Reinforcement Learning

Christos Dimitrakakis

December 12, 2023

The multi-armed bandit (MAB) problem

- At time t:
- Select action $a_t \in A$
- Obtain reward $r_t \in \mathbb{R}$

Basic objective

Maximise total reward

$$U = \sum_{t=1}^{T} r_t,$$

where T is the horizon. It may be unknown, or random.

Regret

We can instead minimise total regret

$$L = \sum_{t=1}^{T} [r_t^* - r_t],$$

where r^* is the reward an oracle that knew the "best" arm would have obtained.

No let's make this more precise.

The stochastic MAB

For each arm $i \in A$:

- $r_t \mid a_t = i \sim \mu_i$ is the reward distribution
- $\rho_i \triangleq E_\mu[r_t \mid a_t = i]$ the expected reward
- $\rho^* \triangleq \max_i \rho_i$.

Policy

The policy $\pi \in \Pi$ is a adaptive: $\pi(a_t \mid a_{t-1}, r_{t-1}, \ldots, a_1, r_1)$

Objective

Maximise expected total reward

$$\mathbb{E}^\pi_\mu[U] = \mathbb{E}^\pi_\mu\left[\sum_{t=1}^T r_t\right]$$

The total expected regret is

$$\mathbb{E}^\pi_\mu[L] = \mathbb{E}^\pi_\mu\left[\sum_{t=1}^T \rho^* - \rho_t\right]$$

The horizon

Discounted T

- $U = \sum_{t=1}^{T} \gamma^{t-1} r_t$
- Same as non-discounted with stopping probability $(1 - \gamma)$.

Arbitrary T

To compare algorithms, we use the notion of regret growth

- Linear regret: $L_T = O(T)$. i.e. insufficient learning
- Sub-linear regret, e.g. $L_T = O(\sqrt{T})$ or $O(\ln T)$.

Algorithms

ϵ-greedy

UCB

Thompson sampling

The Markov decision process

The value of a policy