Mathematical machine learning part IV : active and online learning

Prof. Dr. Gilles Blanchard, Dr. Alexandra Carpentier*, Dr. Jana de Wiljes, Dr. Martin Wahl

In most classical problems, one considers settings where all data are available before hand. But this is not always the case and in many applications, the data become available gradually - this is the *online learning* setting. Sometimes, the learner does even have an impact on how the data is collected - this is a specific case of the online learning setting, which is called *active learning*. We are going to investigate a specific and relatively simple (yet mathematically challenging) example of active learning, which is called the *bandit problem*.

1. The stochastic bandit problem

`Useful material : See Bubeck et.al (2012), and also Cesa-Bianchi et.al (2006) for a broader perspective - see also https://blogs.princeton.edu/imabandit/2016/05/11/bandit-theory-part-i/ (and part ii) for a helpful blog post.`

1.1. The problem

The bandit setting is an online learning setting, where the learner gets to choose which source of data it wants to observe among many. In the stochastic bandit setting, we assume that the sources output data randomly in a i.i.d. fashion.

Now let us say this in a more specific way. The learner can sample K data sources which are often referred to as "arms". At each time t, the learner chooses one of the systems $k_t \in \{1, \ldots, K\}$ it wants to observe. This decision is not based on the data observed in the past $(X_u)_{u<t}$. After choosing k_t, it receives $X_t \sim \nu_{k_t}$. At the end of the game at time n (the game is said to be of horizon n), the performance of the learner is measured by

$$L_n = \sum_t X_t.$$

Let us assume in the sequel that the distributions $(\nu_k)_k$ have support in $[0,1]$, and let us write μ_k for their means, $\mu^* = \max_k \mu_k$, $\Delta_k = \mu^* = \mu_k$ and $T_{k,n}$ for the number of times arm k was chosen. A related objective is to make the *expected regret* with respect to the best arm

$$\bar{R}_n = n\mu^* - \mathbb{E}\sum_t X_t = \sum_k \Delta_k \mathbb{E}T_{k,n},$$

as small as possible, with respect to the arm selection $(k_t)_t$ of the learner.

*Contact : carpentier@math.uni-potsdam.de. **Webpage with course material TBA :** `http://www.math.uni-potsdam.de/~carpentier/page3.html`

In order to simplify the analysis of this game, we write $X_{k,u}$ for the uth data observed from system k if system k is sampled at least u times.

Game 1: The stochastic bandit game.

```
Unknown infos: $(\nu_k)_k$
Known parameters: $K$ and $n$
for $t = 1, \dots, n$ do
    The player chooses $k_t \in \{1, \dots K\}$
    The system $k_t$ reveals the reward $X_t \sim \nu_{k_t}$
end for
Goal : Maximize over $(k_t)_t$ the sum $L_n = \sum_{t \leq n} X_t$.
```

Global objective : Propose good strategies for solving Game 1 and minimizing $\bar{R}_n$:

- Propose strategies : upper bounds.
- Prove optimality of these strategies : lower bounds.

1.2. The stochastic bandit problem - upper bounds

A popular simple strategy for this problem is the UCB-strategy .

Write $T_{k,t}$ for the number of times arm k has been pulled at time t and $\hat{\mu}_{k,t}$ for the empirical mean of arm k at time t.

Algorithm 1: The UCB strategy.

```
Initialisation: Pull a sample from each distribution
for $t = K + 1, \dots, n$ do
    Set $UCB_{k,t} = \hat{\mu}_{k,t} + 4\sqrt{\frac{\log(n)}{T_{k,t}}}$.
    Set $k_t = \arg\max_k UCB_{k,t}$ and collect $X_t \sim \nu_{k_t}$.
end for
```

The following theorem holds for bounding the pseudo-regret.

Theorem 1. *Assume that $n \geq 4K$. It holds that*

$$\bar{R}_n \leq 10\Big(1 + \sum_{k:\Delta_k > \sqrt{K\log(n)/n}} \frac{\log(n)}{\Delta_k} + \mathbf{1}\{\exists k : 0 < \Delta_k < \sqrt{K\log(n)/n}\}\sqrt{nK\log(n)}\Big).$$

This implies in particular the so-called problem-dependent bound

$$\bar{R}_n \leq 20\Big(1 + \sum_{k:\Delta_k > 0} \frac{\log(n)}{\Delta_k}\Big),$$

and the problem-independent bound

$$\bar{R}_n \leq 30\sqrt{nK\log(n)}.$$

1.3. Exercises : part 1

Bandit problem with a different objective. Consider the stochastic bandit setting where now the objective is not to maximise the sum of collected samples, but to find with probability as high as possible the arm with highest mean $k^* = \arg\max_k \mu_k$. If at the end of the budget the learner guesses $\hat{k}_n$, the objective is to minimise

$$\mathbb{E}\Delta_{\hat{k}_n} \quad \text{or also} \quad \mathbb{P}(\hat{k}_n \neq k^*).$$

We assume that there is just one optimal arm, i.e. that $\forall k \neq k^*$, then $\mu_k < \mu_{k^*}$. Toward this end, we consider the strategy

Algorithm 2: UCB-A strategy (UCB-A).

```
Parameter: $a > 0$
Initialisation: Pull a sample from each distribution
for $t = K+1, \dots, n$ do
    Set $UCBA_{k,t} = \hat{\mu}_{k,t} + 2\sqrt{\frac{na}{T_{k,t}}}$.
    Set $k_t = \arg\max_k UCBA_{k,t}$ and collect $X_t \sim \nu_{k_t}$.
end for
Output : $\hat{k}_n = \arg\max_k T_{k,n}$.
```

1. Here the $\log(n)$ from UCB is replaced by na, which is much larger if $a > 0$ is a fixed constant. Do you think it is a good idea and why?
2. Prove that
$$\xi = \{\forall k, \forall t, |\hat{\mu}_{k,t} - \mu_k| \leq 2\sqrt{\frac{na}{T_{k,t}}}\}$$
is such that $\mathbb{P}(\xi) \geq 1 - nK\exp(-na)$.
3. Let us now pose for the parameter
$$a^{-1} = \sum_{k \neq k^*} \frac{1}{20\Delta_k^2}.$$
Prove that for some $n \geq CH$ where $C > 0$ is an universal constant, for any sub-optimal arm (i.e. that is not k^*), it holds on ξ that
$$T_{k,t} < 40\frac{n}{\Delta_k^2 \sum_{i \neq k^*} \Delta_i^{-2}}.$$
4. Deduce from this a useful problem-dependent bound on $\mathbb{E}\Delta_{\hat{k}_n}$ or also $\mathbb{P}(\hat{k}_n = k^*)$.
5. From the bounds on $T_{k,t}$ on ξ, deduce a problem independent bound on $\mathbb{E}\Delta_{\hat{k}_n}$.
6. What happens if a^{-1} is taken smaller than $\sum_{k \neq k^*} \frac{1}{20\Delta_k^2}$? And if it is taken larger?

References

Bubeck, Sebastien, and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1-122, 2013.

Cesa-Bianchi, Nicolo, and Gabor Lugosi. Prediction, learning, and games. *Cambridge University Press*, 2006.