A Bayesian Modern look at multi armed bandits

ENSAE PARIS

Charlotte DE ROMÉMONT
charlotte.deromemont@ensae.fr

Suzie GRONDIN
suzie.grondin@ensae.fr

Marion CHABROL
marion.chabrol@ensae.fr

To run the code and explore the results, please refer to the notebooks available in the following GitHub repository : https://github.com/mchabrol/Bayesian-Statistics-

Année 2024 - 2025

I Definition of the problem under study

The Multi-Armed Bandit Problem

The multi-armed bandit problem describes a sequential decision-making scenario where the goal is to maximize rewards from a set of actions (or "arms") with unknown payoff distributions. At each step, the experimenter must decide which arm to pull, balancing :

— Exploitation : Selecting the arm believed to offer the highest reward based on current knowledge ;

— Exploration : Trying less-tested arms to learn more about their potential.

The example given in the paper is a website testing scenario where different layouts are being evaluated to maximize the conversion rate (e.g., user sign-ups). Each layout represents an "arm" in the multi-armed bandit problem, with an unknown probability of success. At each step, visitors to the website are assigned to one of the layouts, and the observed outcomes (conversion or no conversion) provide feedback about its effectiveness.

Formally, let $y_t = (y_1, \ldots, y_t)$ denote the sequence of observed rewards and a_t the arm chosen at time t. Each y_t is sampled from a distribution $f_{a_t}(y|\theta)$, where θ is an unknown parameter vector. The goal is to maximize cumulative rewards over time while incrementally learning the true parameters of θ.

Performance of an algorithm

The performance of algorithms in the multi-armed bandit problem is typically evaluated using the regret metric. Regret measures the cumulative difference between the reward that could have been obtained by always selecting the optimal arm and the reward actually accumulated by the algorithm.

Formally, let $\mu^*(\theta) = \max_a \mu_a(\theta)$, where $\mu_a(\theta) = \mathbb{E}[y|\theta, a]$ is the expected reward for arm a. If $n_a(t)$ represents the number of times arm a has been played by time t, then the cumulative regret at time T is defined as :

$$L_T = \sum_{a=1}^{k} n_a(T) \left[\mu^*(\theta) - \mu_a(\theta)\right].$$

— Regret is zero if the algorithm always plays the optimal arm from the very beginning,

— Regret grows when the algorithm spends time exploring suboptimal arms, incurring a loss in reward relative to the optimal strategy.

In practice, we evaluate the cumulative regret L_T, which quantifies the total reward lost over T time steps compared to the optimal policy. This is used to compare the efficiency of different algorithms in identifying and prioritizing the best arm.

II Bayesian solution : Randomized Probability Matching (RPM)

The Randomized Probability Matching (RPM) algorithm manages the exploration-exploitation trade-off in the multi-armed bandit problem by allocating observations proportionally to the posterior probability of each arm being optimal. The steps are as follows :

— **Initialization :** Specify a prior distribution $p(\theta)$ for the parameters θ governing the reward distributions $f_a(y|\theta)$ of the arms $a = 1, \ldots, k$.

— **Posterior Update :** At each time t, given the observed sequence of rewards $y_t = (y_1, \ldots, y_t)$ and selected arms a_t, update the posterior distribution :

$$\begin{aligned} p(\theta|y_t) &\propto p(\theta) p(\mathbf{y}_t \mid \theta) \\ &\propto p(\theta) \prod_{\tau=1}^{t} f_{a_\tau}(y_\tau|\theta) \end{aligned}$$

— **Optimality Probabilities :** For each arm a, compute the probability that it is optimal :

$$w_{a,t} = \Pr(\mu_a = \max(\mu_1, \ldots, \mu_k)|y_t),$$

where μ_a is the expected reward for arm a.

