\Big[R_n - Q_n\Big]$$

- This is a standard form for learning/update rules:

$$NewEstimate \leftarrow OldEstimate + StepSize\Big[Target - OldEstimate\Big]$$

Derivation of incremental update

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

$$\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big],
\end{aligned}$$

Tracking a Non-stationary Problem

- Suppose the true action values change slowly over time
  - then we say that the problem is *nonstationary*
- In this case, sample averages are not a good idea (Why?)
- Better is an "exponential, recency-weighted average":

$$Q_{n+1} = Q_n + \alpha\Big[R_n - Q_n\Big]$$

$$= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$$

where α is a constant, *step-size parameter*, $0 < \alpha \leq 1$

- There is bias due to Q_1 that becomes smaller over time

Standard stochastic approximation convergence conditions

- To assure convergence with probability 1:

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

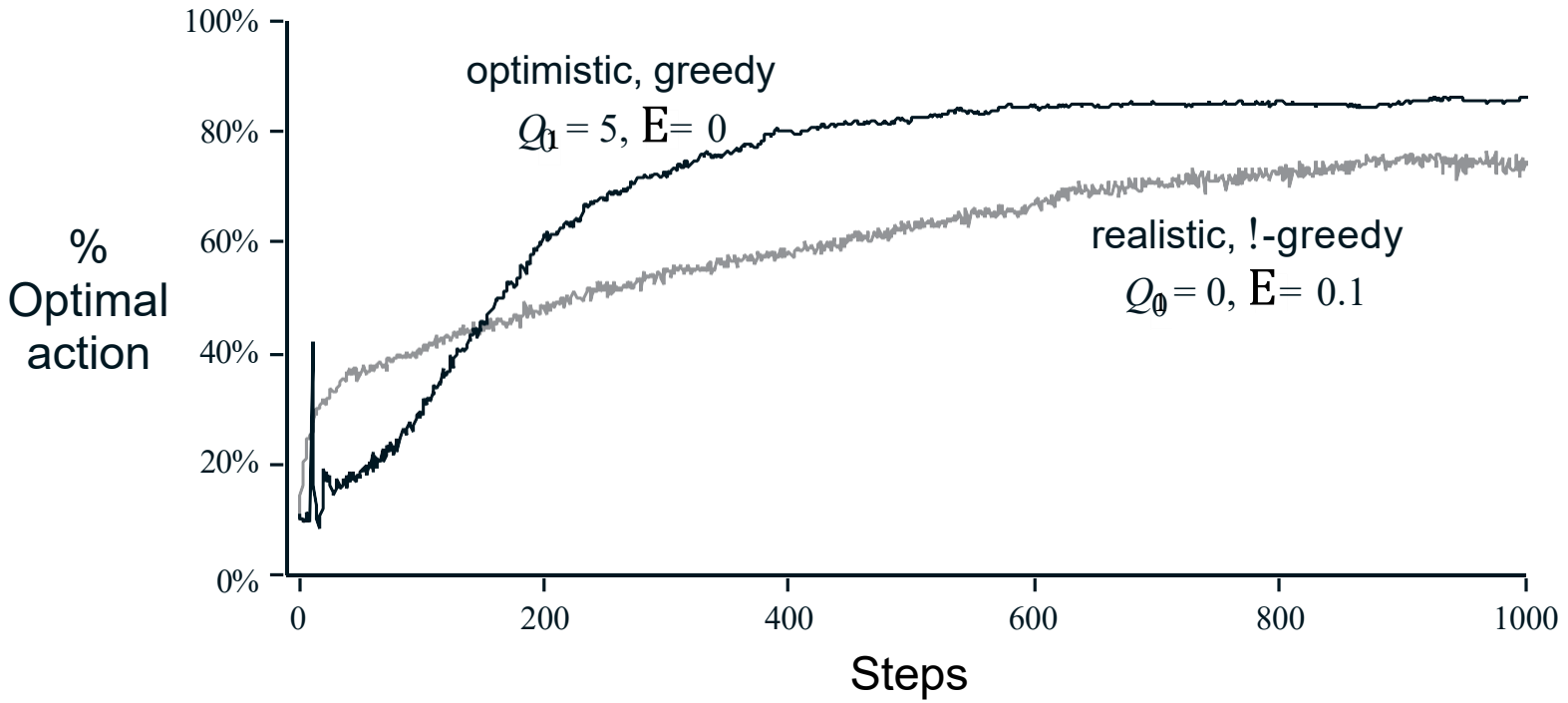
- e.g., $\alpha_n = \frac{1}{n}$
- not $\alpha_n = \frac{1}{n^2}$

