CHAPTER 2: MULTI-ARMED BANDITS

2.1 K-ARMED BANDIT PROBLEM

WHY STUDY IT: SIMPLE NON-ASSOCIATIVE SETTING, MEANING THAT THE OPTIMAL ACTION DOES NOT DEPEND ON A STATE. SO IT IS A GOOD START, AND USEFUL IN PRACTISE

MODEL: K ACTIONS, WHICH WE TAKE REPEATEDLY

$A_1, A_2, A_3, \ldots$ LEADING TO REWARDS

$R_1, R_2, R_3, \ldots$ DRAWN BY SOME UNKNOWN DISTRIBUTION

(NOTE: NO STATE IN THIS CHAPTER)

VALUE OF ACTION a:

$$q_*(a) \stackrel{\Delta}{=} E\left[R_t \mid A_t = a \right] \quad \text{UNKNOWN!}$$

WE ONLY HAVE ESTIMATES OF THE REWARDS AT TIME t:

$$Q_t(a)$$

TWO STRATEGIES:

1) EXPLOITATION: AT TIME $t+1$, PICK ACTION a THAT MAXIMIZES ESTIMATE $Q_t(a)$

THIS IS GREEDY, AND SHORT-TERM OPTIMUM

2) EXPLORATION: PICK SUBOPTIMAL ACTION, WITH THE AIM TO IMPROVE ESTIMATE $Q_t(a)$

LONG TERM, THIS IS ADVANTAGEOUS

$\Longrightarrow$ WE NEED TO STRIKE BALANCE BETWEEN THE TWO

EXAMPLES OF APPLICATIONS:

1. CHOOSING THE ONE-ARMED BANDIT TO PLAY IN CASINOS (THOSE CLOSER TO THE EXITS PERFORM BETTER)
2. WHICH ROULETTE AND WHICH NUMBER TO CHOOSE IN A CASINO (THIS IS IMPORTANT)
3. WHICH DRUG TO USE, FOR A PATIENT
4. WHICH RESTAURANT/BEACH TO VISIT

AIM OF THIS CHAPTER:
DEVELOP GOOD BALANCING STRATEGIES

2.2 ACTION VALUE METHODS

DEFINITION: ACTION VALUE METHODS
1. ESTIMATE ACTION VALUES
2. USE ESTIMATES TO SELECT ACTION

DEFINITION: SAMPLE AVERAGE METHOD USES THE SAMPLE AVERAGE

$$Q_t(a) \triangleq \frac{\text{SUM OF REWARDS WHEN } a \text{ IS TAKEN PRIOR TO } t}{\text{NUMBER OF TIMES } a \text{ TAKEN PRIOR TO } t}$$

$$= \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

NOTE: $\mathbb{1}_{A_i=a}$ AND $Q_t(a)$ ARE RVs

IF DENOMINATOR = 0, SET $Q_t(a)$ = DEFAULT, E.G. 0

INDICATOR FUNCTION

TWO METHODS:

GREEDY METHOD:

$$A_t = \arg\max_a Q_t(a)$$

ε-GREEDY METHOD:

$$A_t = \begin{cases} \arg\max_a Q_t(a), & \text{W.P. } 1-\epsilon \\ \text{RANDOMLY CHOSEN ACTION } a \text{ (UNIFORMLY)} & \end{cases}$$

PROPERTIES: 1) AS $t \to \infty$, ALL ACTIONS WILL BE SELECTED A NUMBER OF TIMES $\to \infty$, SO BY SLLN, WE FIND $q_*(a)$ FOR ALL a.

2) IN THE LONG RUN, OPTIMAL STRATEGY IS SELECTED FOR A PERCENTAGE OF ACTIONS GREATER THAN $1-\epsilon$.