— **Arm Selection :** Randomly select the next arm a_{t+1} with probability proportional to $w_{a,t}$.
Optimality probabilities $w_{a,t}$ are often computed via Monte Carlo methods, such as :

$$w_{a,t} \approx \frac{1}{G} \sum_{g=1}^{G} \mathbb{I}(\mu_a^{(g)} = \max(\mu_1^{(g)}, \dots, \mu_k^{(g)})),$$

where $\mu_a^{(g)}$ are sampled from $p(\theta|y_t)$.

III Experiment 1 : Binomial bandits

The **binomial bandit** is a specific version of the multi-armed bandit problem where each arm produces binary rewards (success or failure, typically modeled using a Bernoulli distribution). The success probabilities of each arm $(\theta_1, \dots, \theta_k)$ are unknown and must be learned over time.

III.1 Derivation of the solution

Computations

We consider the parameters $\theta = (\theta_1, \dots, \theta_K)$, where each $\theta_a \in [0, 1]$ represents the probability of success for arm $a \in \{1, \dots, K\}$. The model assumptions are as follows :
— For simplicity, the parameters θ_a are independant accross a
— Observations y_t up to time t are conditionally independent given θ
— For each arm a at time τ, the rewards are distributed from a Bernoulli, i.e. $y_{a\tau} \sim \text{Bernoulli}(\theta_a)$
We define the following summary statistics :
— $Y_{a,t} := \sum_{\tau\,:\,a_\tau = a} y_\tau$, the number of successes observed for arm a up to time t,
— $N_{a,t} := \sum_{\tau\,:\,a_\tau = a} 1$, the total number of trials (successes + failures) for arm a up to time t.
Thus, the observations contained in y_t are summarized by the vectors $\mathbf{Y}_t = (Y_{1,t}, \dots, Y_{K,t})$ and $\mathbf{N}_t = (N_{1,t}, \dots, N_{K,t})$.

Link between posterior of θ and posterior of θ_a

We have :

$$p(\theta \mid \mathbf{y}_t) = \prod_{a=1}^{K} p(\theta_a \mid \mathbf{y}_t) = \prod_{a=1}^{K} p(\theta_a \mid \mathbf{Y}_t, \mathbf{N}_t) = \prod_{a=1}^{K} p(\theta_a \mid \mathbf{Y}_{a,t}, \mathbf{N}_{a,t})$$

Posterior of θ_a

As said in the paper, the prior of θ_a is a uniform $\mathcal{U}(0, 1)$ (note that it corresponds to a Beta$(1, 1)$ which is a conjugate prior).

Moreover, $Y_{a,t}$ and $N_{a,t}$ respectively correspond to the number of successes and the number of trials for arm a up to time t. Therefore, $Y_{a,t}|\theta_a, N_{a,t} \sim \text{Bin}(N_{a,t}, \theta_a)$.

$$p(Y_{a,t} = y_{a,t}|\theta_a, N_{a,t}) = \binom{N_{a,t}}{y_{a,t}} \theta_{a,t}^{y_{a,t}} (1 - \theta_{a,t})^{N_{a,t} - y_{a,t}}$$

So by Bayes formula,

$$\begin{aligned} p(\theta_a \mid Y_{a,t} = y_{a,t}, N_{a,t} = n_{a,t}) &\propto p(Y_{a,t} = y_{a,t} \mid \theta_a, N_{a,t} = n_{a,t})\, p(\theta_a \mid N_{a,t} = n_{a,t}) \\ &\propto p(Y_{a,t} = y_{a,t} \mid \theta_a, N_{a,t} = n_{a,t})\, p(\theta_a) \\ &\propto \theta_{a,t}^{y_{a,t}} (1 - \theta_{a,t})^{n_{a,t} - y_{a,t}} 1_{\{\theta_a \in [0,1]\}} \end{aligned}$$

We recognize the form of a Beta and conclude that $\theta_a|Y_{a,t}, N_{a,t} \sim \text{Beta}(Y_{a,t} + 1, N_{a,t} - Y_{a,t} + 1)$.

