Bourbaki vs Pragmatism
A methodological comparison through the multi-armed bandits problem

Sebastiano Ferraris*

March 1, 2020

In these pages we compare two different mathematical methodologies in approaching the well known multi-armed bandits problem: the Bourbakist and the pragmatic. The Bourbakist way is concerned with the mathematical foundation upon which a formal solution is derived in the shape of an axiomatic structure. In contrast, the pragmatist approach aims at finding the shortest path towards a solution, reducing the mathematical formalisms to the bare minimum. The article introduces the problem, shows the two approaches and ends with a critical comparison.

If you came across this article when searching for an introduction to the multi-armed bandit problem, and not a methodological comparison, you may still find what you are looking for in section 1, 3 and in the bibliography (please do ignore completely section 2). The code to create the figures and run a range of algorithms to solve the problem is implemented in text.

1 Multi armed bandits problem

Consider of being in the situation of having to repeatedly choose between K different options, each having a cost and a possible cash reward. For each option the cost is fixed and the reward is drawn from an *unknown* probability distribution, constant across time.

The problem of finding a strategy to maximise the reward is named multi armed bandits (MAB) after the situation of playing at a row of K slot machines (or single armed bandit). Given an initial amount of money of \$1000 and a costs of \$1 for each draw, the

*sebastiano.ferraris@gmail.com

player must balance an exploration phase when estimating the unknown distributions of each arm, with an exploitation phase, when the acquired knowledge is used for a gain[1].

The problem can be generalised to clinical or pre-clinical trials, control engineering, mechanical and software testing, stock market investments, behavioural modelling, dynamic pricing, and many more[2].

2 The Bourbakist perspective

Let $(\Omega, \mathcal{A})$ be a σ-algebra defined as a couple of a non-empty set Ω paired with a subset of its power set, containing the empty set and closed under numerable union and complement set. Let the latter be called *action space*. Let $\mathcal{I}_K = \{1, 2, \dots, K\} \subset \mathbb{N}$ be a set of indexes whose generic element k is called *arm* by convention. Let $\mathcal{I}_T = \{1, 2, \dots, T\} \subset \mathbb{N}$ be another set whose elements are called conventionally *time*. Let $\mathbb{R}_+$ the positive real axis including the zero. The relationship between the above defined elements are given by a function A defined as:

$$\begin{aligned} A : \mathcal{I}_T \times \mathcal{A} &\longrightarrow \mathcal{I}_K \\ (t, \omega) &\longmapsto A(t, \omega) = A_t(\omega) \end{aligned}$$

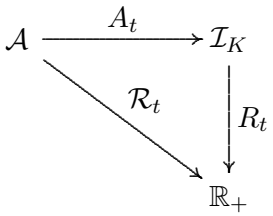
and by a function R, defined as:

$$\begin{aligned} R : \mathcal{I}_T \times \mathcal{I}_K &\longrightarrow \mathbb{R}_+ \\ (t, \omega) &\longmapsto R(t, k) = R_t(k) \end{aligned}$$

Let the former be called *action* and the latter be called *reward*. Let

$$\begin{aligned} \mathcal{R} : \mathcal{I}_T \times \mathcal{A} &\longrightarrow \mathbb{R}_+ \\ (t, \omega) &\longmapsto R(t, \omega) = \mathcal{R}_t(\omega) \end{aligned}$$

another function, defined as the only possible function making the diagram below commutative.



for each $t \in \mathcal{I}$. Let $\mathcal{R}_t$ be called, again by convention, *reward map at the time t*.
We observe that $\mathcal{R}$ maps the events of the σ-algebra, while R maps the corresponding

[1] See Thompson [Tho33] for an early approach where the two arms are two medical treatments, Bellman [Bel56] where the problem is formulated in a Bayesian perspective for two arms, and the more recent Sutton [SB18], chapter 1, for a reinforcement learning perspective of MAB.

[2] See Bouneffouf [BR19] for a survey with a list of applications of the main algorithms solving the MAB problem.

indexes. This is an analogous of the definition of probability respect to the one of random variable and probability density function, for when the real axis image is normalised.