2.3 10-ARMED TESTBED

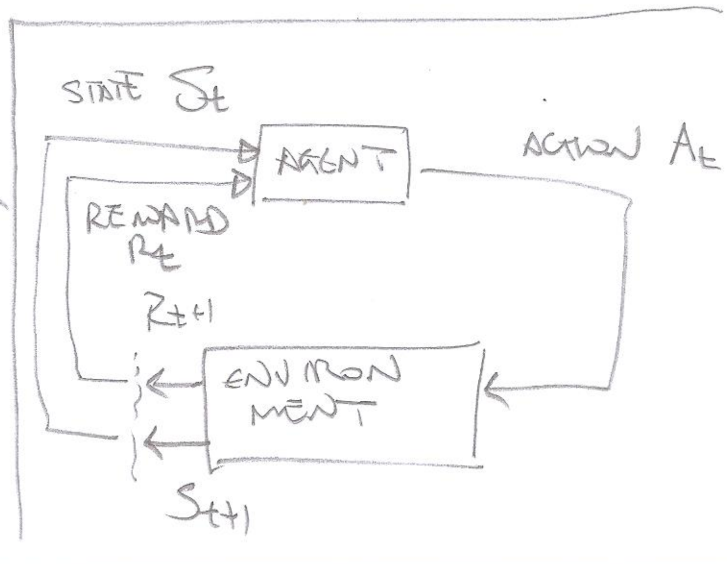
SIMULATION SETTING: 1) 10 ACTIONS, SAMPLED FROM $N(0,1)$ DISTRIBUTION

2) INITIAL ESTIMATES = 0

3) 1000 TIME STEPS

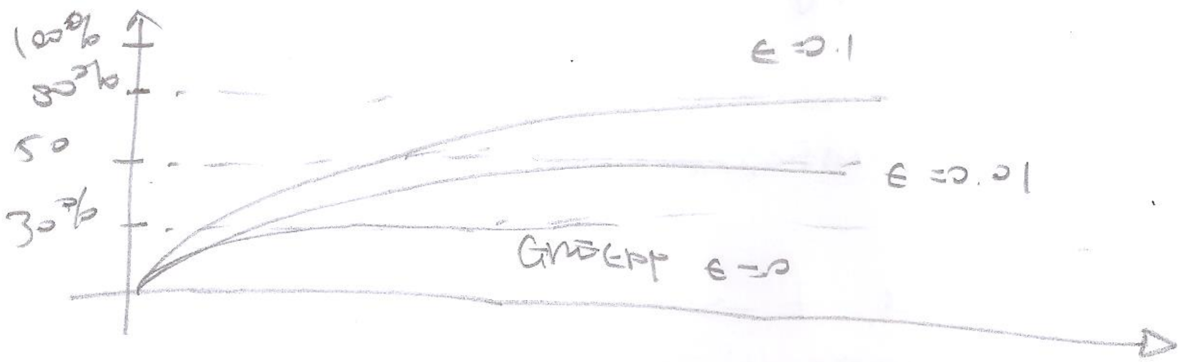
4) 2000 INDEPENDENT RUNS

5) $\epsilon = 0,\ 0.1,\ 0.01$



AVERAGE REWARD

OPTIMAL ACTION



2.4 INCREMENTAL IMPLEMENTATION

THE ACTION VALUE ESTIMATES ARE COMPUTED USING

$$Q_n = \frac{R_1 + R_2 + \dots R_{n-1}}{n-1}$$

(OBSERVE THAT WE CHANGED THE NOTATION)

PROBLEM: WE NEED TO STORE A LOT OF INFORMATION THAT INCREASES WITH n: $R_1, R_2, \dots, R_{n-1}$

SOLUTION:

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^{n} R_i = \frac{1}{n} \left(\sum_{i=1}^{n-1} R_i + R_n \right) =$$

$$\frac{1}{n} \left((n-1) \frac{\sum_{i=1}^{n-1} R_i}{n-1} + R_n \right) = \frac{1}{n} \left((n-1) Q_n + R_n \right)$$

$$= Q_n + \frac{1}{n} \left[R_n - Q_n \right]$$

$$\Rightarrow \boxed{Q_{n+1} = Q_n + \frac{1}{n}\left[R_n - Q_n\right]} \quad (2.3)$$

THIS FORMULA HAS A VERY GENERAL FORM

NEW ESTIMATE = OLD ESTIMATE + STEPSIZE · [NEW INFO − OLD ESTIMATE]

(IT IS ALSO SIMILAR TO STANDARD <u>SGD</u> ITERATION)

2.5 TRACKING A NONSTATIONARY PROBLEM

WE CAN USE (2.3) BUT WITH OTHER CHOICES OF THE COEFFICIENT.

