Mathematical machine learning part IV : active and online learning

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

Here are some precisions on the two-armed bandit case. At the end of this document are two very useful classical tools : Hoeffding's inequality, and union bounds (plus application to bandit data).

Special notes : the two armed stochastic bandit problem

Consider the two armed bandit problem with distributions ν_1 and ν_2, that both have support on $[0,1]$, and associated means μ_1, μ_2 and gap $\Delta = |\mu_1 - \mu_2| > 0$. Write $k^* = \arg\max(\mu_1, \mu_2)$.

First naive strategy

Consider the strategy that tries both distributions u times (up to time $t = 2u$ then), and then from time $t = 2u+1$ on always picks the distribution $\hat{k} = \arg\max(\hat{\mu}_1, \hat{\mu}_2)$ until time n. This is described in Algorithm 3.

Algorithm 2: First naive two-armed strategy.

Parameter: u
Initialisation: Pull u samples from each distribution
for $t = 2u+1, \dots, n$ **do**
  Set $k_t = \hat{k} = \arg\max(\hat{\mu}_{1,2u}, \hat{\mu}_{2,2u})$ and collect $X_t \sim \nu_{\hat{k}}$.
end for

Without loss of generality, assume that $\mu_1 < \mu_2$. It holds that

$$\begin{aligned}\mathbb{P}(\hat{k} \neq k^*) &= \mathbb{P}(\hat{\mu}_{1,2u} > \hat{\mu}_{2,2u}) \\ &= \mathbb{P}(\hat{\mu}_{1,2u} - \hat{\mu}_{2,2u} > 0) = \mathbb{P}\Big(\frac{1}{u}\sum_{i\leq u}(X_{1,i} - X_{2,i} + \Delta) > \Delta\Big).\end{aligned}$$

We now apply Hoeffding's inequality from Theorem 1 to obtain that

$$\mathbb{P}\Big(\frac{1}{u}\sum_{i\leq u}(X_{1,i} - X_{2,i} + \Delta) > \Delta\Big) \leq 2\exp(-u\Delta^2/2),$$

as the $Z_i = X_{1,i} - X_{2,i} - \Delta$ are in $[-2,2]$ and have mean 0.

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

The expected regret of this strategy is

$$\begin{aligned}\bar{R}_n &= u\Delta + (n-2u)\Delta\mathbb{P}(\hat{k} \neq k^*) \\ &\leq u\Delta + n\Delta 2\exp(-u\Delta^2/2).\end{aligned}$$

Assume first that $8\log(n\Delta^2)/\Delta^2 \leq n$. Then setting $u = \lfloor 2\log(n\Delta^2)/\Delta^2 \rfloor + 1$ we get

$$\bar{R}_n \leq 3\frac{\log(n\Delta^2)}{\Delta} + \frac{2}{\Delta} \leq 3\frac{\log(n\Delta^2)}{\Delta}.$$

Otherwise if $8\log(n\Delta^2)/\Delta^2 \geq n$, we set $u = n/2$ and have

$$\bar{R}_n = n\Delta/2 \leq 4\sqrt{n}.$$

Remarks :

- Both bounds are sub-linear - in particular they are much smaller that the average gain of the best oracle strategy $n\max(\mu_1\mu_2)$ if $\max(\mu_1\mu_2) > 0$ does not depend on n.
- These values of u minimise the bound - we will see later that they are in some sense optimal.
- The first bound depends on Δ - it is called problem dependent. The second one does not - it is called problem independent.

Porblem : u is then a parameter of the algorithm - we cannot callibrate it with Δ if we do not know Δ.

Better strategy

Consider the strategy that picks both distributions the same amount of time until one of them seems clearly better than the other, the comparison being made using a confidence interval.

Algorithm 2: Better two-armed strategy.

Initialisation: Pull one sample from each distribution
while $|\hat{\mu}_{1,t} - \hat{\mu}_{2,t}| \leq 2\sqrt{\frac{\log(4n^2)}{t}}$ **do**
  Pull one sample from each distribution
  $t \leftarrow t + 2$
end while
Set $T + 1$ for the moment of exit of the while loop.
for $t = T + 1, \ldots, n$ **do**
  Set $k_t = \hat{k} = \arg\max(\hat{\mu}_{1,T}, \hat{\mu}_{1,T})$ and collect $X_t \sim \nu\hat{k}$
  $t \leftarrow t + 1$
end for

Let us write

$$\xi = \Big\{\forall k \leq 2, \forall t \leq n, \big|\frac{1}{t}\sum_{i\leq t} X_{k,i} - \mu_k\big| \leq \sqrt{\frac{\log(4n^2)}{2t}}\Big\}.$$

