Settings:

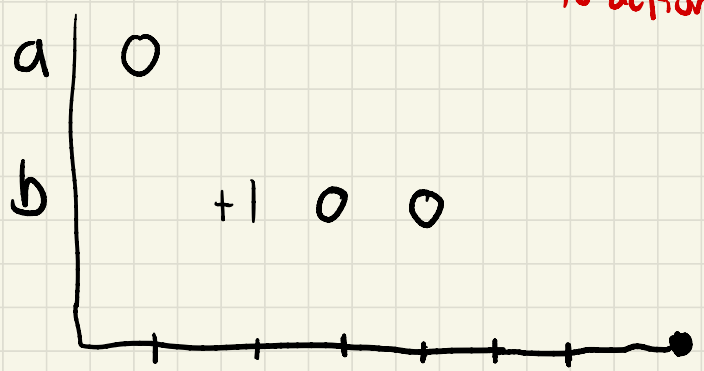
- Env. has a single state → returns reward, no observation
  - actions no longer have long-term consequences
  - actions gives immediate reward
  - can ignore other observations

How to learn policy in this setting?

Example:

rewards = 1 or 6

when do we switch back to action a?



- Learning agents must balance between 2 things

Exploitation :: maximise performance based on current knowledge

Exploration :: Increase Knowledge

Gather info. to make better decisions

- long term strategy may involve short term sacrifice

Formalising: Multi-Armed Bandit

- Multi-Armed Bandit is a set of distributions $\{R_a | a \in A\}$

R_a: distribution on rewards; a: known actions

- Each time step t, agent selects $A_t \in A$
- Env. generates $R_t \sim R_{A_t}$
- goal = maximise $\sum_{i=1}^{t} R_i$ → do this by learning policy

Action value: $q(a) = E[R_t | A_t = a]$

Optimal value: $V_* = \max_{a \in A} q(a) = \max_a E[R_t | A_t = a]$

regret : $\Delta_a = v_* - q(a)$

Want to minimise total regret

$$L_t = \sum_{n=1}^{t} v_* - q(A_n) = \sum_{n=1}^{t} \Delta A_n$$

Maximise cumulative reward $\equiv$ minimise total regret

Algorithms

several:

1) Greedy
2) ϵ-Greedy
3) UCB
4) thompson sampling
5) Policy gradients

} use action value estimates

$Q_t(a) \approx q(a)$

↳ estimate

simple one:

$$Q_t(a) = \frac{\sum_{n=1}^{t} I(A_n = a) R_n}{\sum_{n=1}^{t} I(A_n = a)}$$

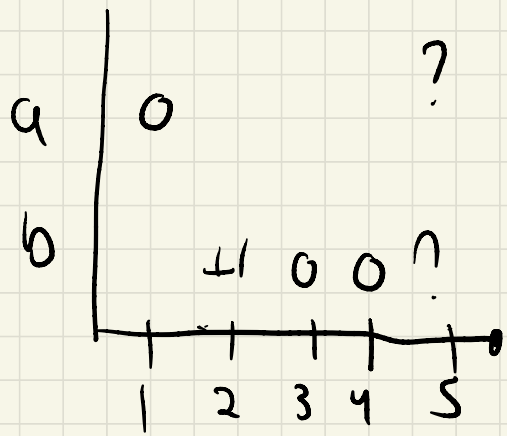
↳ average

$I(\text{true}) = 1$
$I(\text{false}) = 0$

1) Greedy

Select action with highest value: $A_t = \arg\max_a Q(a)$

regret:



$Q_4(a) = 0$

$Q_4(b) = 1/3$

↳ greedy would keep choosing action b

$p(+1|a) = 0.8 = q(a)$

$p(+1|b) = 0.2 = q(b)$

$\Delta a = V_{**} - q(a)$

$\Delta a = 0$ $\Delta b = 0.6$

?

2) ϵ-greedy

- Greedy gets stuck on suboptimal action forever
  ⇒ linear expected total regret.

Comes ϵ-greedy!

- with probability $1-\epsilon$ select optimal
- " " ϵ select random action

$$\pi_t(a) = \begin{cases} (1-\epsilon) + \epsilon/|A| & \text{if } Q_t(a) = \max_b Q_t(b) \\ \epsilon/|A| & \text{otherwise} \end{cases}$$

Continues to explore, still linear expected total regret

Policy Search → can we learn $\pi(a)$ directly, instead of learning values

Action preferences $H_t(a)$ and policy

$$\pi(a) = \frac{e^{H_t(a)}}{\sum_b e^{H_t(b)}}$$ softmax

↳ goal: optimize preferences

preferences $\neq$ values; learnable policy parameters

Use gradient ascent

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta E[R_t | \pi_{\theta_t}]$$

↳ vector

θ_t → current policy parameters

↳ how to compute gradient

Log-likelihood trick (see slides)

↳ can sample this

$$\nabla_\theta E[R_t | \theta] = E[R_t \nabla_\theta \log \pi_\theta(A_t)]$$

$$\theta = \theta + \alpha R_t \nabla_\theta \log \pi_\theta(A_t)$$

↳ stochastic gradient ascent

Note:

$$\sum_a \pi_\theta(a) = 1 \Rightarrow \forall b$$

$$\sum_a b \nabla_\theta \pi_\theta(a) = b \nabla_\theta \sum_a \pi_\theta(a) = 0$$

as long as b does not depend on θ or a.