FOR EXAMPLE, LET $\alpha \in (0, 1]$ THEN

$$\begin{aligned}
Q_{n+1} &\stackrel{\Delta}{=} Q_n + \alpha\left[R_n - Q_n\right] \\
&= \alpha R_n + (1-\alpha) Q_n \\
&= \alpha R_n + (1-\alpha)\left[\alpha R_{n-1} + (1-\alpha) Q_{n-1}\right] \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\
&= \ldots \\
&= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i
\end{aligned}$$

BIAS

THIS IS A MUCH MORE REASONABLE CHOICE FOR NON-STATIONARY PROBLEMS, WHERE YOU WANT TO PLACE MORE EMPHASIS ON MORE RECENT REWARDS

MORE GENERALLY: WE CAN USE ANY SEQUENCE $\{\alpha_n(a)\}$ PROVIDED

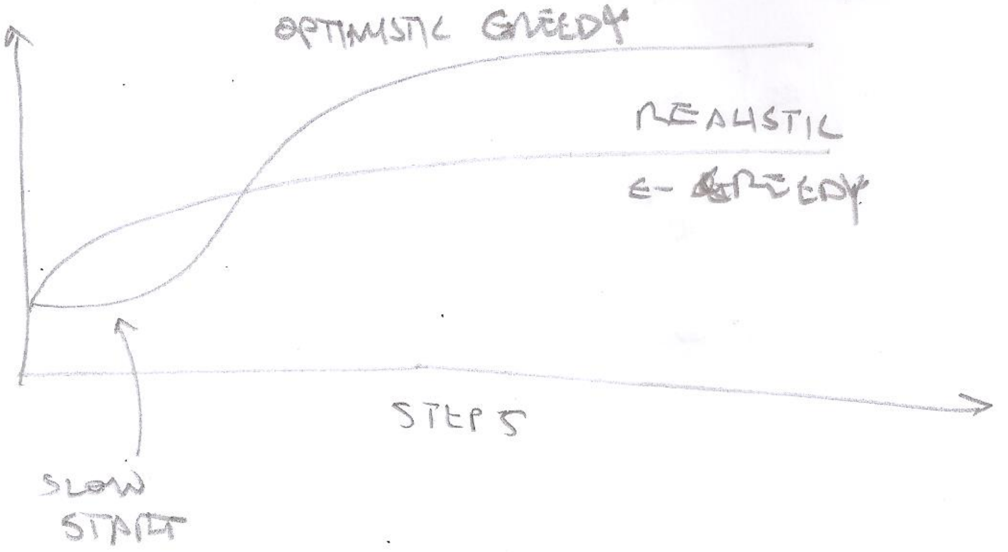
$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \quad , \quad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

STEPS LARGE ENOUGH SO THAT WE CAN EXPLORE

STEPS EVENTUALLY SMALL ENOUGH TO ENSURE CONVERGENCE

2.6 OPTIMISTIC INITIAL VALUES

SIMPLE IDEA: INSTEAD OF SETTING $Q_1 = 0$, SET THEM Q_1 = LARGE. THIS FORCES EXPLORATION



2.7 UPPER-CONFIDENCE-BOUND ACTION SELECTION

WITH ε-GREEDY APPROACHES, WHEN WE DO NOT CHOOSE THE BEST ACTION, WE SELECT ONE OTHER ACTION, WITHOUT PREFERENCE. THIS IS SUBOPTIMAL, BECAUSE SOME ARE REALLY BAD

BETTER IDEA:

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

IF $N_t(a) = 0$, THEN a IS ALWAYS OPTIMAL

SO ALL ACTIONS WILL BE SAMPLED INFINITE TIMES, BUT SUBLINEARLY. AGGRESSIVENESS OF SAMPLING DEPENDS ON PARAMETER c

SEE PAGE 20

2.8 GRADIENT BANDIT ALGORITHMS

WE DO NOT HAVE TO USE THE AVERAGES OF REWARDS AS ESTIMATES. ALTERNATIVE METHODS EXIST. FOR EXAMPLE: USE VECTOR OF PREFERENCES: $H_t(a)$ AS FOLLOWS

1) SET $H_1(a) = 0, \quad \forall a \in \mathcal{A}$
2) AT TIME t SELECT ACTION ACCORDING TO DISTRIBUTION

$$\pi_t(a) \triangleq \frac{e^{H_t(a)}}{\sum_{b \in \mathcal{A}} e^{H_t(b)}}$$

3) AFTER SELECTING A_t AT TIME t,

$$H_{t+1}(A_t) = H_t(A_t) + \alpha \left(R_t - \bar{R}_t\right)\left(1 - \pi_t(A_t)\right)$$

$$H_{t+1}(a) = H_t(a) - \alpha \left(R_t - \bar{R}_t\right) \pi_t(a) \quad \forall a \neq A_t$$