if $\alpha_n = n^{-p}, \quad p \in (0,1)$
then convergence is
at the optimal rate:
$O(1/\sqrt{n})$

Optimistic InitialValues

- All methods so far depend on $Q_1(a)$, i.e.,they are biased.
  So far we have used $Q_1(a) = 0$

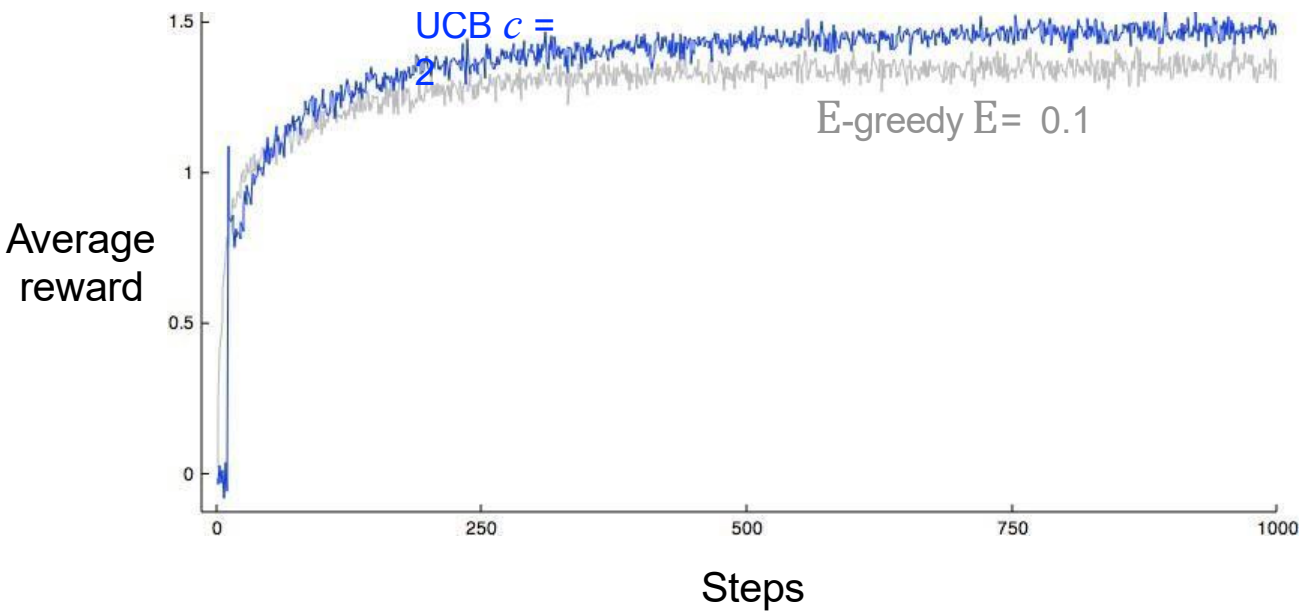
- Suppose we initialize the action values *optimistically* ($Q_1(a) = 5$), e.g., on the 10-armed testbed (with alpha= 0.1)



