Contents

Chapter 1

Multi-armed bandits

1.1 Problem formulation

In the multi-armed bandit problem, there are k distributions ('arms'), $\{P_1, P_2, \ldots, P_k\}$ with unknown means $\{\mu_1, \mu_2, \ldots, \mu_k\}$. For a given number of turns T, the goal is to maximise the expected reward by iteratively selecting a distribution to sample from. In particular, the goal is to minimise the regret defined as

$$R(T) = \sum_{t=1}^{T} \mu^* - \mu_t, \tag{1.1}$$

where μ^* is the mean of the distribution with the highest mean, and μ_t is the mean of the distribution selected at time t. The number of turns T is often referred to as the horizon and can be assumed to be greater than k. This poses a constant struggle between exploration and exploitation, where exploration is the process of trying out new distributions, and exploitation is the process of using the distribution with the highest mean.

Almost always, assumptions are made about the distributions. Otherwise, composing algorithm with any sort of optimality guarantee would be futile. A common assumption, for example, is that the distributions are Bernoulli. Often they are assumed to be Gaussian with unknown mean and maybe some restrictions on the variance. If no assumptions are made to the type of distributions, there are likely to be assumptions made to the variance or support of the distributions.

1.1.1 Alternative problems

An alternative problem is to find the best arm with as few turns as possible. In this version, a δ is given, and the goal is to find the best arm with probability at least $1 - \delta$.

1.1.2 Generalisations

There are many generalisations to the multi-armed bandit problem. For instance, the distributions may not be stationary, but instead change throughout the game. Alternatively, with contextual bandits, information about a context is given before each turn, which must then be taken into account when selecting an arm. Adversarial bandits complicates matters further, where the rewards are not stochastic from some distribution, but are instead selected by an adversary.

1.2 Strategies

1.2.1 Explore-only

Pure exploration is obviously a suboptimal strategy, but it is a good baseline to compare against. It can be implemented by selecting an arm uniformly or in order, but it will perform poorly either way. The arm-selection procedure is described by algorithm 2.

Algorithm 1 Random arm selection

```
procedure SelectArm(t)
    Sample i from {1, ..., k} uniformly
    return i
```

It is easy that the expected regret is

$$R(T) = T\left(\mu^* - \frac{1}{k}\sum_{i=1}^{k}\mu_i\right), \tag{1.2}$$

which is $\Theta(T)$. This motivates the search for an algorithm with sublinear regret.

1.2.2 Greedy

A simple algorithm and a good baseline is the greedy algorithm. Here, all arms are tried N initial times, and the empirical means are used to select the best arm. Afterwards, the arm with the highest empirical mean is selected for all

remaining turns. The arm-selection procedure is listed in algorithm 2, where $\hat{\mu}_i$ is the empirical mean of arm i.

Algorithm 2 Greedy arm selection

```
procedure SelectArm(t)
    if t ≤ Nk then
        return (t mod k) + 1
    else
        return argmax_{i=1,...,k} μ̂_i
```

With greedy selection, the expected regret is

$$R(T) = \sum_{t=1}^{T} \mu^* - \hat{\mu}_t, \tag{1.3}$$

1.2.3 Epsilon-greedy

The problem with the greedy algorithm is that it may be unlucky and not discover the best arm in the initial exploration phase. To mitigate this, the epsilon-greedy algorithm may be used. In this algorithm, the estimated arm is pulled with probability $1 - \epsilon$ and a random arm is pulled with probability ϵ. This ensures convergence to correct exploitation as the horizon increases, and it will generally reduce the regret. Choosing ϵ is a trade-off between exploration and exploitation and can significantly affect the regret.

Epsilon-decay

If one allows the ϵ to decay over time, the regret can be reduced even further. This makes sense, as exploration is less worthwhile as estimates become ever more accurate. A common choice is to decay ϵ as $\epsilon_t = \epsilon_0/t$.

1.2.4 UCB

The upper confidence bound (UCB) algorithm is a more sophisticated algorithm based on estimating an upper bound for the mean of each arm. In the original formulation, where support on only $[0, 1]$ is assumed, the upper confidence bound is given by

$$\mathrm{UCB}_t(a) = \hat{\mu}_t(a) + \sqrt{\frac{2 \ln t}{N_t(a)}}, \tag{1.4}$$

where $\hat{\mu}_t(a)$ is the empirical mean of arm a at time t, $N_t(a)$ is the number of times arm a has been pulled at time t. It generally achieves logarithmic regret.

Many variants of the UCB algorithm exist. Different assumptions about the distributions changes the confidence bounds.

1.2.5 Bayesian: Thompson sampling

Thompson sampling is a Bayesian approach to the multi-armed bandit problem, originally described in 1933 as a way to handle the exploration-exploitation dilemma in the context of medical trials. The idea is to sample from the posterior distribution of the means of the arms and pull the arm with the highest sample.