WHERE α LEARNING RATE AND

$\bar{R}_t$ AVERAGE OF REWARDS UP TO TIME t (BUT NOT INCLUDING t)

INTUITIVELY, THIS MAKES SENSE. THERE IS ALSO MATHEMATICAL JUSTIFICATION, AS FOLLOWS

OPTIMIZATION THEORY FRAMEWORK:

WE WANT TO MAXIMIZE

$$g(H) = \sum_{x=1}^{K} \pi(x)\, q_*(x),$$

$$\pi(x) = \frac{e^{H_x}}{\sum_{b=1}^{K} e^{H_b}}$$

WHERE $x = 1, \dots, K$ CHOSEN ACTION, $H = (H_1, H_2, \dots, H_K) \in \mathbb{R}^K$ PREFERENCES OF ACTIONS.

OPTIMAL IS OBVIOUS: FIND ANY $x_0 = \arg\max q_*(x)$ AND SET

$$H_x = \begin{cases} \infty, & x = x_0 \\ 0, & x \neq x_0 \end{cases} \Rightarrow \pi_x = \begin{cases} 1, & x = x_0 \\ 0, & x \neq x_0 \end{cases}$$

LET US FIND $\nabla g(H)$

$i = 1, \dots, k$

$$\frac{\partial g(H)}{\partial H_i} = \frac{\partial}{\partial H_i} \left(\sum_{x=1}^{k} \pi(x)\, q_*(x) \right)$$

$$= \frac{\partial}{\partial H_i} \left(\sum_{x=1}^{k} \pi(x) \left(q_*(x) - B \right) \right)$$

B ARBITRARY CONSTANT THAT WE WILL USE LATER. NOTE THAT $\sum_{x=1}^{k} \pi(x) = 1$

$$= \sum_{x=1}^{k} \left(q_*(x) - B \right) \frac{\partial}{\partial H_i} \pi(x) =$$

$$\sum_{x=1}^{k} \left(q_*(x) - B \right) \frac{\left(1_{i=x}\, e^{H_x} \right) \left(\sum_{b=1}^{k} e^{H_b} \right) - e^{H_x} \cdot e^{H_i}}{\left(\sum_{b=1}^{k} e^{H_b} \right)^2}$$

$$= \sum_{x=1}^{k} \left(q_*(x) - B \right) \left[1_{i=x}\, \pi(x) - \pi(x)\, \pi(i) \right] \Rightarrow$$

$$\boxed{\frac{\partial g(H)}{\partial H_i} = \sum_{x=1}^{k} \left(q_*(x) - B \right) \pi(x) \left[1_{i=x} - \pi(i) \right]}$$

FOR ANY $B \in \mathbb{R}$ A

OBSERVE THAT

$$H_{t+1}(i) = H_t(i) + \underbrace{\left(R_t - \bar{R}_t \right) \left[1_{i=A_t} - \pi(i) \right]}_{\text{RANDOM}}$$

RANDOM BUT WHAT IS THE MEAN?

(i IS PARAMETER)

$$E_{R_t, A_t}\left[(R_t-\bar{R}_t)(\mathbb{1}_{i=A_t}-\pi(i))\right] =$$ (NESTED EXPECTATION)

$$E_{A_t}\left[E_{R_t}\left[(R_t-\bar{R}_t)(\mathbb{1}_{i=A_t}-\pi(i)) \mid A_t=h\right]\right]$$

DOES NOT DEPEND ON R_t

$$= E_{A_t}\left[(\mathbb{1}_{i=A_t}\ \pi(i))\ E_{R_t}\left[(R_t-\bar{R}_t) / A_t=h\right]\right]$$

$$= E_{A_t}\left[(\mathbb{1}_{i=A_t}-\pi(i))(q_*(h)-\bar{R}_t)\right] =$$

$$\sum_{x=1}^{k} \pi(x)(\mathbb{1}_{i=x}-\pi(i))(q_*(h)-\bar{R}_t) \overset{A}{=} \frac{\partial g(h)}{\partial H_i}$$

THEREFORE, THIS IS A STOCHASTIC GRADIENT ASCENT METHOD!

B COULD BY CHOSEN DIFFERENTLY, BUT <u>VARIANCE</u> OF GRADIENT IS KEPT LOW

$\boxed{\text{HW \#1 : REPRODUCE FIGURE 2.6}}$