Upper Confidence Bound (UCB) action selection

- A clever way of reducing exploration over time
- Focus on actions whose estimate has large degree of uncertainty
- Estimate an upper bound on the true action values
- Select

$$A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$$



Complexity of UCB Algorithm

Theorem

The UCB algorithm achieves logarithmic asymptotic total regret

$$\lim_{t \to \infty} L_t \leq 8 \log t \sum_{a | \Delta_a > 0} \Delta_a$$

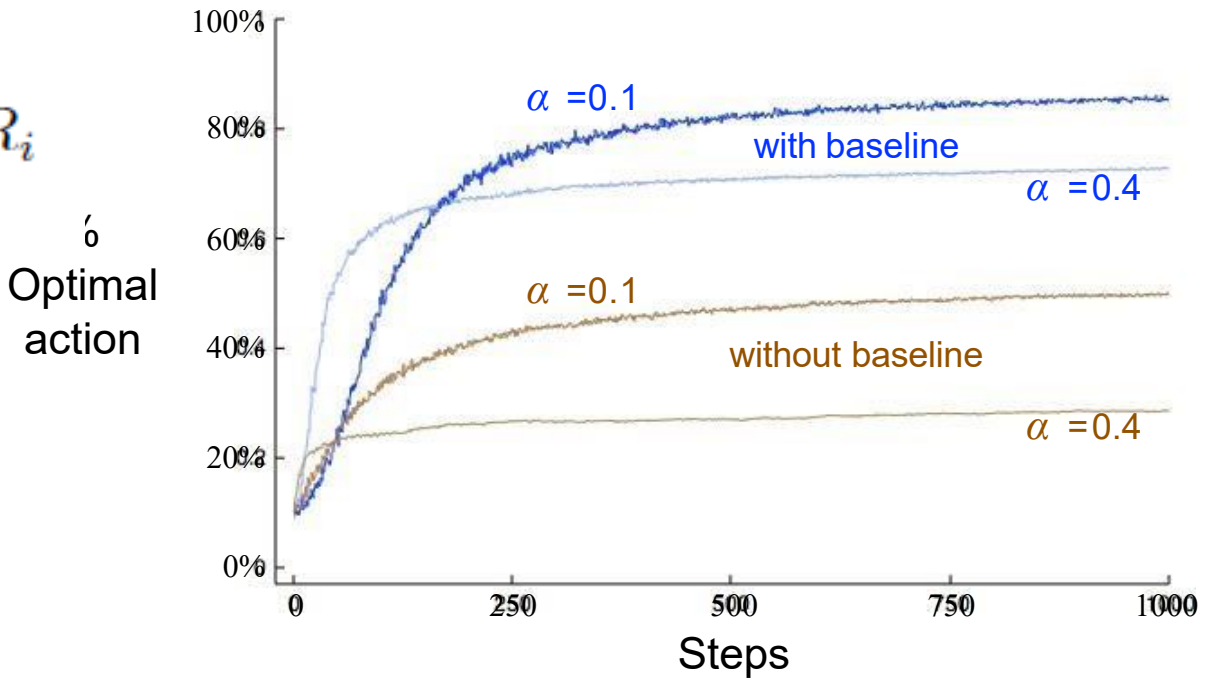
Gradient-Bandit Algorithms

- Let $H_t(a)$ be a learned *preference* for taking action a

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

$$\begin{aligned} H_{t+1}(A_t) &\doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), \quad \text{and} \\ H_{t+1}(a) &\doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \quad \text{for all } a \neq A_t, \end{aligned} \tag{2.10}$$

$$\bar{R}_t \doteq \frac{1}{t}\sum_{i=1}^{t} R_i$$



Derivation of gradient-bandit algorithm

In exact *gradient ascent*:

$$H_{t+1}(a) \doteq H_t(a) + \alpha \frac{\partial \mathbb{E}\left[R_t\right]}{\partial H_t(a)}, \tag{1}$$

where:

$$\mathbb{E}[R_t] \doteq \sum_b \pi_t(b) q_*(b),$$

$$\begin{aligned}
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \left[\sum_b \pi_t(b) q_*(b) \right] \\
&= \sum_b q_*(b) \frac{\partial \pi_t(b)}{\partial H_t(a)} \\
&= \sum_b \left(q_*(b) - X_t\right) \frac{\partial \pi_t(b)}{\partial H_t(a)},
\end{aligned}$$

where X_t does not depend on b, because $\sum_b \frac{\partial \pi_t(b)}{\partial H_t(a)} = 0$.