By Corollary 1, it holds that

$$\mathbb{P}(\xi) \geq 1 - 1/n.$$

Note first that on ξ, it holds by definition that $\hat{k} = k^*$. Note also that on ξ, we have that

$$\Delta \leq 4\sqrt{\frac{\log(4n^2)}{T}}, \quad \text{i.e.} \quad T \leq 16\frac{\log(4n^2)}{\Delta^2}.$$

So the expected regret of this strategy is

$$\begin{aligned}\bar{R}_n &\leq \mathbb{E}T\Delta + n\Delta\mathbb{P}(\hat{k} \neq k^*) \\ &\leq 16\frac{\log(4n^2)}{\Delta^2} + \Delta + 1 \leq 100\frac{\log(2n)}{\Delta^2}.\end{aligned}$$

Remarks :

- No parameter in this strategy, and same results as before!
- Very close to the UCB-strategy, i.e. sample the distribution that maximise the upper confidence bound. Such a proof is very similar to the proof of UCB, but with two distributions.

Useful classical results

Hoeffding's inequality

We recall Hoeffding's inequality.

Theorem 1 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables such that their distribution ν has support in $[0,1]$, and mean μ. It holds that*

$$\mathbb{P}\Big(|\frac{1}{n}\sum_i X_i - \mu| > \epsilon\Big) \leq 2\exp\Big(-2n\epsilon^2\Big).$$

Equivalently, it holds with probability larger than $1-\delta$ that

$$|\frac{1}{n}\sum_i X_i - \mu| \leq \sqrt{\frac{\log(2/\delta)}{2n}}.$$

See e.g. in Bubeck et.al (2012) for a proof.

Remarks :

- Hoeffding's inequality is proven through the MGF, basically bounding the MGF of the $X_i - \mu$ by the one of a Rademacher random variable. Intuitvely, the idea is that bounded random variable on $[-1,1]$ and of mean 0 are dominated stochastically by a Rademacher random variable - and so the sum of the $\sum_i X_i - n\mu$ is stochastically dominated by the sum of n i.i.d. Racemacher. Hoeffding's inequality is tight up to constants in this case.
- The bound on the probability of deviations of maginitude more than ϵ of $\frac{1}{n}\sum_i X_i - \mu$, namely $2\exp\Big(-2n\epsilon^2\Big)$, is of same order as the probability of deviations of maginitude more than ϵ of a Gaussian of mean 0 and variance n. For this reason, it is said that $\frac{1}{n}\sum_i X_i - \mu$ is sub-Gaussian under the assumption of the previous theorem.

Union bound

We recall the union bound principle (first principles in probability).

Theorem 2 (Union bound)**.** *Let $\xi_1, \dots, \xi_n$ be events of probability respcetively $1-\delta_1, \dots, 1-\delta_n$. It holds that*

$$\mathbb{P}\Big(\bigcap_i \xi_i\Big) \geq 1 - \sum_i \delta_i.$$

This simple basic inequality is not always tight, but it can be extremely useful when combined with Hoeffding's inequality.

Corollary 1. *Let for any $k \leq K$, $X_{k,1}, \dots, X_{k,n}$ be i.i.d. random variables such that their distribution ν_k has support in $[0,1]$, and mean μ_k. It holds that*

$$\mathbb{P}\Big(\exists k \leq K, \exists t \leq n, |\frac{1}{t}\sum_{i\leq t} X_{k,i} - \mu_k| > \frac{\epsilon}{\sqrt{t}}\Big) \leq 2nK \exp\Big(-2\epsilon^2\Big).$$

Equivalently, it holds with probability larger than $1 - nK\delta$ that $\forall k \leq K, \forall t \leq n$

$$|\frac{1}{t}\sum_{i\leq t} X_{k,i} - \mu_k| \leq \sqrt{\frac{\log(2/\delta)}{2t}}.$$

Or again, equivalently, it holds with probability larger than $1-\delta$ that $\forall k \leq K, \forall t \leq n$

$$|\frac{1}{t}\sum_{i\leq t} X_{k,i} - \mu_k| \leq \sqrt{\frac{\log(2nK/\delta)}{2t}}.$$

Remarks :

- Union bound plus Hoeffding is basically the only probabilistic tools that are needed for the understanding of standard stochastic bandit algorithms.
- The union bound is not tight and can be slightly improved - but not by a large extent.
- The corollary above is the first step in most standard stochastic bandit proofs - and all stochastic bandit proofs in this class. Once this step is done, bandit proofs usually amount in proving that on the large event where empirical means concentrate well around their means, the algorithm behaves as wished *even in the worst case of this event.*

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.