① バンディット問題

bandit：スロットマシン
レバーを引くとランダムに絵柄が変わる
マークの並びによってコインがもらえる（0〜n）枚

9.1 腕バンディット問題

設定

1本レバースロットマシンが複数台ある
スロットマシン毎に絵柄の出方が異なる.
プレイヤーはスロットマシンの情報を何も知らない
(例えば)1000回プレイして得るコインの枚数を最大化したい

用語

環境 Environment	スロットマシン
エージェント Agent	プレイヤー
行動 Action	1台選んでプレイする
報酬 Reward	スロットマシンから出るコイン

スロットマシンの良さとは？

スロットマシン毎にもらえるコインの確率分布が異なる.

SMa

coin	0	1	5	10
p	0.7	0.15	0.12	0.03

SMb

coin	0	1	5	10
p	0.5	0.4	0.09	0.01

常に期待値の高いマシンを選ぶようプレイすればよい

価値 Value　報酬の期待値
行動価値 Action Value　行動に対して得られる報酬の期待値

$R_t \in \{0, 1, 5, 10\}$　：t回目に得られる報酬 Reward
$A_t \in \{a, b\}$　：t回目の行動 Action，a, bはスロットの種類

$\mathbb{E}[R]$　　Rの期待値

$\mathbb{E}[R|A]$　　Aという行動をした場合のRの期待値

$\mathbb{E}[R|A=a] = \mathbb{E}[R|a]$

$q(A) = \mathbb{E}[R|A]$　行動価値 (Quality)

行動Aの価値

$q(A)$　　真の値　理論値 ← エージェントは知らない

$Q(A)$　　推定値

アルゴリズム

エージェントはスロットマシンの価値（確率分布か期待値）を知らない

→ 各スロットマシンの価値を推定する必要がある.

推定方法

SM	1回目	SMの価値
a	0	0
b	1	1

SM	1回目	2回目	3回目	SMの価値
a	0	1	5	2
b	1	0	0	0.33

$Q(A=a) = 2$, $Q(A=b) = 0.33$

この期待値 $Q(A)$ は推定値だけど SM$_a$ の方が価値が高そう

標本平均

大数の法則：無限回のサンプリングで標本平均は真の値に一致する

実装

$R_1, R_2, \cdots, R_n$ から Q_n を求める.

単に平均すると $n+1$ 回目の行動で R_{n+1} を得たとき $R_1 \sim R_{n+1}$ を再度 必要としてしまう.

→ 漸化的に求める

$$Q_{n-1} = \frac{R_1 + \cdots + R_{n-1}}{n-1}$$

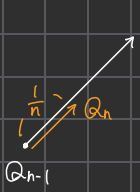
$$R_1 + \cdots + R_{n-1} = (n-1) Q_{n-1}$$

$$\begin{aligned} Q_n &= \frac{1}{n}(R_1 + \cdots + R_{n-1} + R_n) \\ &= \frac{1}{n}\left((n-1)Q_{n-1} + R_n\right) \\ &= \left(1 - \frac{1}{n}\right) Q_{n-1} + \frac{1}{n} R_n \\ &= Q_{n-1} + \frac{1}{n}(R_n - Q_{n-1}) \end{aligned}$$

$$Q_n = Q_{n-1} + \frac{1}{n}(R_n - Q_{n-1})$$

次元も時間も小さい

$$Q_n = Q_{n-1} + \frac{1}{n}(R_n - Q_{n-1})$$



$\frac{1}{n}$：学習率としての役割

プレイヤーの戦略

- greedy（貪欲な）　各SMの価値の推定値が最大のものを常に選ぶ
  実験が少なく推定値の大小と真の値の大小が一致していると決めきれないため「不確かさ」
- SMの価値を精度よく推定するため　様々なSMを試す．

活用 Exploitation　経験から最善な行動をする（greedy）
　← 真の最善を見逃してるかも
探索 Exploration　greedyでない行動を試す．
　↔ トレードオフ

教化学習：活用と探索のバランスをいかに取るか

- ε-greedy法
  ε = 0.1（例）の確率で探索（ランダムに行動を選ぶ），1−εでそれ以外で活用

定常問題：報酬の確率分布が定常
スロットマシンの勝率は定常だった.
非定常問題：毎プレイで勝率が変化するなど

Q_n：価値，R_n：報酬

$$Q_n = \frac{R_1 + \cdots + R_n}{n} = \frac{1}{n}R_1 + \frac{1}{n}R_2 + \cdots + \frac{1}{n}R_n$$

（$\frac{1}{n}$：重み）

定常　$Q_n = Q_{n-1} + \frac{1}{n}(R_n - Q_{n-1})$

非定常　$Q_n = Q_{n-1} + \alpha(R_n - Q_{n-1})$

$$Q_n = \alpha R_n + \alpha(1-\alpha)R_{n-1} + \alpha(1-\alpha)^2 R_{n-2} + \cdots + \alpha(1-\alpha)^{n-1}R_1 + (1-\alpha)^n Q_n$$

$R_1 \sim R_n$ の指数（加重）移動平均