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_b (q_*(b) - X_t) \frac{\partial \pi_t(b)}{\partial H_t(a)}$$

$$= \sum_b \pi_t(b)(q_*(b) - X_t) \frac{\partial \pi_t(b)}{\partial H_t(a)} / \pi_t(b)$$

$$= \mathbb{E}\left[(q_*(A_t) - X_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right]$$

$$= \mathbb{E}\left[(R_t - \bar{R}_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right],$$

where here we have chosen $X_t = \bar{R}_t$ and substituted R_t for $q_*(A_t)$, which is permitted because $\mathbb{E}[R_t|A_t] = q_*(A_t)$.
For now assume: $\frac{\partial \pi_t(b)}{\partial H_t(a)} = \pi_t(b)(\mathbf{1}_{a=b} - \pi_t(a))$. Then:

$$= \mathbb{E}\left[(R_t - \bar{R}_t) \pi_t(A_t)(\mathbf{1}_{a=A_t} - \pi_t(a)) / \pi_t(A_t) \right]$$
$$= \mathbb{E}\left[(R_t - \bar{R}_t)(\mathbf{1}_{a=A_t} - \pi_t(a)) \right].$$

$$H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t)(\mathbf{1}_{a=A_t} - \pi_t(a)), \text{ (from (1), QED)}$$

Thus it remains only to show that

$$\frac{\partial \pi_t(b)}{\partial H_t(a)} = \pi_t(b)\big(\mathbf{1}_{a=b} - \pi_t(a)\big).$$

Recall the standard quotient rule for derivatives:

$$\frac{\partial}{\partial x}\left[\frac{f(x)}{g(x)}\right] = \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x)\frac{\partial g(x)}{\partial x}}{g(x)^2}.$$

Using this, we can write...

Quotient Rule: $\frac{\partial}{\partial x}\left[\frac{f(x)}{g(x)}\right] = \frac{\frac{\partial f(x)}{\partial x}g(x) - f(x)\frac{\partial g(x)}{\partial x}}{g(x)^2}$

$$\frac{\partial \pi_t(b)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)}\pi_t(b)$$

$$= \frac{\partial}{\partial H_t(a)}\left[\frac{e^{h_t(b)}}{\sum_{c=1}^{k} e^{h_t(c)}}\right]$$

$$= \frac{\frac{\partial e^{h_t(b)}}{\partial H_t(a)}\sum_{c=1}^{k} e^{h_t(c)} - e^{h_t(b)}\frac{\partial \sum_{c=1}^{k} e^{h_t(c)}}{\partial H_t(a)}}{\left(\sum_{c=1}^{k} e^{h_t(c)}\right)^2} \qquad \text{(Q.R.)}$$