Now consider

$$\begin{aligned} Q : \mathcal{I}_T \times \mathcal{I}_K &\longrightarrow \mathbb{R}_+ \\ (t, \omega) &\longmapsto \mathcal{Q}(t,k) = \mathcal{Q}_t(k) \end{aligned}$$

the *estimated reward of the action* ω *up to time* t, for $\omega = A_t^{-1}(k)$ for a fixed $k \in \mathcal{I}_K$, with the corresponding function $\mathcal{Q} : \mathcal{I}_T \times \mathcal{A} \to \mathbb{R}$. It follows that Q is therefore defined application of the mean value in a Lebesgue space over $(\Omega, \mathcal{A})$, that is now a Borel σ-algebra[3] as:

$$Q_t(k) = \mathbb{E}\left[\mathcal{R}_\tau(\omega) \mid A_\tau^{-1}(k) = \omega, \tau \in \mathcal{I}_T, t \leq t\right] \qquad t \in \mathcal{I}_T \qquad k \in \mathcal{I}_k \tag{1}$$

and therefore

$$\mathcal{Q}_t(\omega) = \mathbb{E}\left[R_\tau(k) \mid A_\tau(\omega) = k, \tau \leq t\right] \qquad \omega \in \mathcal{A} \qquad t \in \mathcal{I}_T \tag{2}$$

As $\mathcal{R}$ and R, also $\mathcal{Q}$ and Q satisfies the commutativity of an analogous of the diagram shown above. The notation can be simplified for brevity to:

$$\mathcal{Q}_t(\omega) = \mathbb{E}\left[R_t(k) \mid A_t = k\right] \tag{3}$$

where the mean value is for all the time indexes up to t and where the domain values of A_t is clear from the context.

We now define the *total reward* $Q_\infty : \mathcal{I}_K \to \mathbb{R}$ as

$$Q_\infty(k) = \mathbb{E}\left[\mathcal{Q}_t(\omega) \mid A_t^{-1}(k) = \omega, t \in \mathcal{I}_T\right] \qquad k \in \mathcal{I}_K \tag{4}$$

or with the simplified notation as[4]:

$$Q_\infty(k) = \mathbb{E}\left[R_t(k) \mid A_t = k\right] \tag{5}$$

So far we have been considering the reward and the total reward for a fixed choice of k. We can vary $k \in \mathcal{I}_K$ in function of the time index. So let $\mathbf{k}$ an element of

$$\mathcal{I}_K^T = \underbrace{\mathcal{I}_K \times \mathcal{I}_K \times \cdots \times \mathcal{I}_K}_{T\text{-times}}$$

or equivalently a function of $\mathcal{I}_T$ to $\mathcal{I}_K$. Definitions 2 and 3 are so generalised to $Q_t : \mathcal{I}_K^T \to \mathbb{R}$ for $t \in \mathcal{I}_{\leq T}$, having defined $\mathcal{I}_{\leq T}$ any interval of positive integers between 1 and T, and

$$Q_t(\mathbf{k}) = \mathbb{E}\left[\mathcal{R}_\tau(\omega) \mid A_\tau^{-1}(\mathbf{k}_\tau) = \omega, \tau \leq t\right] \qquad \mathbf{k} \in \mathcal{I}_K^T \qquad t \in \mathcal{I}_T$$

[3] For a foundational perspective, see Bourbaki [Bou04a].

[4] The simplified notation is often the only notation appearing in engineering textbooks (e.g. Sutton [SB18]), although this would not allow the reader to understand the subtle formalisation of assigning to an event ω its index k. More tragically the simplified notation makes most of the concepts introduced so far pedantic and irrelevant.

and therefore

$$Q_\infty(\mathbf{k}) = \mathbb{E}\left[\mathcal{R}_t(\omega) \mid A_t^{-1}(\mathbf{k}_t) = \omega, t \in T\right] \qquad \mathbf{k} \in \mathcal{I}_K^T$$

The mean value computed with a Lebesgue measure, over the Borel space generated as the sets of images R of [5]

3 The pragmatic perspective

4 Conclusions

References

[Bel56] Richard Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 16(3/4):221–229, 1956.

[Bou04a] Nicolas Bourbaki. Integration. Springer Berlin, 2004.

[Bou04b] Nicolas Bourbaki. Theory of sets. Springer Berlin, 2004.

[BR19] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*, 2019.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Tho33] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[TZ82] Gaisi Takeuti and Wilson M Zaring. Introduction to axiomatic set theory. Springer, 1982.

[5] We consider the definition under the accordance with the axiom of choice as in the ZFC axiomatic set theory, in order to avoid "virages dangereux". See also Bourbaki [Bou04b] and [TZ82].