We can substract a baseline and use:

$$\theta = \theta + \alpha (R_t - b) \nabla_\theta \log \pi_\theta(A_t)$$

↳ changes variance.

↳ gradient can be stuck on local optimum

theory! what is possible!

Theorem

$$\lim_{t\to\infty} \frac{L_t}{\log t} \geq \sum_{a|\Delta_a>0} \frac{\Delta_a}{KL(R_a \| R_{a^*})} \propto \Delta_a^2$$

↳ regret grows at least logarithmically.

↳ can we hit that lower bound?
↳ with log upper bound!

$$\Delta_a = V_{k^*} - q(a)$$

total regret depends on Δ_a and action count

$$L_t = \sum_{n=1}^{t} \Delta_{A_n} = \sum_{a \in A} N_t(a)\, \Delta_a$$

↳ good algorithms ensure small counts for large action regrets

UCB

- Estimate upper confidence $U_t(a)$ of action value

  s.t. $q(a) \leq Q_t(a) + U_t(a)$ with high probability

- Select action maximizing upper confidence bound (UCB)

$$a_t = \underset{a \in A}{\operatorname{argmax}}\ Q_t(a) + U_t(a)$$

↳ uncertainty depends on $N_t(a)$

↳ small $N_t(a)$ = high $U_t(a)$

↳ large = small $U_t(a)$

a only selected if - $Q_t(a)$ large

or

- $U_t(a)$ large

Hoeffding's inequality → see slides

UCB derivation

note: $\Delta_{a^*} = 0$

$A_t = \arg\max_a Q(a) + U_t(a)$

$L_t = \sum_a N_t(a)\,\Delta_a$

pseudo random arty

$\forall a \neq a^*$ $\quad N_t(a)\,\Delta_a \leq x_a \log t$

$m \leq t$ $\quad N_m(a)\,\Delta_a \leq x_a \log m \leq x_a \log t$

Latest possible time step which this holds.

$n \in \{m+1, \dots, t\} : N_n(a)\,\Delta_a > x_a \log n$

$$\mathbb{E}[N_t(a)] = \mathbb{E}\left[\sum_{n=1}^{t} I(A_n = a)\right]$$

$$= \mathbb{E}\left[N_m(a) + \sum_{n=m+1}^{t} I(A_n = a)\right]$$

$$= \mathbb{E}\left[x_a \cdot \frac{\log t}{\Delta_a} + \sum_{n=m+1}^{t} I(A_n = a)\right]$$

$$= x_a \frac{\log t}{\Delta_a} + \sum_{n=m+1}^{t} \mathbb{E}\left[I(A_n = a) \mid N_n(a)\,\Delta_a > x_a \log n\right]$$

$$= x_a \frac{\log t}{\Delta_a} + \sum_{n=m+1}^{t} p(A_n = a \mid \dots)$$

$p(A_n = a)$? given $N_m(a)\Delta_a > x_a \log n$ ①

$\leq P(Q_t(a) + U_t(a) \geq Q_t(a^*) + U_t(a^*))$

$= P(Q + U \geq Q^* + U^*)$? ≤ 1

inequality seen before! $\leq \frac{1}{n}$

$= P(Q + U \geq Q^* + U^* \mid Q^* + U^* \leq y) \cdot P(Q^* + U^* \leq y)$ ≤ 1

$+ P(Q + U \geq Q^* + U^* \mid Q^* + U^* \geq y) \cdot P(Q^* + U^* \geq y)$ $\leq \frac{1}{n}$

$\leq \frac{1}{n(a^*)}$ missing a 2 $e^{-N_n(a^*) U_n(a^*)^2}$ PICK $y = q(a^*)$

$P(Q^* + U^* \leq q^*) \leq e^{-N U^2}$

$\sum_{n=1}^{t} \frac{1}{n} < \log t + 1$

$e^{-NU^2} = \frac{1}{n} \Rightarrow U \geq \sqrt{\frac{\log n}{N}}$

Now Bound:

$P(Q + U \geq Q^* + U^* \mid Q^* + U^* \geq q^*) \leq \frac{1}{n}$?

$\leq P(Q + U \geq q^*) = P(-Q - U \leq \underbrace{-q^* + q}_{= -\Delta q} - q)$

$= P(-Q + \Delta a - U \leq q)$ Round this

$\Delta a \geq 2U$?

(1): $N \cdot \Delta a \geq c \log n$

$\Delta a \geq c \frac{\log n}{N_t(a)}$

$= x_a (U_t(a))^2$

if we pick $t_a = \frac{1}{\Delta_a} \Rightarrow \Delta_a^2 (u+a)^2$

$\Delta_a (u+a)$!

pick $t_a = \frac{u}{\Delta_a} \Rightarrow \Delta_a (2u)$ wow

$\Rightarrow P(-Q + \Delta_a - u \le -u) \le P(-Q + 2u - u \le -a)$

$= P(-Q + u \le -a) \le e^{\frac{-N \cdot u^2}{n}}$ | $u = \sqrt{\frac{\log n}{N}}$

$\frac{1}{n}$

$\Rightarrow P(A_n > a) \le \frac{2}{n}$

$\Rightarrow \sum_{n=N+1}^{t} P(A_n > a) \le 2(\log t + 1)$