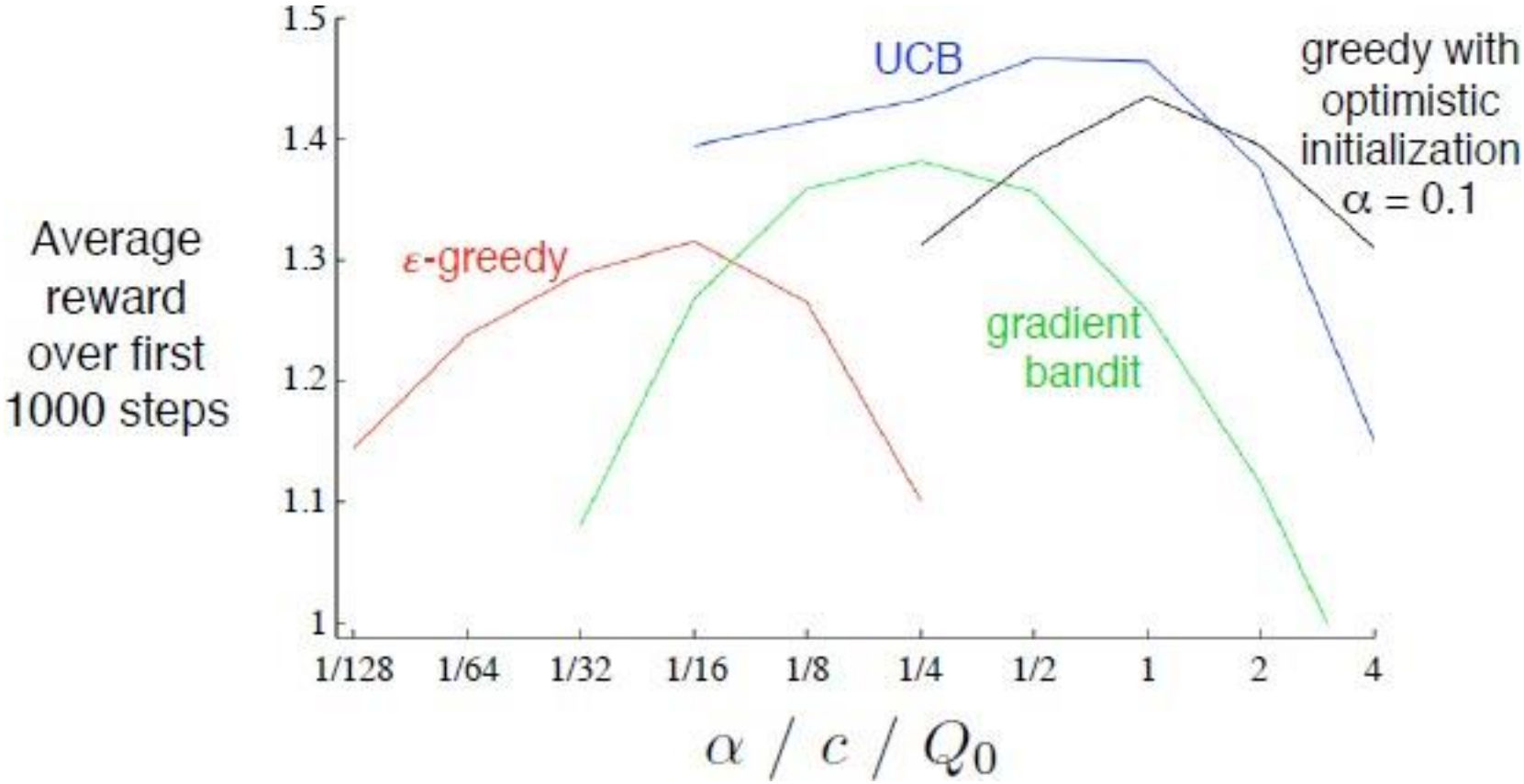
$$= \frac{\mathbf{1}_{a=b}e^{h_t(a)}\sum_{c=1}^{k} e^{h_t(c)} - e^{h_t(b)}e^{h_t(a)}}{\left(\sum_{c=1}^{k} e^{h_t(c)}\right)^2} \qquad \left(\frac{\partial e^x}{\partial x} = e^x\right)$$

$$= \frac{\mathbf{1}_{a=b}e^{h_t(b)}}{\sum_{c=1}^{k} e^{h_t(c)}} - \frac{e^{h_t(b)}e^{h_t(a)}}{\left(\sum_{c=1}^{k} e^{h_t(c)}\right)^2}$$

$$= \mathbf{1}_{a=b}\pi_t(b) - \pi_t(b)\pi_t(a)$$

$$= \pi_t(b)\left(\mathbf{1}_{a=b} - \pi_t(a)\right). \qquad \text{(Q.E.D.)}$$

Summary Comparison of Bandit Algorithms



Reference:

1. Richard S. Sutton and Andrew G. Batto, "Reinforcement leatning: An introduction", Second Edition., MIT Press, 2019
2. Probability, Statistics, and Random Processes for Electrical Engineeling, 3rd Edition, Albe1to Leon-Garcia, 2009