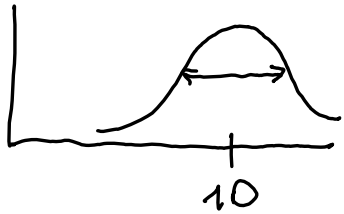
M.A.B

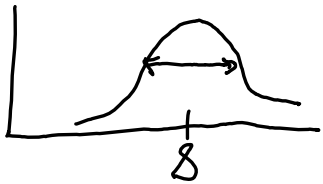
Restaurant example

STANDARD DEV



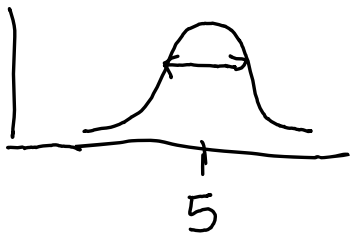
$S_X = 5$

$\bar{X} = 10$



$SX = 4$

$\bar{X} = 8$



$SX = 2.5$

$\bar{X} = 5$

QUESTE INFO SONO NASCOSTE

X è il REWARD DEI RISTORANTI IN termini di SODDISFAZIONE

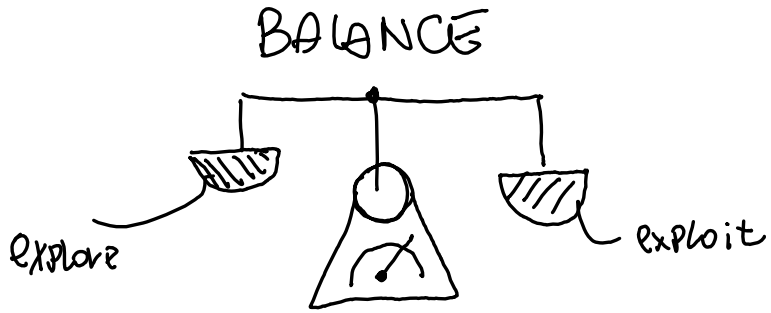
NOI ANDIAMO IN QUESTI RISTORANTI, SENZA SAPERE LA LORO CAPACITÀ DI REWARD

(1) EXPLORATION

- Andiamo prima e cosa e poi secondo una logica

(2) EPLOITATION

- UNA volta che capiamo che quel ristorante che abbiamo esplorato più volte ci soddisfa ci CONTINUIAMO AD ANDARE per massimizzare il reward.



'N BASE ALLA SCELTA DI RISOLUZIONE DEL PROBLEMA AVRETTO UN DISCOSTAMENTO DALLA SOLUZIONE IDEALE (S.I)

" S.I = SAPENDO già le INFO exploito solo sul migliore. "

QUESTA metrica di errore la chiamiamo con $P(R\hat{O})$

ε - Greedy STRATEGY

- Se ESPLORIAMO E BASTA, AD UN CERTO PUNTO SAPPIAMO QUALE BANDIT CI DA UN REWARD MIGLIORE, QUINDI PERDIAMO TEMPO CONTINUANDO AD ESPLORARE
- SE EXPLOITIAMO SU UNO NON PRENDIAMO ABBASTANZA DATI PER SAPERE QUALE potrebbe essere il migliore

CI SERVE COSTRUIRE UNA STRATEGIA DI BILANCIAMENTO.

- PRENDIAMO UNA ε SOLITAMENTE BASSA

$$5\% - 10\% - \frac{1}{\#\,TRY}\,\%$$

ε è la % PER CUI SCEGLIAMO RANDOM UN RISTORANTE

$1-\varepsilon$ è la PROBABILITÀ di CONTINUARE l'EXPLOITATION sul RISTORANTE migliore.

ES: 300 TRY

- 30 EXPLORE
- 270 EXPLOIT

} ε 10%

UCB1 Algo / method

Still:

300 days

3 restaurant → happiness reward distributions

the Problem with ε-Greedy

se ci basiamo sul valore medio di happiness, potremmo aver visitato poche volte un ristorante e quindi quella media potrebbe non essere valida.

"

Es: se R_1 è stato visitato 3 volte e $\mu_{R_1} = 2.9$

questo 2.9 non è selezionato da un n° suff di visite

"

QUI ENRA QUINDI UNA NUOVA STRATEGIA DI BILANCIATTENTO.

- AD OGNI TRY (t), PRENDIAMO UN RISTORANTE (R), tale che:

$$\left[\hat{\mu}_R + \sqrt{\frac{2\ln(t)}{N_t(r)}}\right] \text{ È MASSIMO}$$

HOFFDING'S INEQUALITY

LA MEDIA DEL REWARD PER R; BY GIVEN TRY

QUESTA PARTE del polinomio è l'incertezza che abbiamo basata sul n° di visite

- $N_t(r)$
  N° VOLTE CHE ABBIAMO VISITATO R_i
- $\ln(t)$
  NATURAL LOG OF TRY

LA SECONDA PARTE DEL POLINOMIO
È DETTA HOFFDING'S INEQUALITY

- CI FORNISCE UN UPPER BOUND DI PROBABILITÀ CHE LA SOMMA DELLE VARIABILI CASUALI INDIPENDENTI SI DISCOSTI DAL SUO VALORE ATTESO

MA TORNIAMO A NOI ...

SE ABBIAMO 2 RISTORANTI VISITATI CON LA STESSA MEDIA, NON CI BASIAMO SOLO SU DI ESSA MA IL SECONDO ~~FATTORE~~ DEL POLINOMIO SARÀ DETERMINANTE

N.B

QUEL VALORE È DA MASSIMIZZARE

IN TERMINI DI EXPLORATION E EXPLOITATION

$\hat{\mu}_R$: EXPLOITATION HELPER

$\sqrt{\frac{2\ln(t)}{N_t(r)}}$: EXPLORATION HELPER

QUANDO VISITIAMO UN RISTORANTE UNA VOLTA IN +, AUMENTIAMO IL DENOMINATORE IL CHE DETERMINA CHE MORO E MOW DIVENTERA' SEMPRE MENO LIKELY VISITARLO PERCHE' LA FRAZIONE DIMINUIRA'.