Hence,

$$\boxed{p(\theta \mid \mathbf{y}_t) = \prod_{a=1}^{K} \text{Beta}(\theta_a \mid \mathbf{Y}_{at} + 1, \mathbf{N}_{at} - \mathbf{Y}_{at} + 1)}$$

Where $\text{Beta}(\theta_a \mid \alpha, \beta)$ denotes the density of the beta distribution for random variable θ with parameters α and β. This expression corresponds to Equation (10) in the paper.

Probability that arm a is optimal at time t

Next, we need to calculate the probability that arm a is optimal, i.e., the probability that θ_a is the largest among all θ_j for $j = 1, \ldots, k$. This probability is given by :

$$\begin{aligned} w_{at} &= \Pr(\theta_a = \max\{\theta_1, \ldots, \theta_k\} \mid \mathbf{y}_t) \\ &= \int_0^1 p(\theta_a \mid \mathbf{y}_t) \Pr\left(\bigcap_{j \neq a} \theta_j < \theta_a \mid \mathbf{y}_t\right) d\theta_a \\ &= \boxed{\int_0^1 \text{Beta}(\theta_a \mid Y_{at} + 1, N_{at} - Y_{at} + 1) \prod_{j \neq a} \Pr(\theta_j < \theta_a \mid Y_{jt}, N_{jt})\, d\theta_a} \end{aligned}$$

using previous results for the posterior distribution of θ_a and the fact that $\theta_1, \ldots, \theta_k$ are independent.

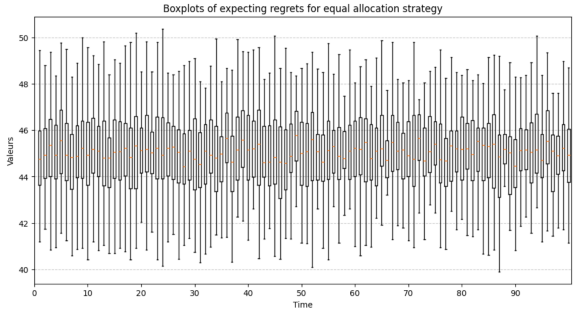
This integral can be computed either by numerical quadrature or by Monte Carlo simulation. This equation represents the probability that arm a is optimal after observing the data up to time t.

III.2 Results

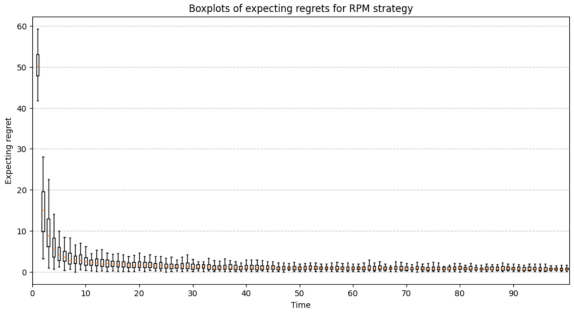
The code and results for this section can be found in the notebook `Binomial_Bandits.ipynb`, available here.

In figure 1, each boxplot in the figure represents the distribution of L_t for experiments still active at time t. At $t = 1$, the regret distributions are identical for both strategies. However, with randomized probability matching, the expected regret rapidly approaches zero as sub-optimal arms are identified and excluded. In contrast, the equal allocation strategy continues to assign observations to sub-optimal arms throughout the experiment. While this allows for more accurate estimation of θ for the sub-optimal arms, it comes at the expense of a significantly higher regret.

Then, figure 2 illustrates how the estimation of the success probability converges towards the true success probability over time. Each subplot corresponds to a specific arm, with the blue dashed line representing the true success probability and the red line showing the estimated probability as the algorithm progresses. Note that the convergence is much better for arms with higher success probabilities, as the algorithm spends more time exploring these "good" arms. Conversely, the convergence is less accurate for arms with lower success probabilities, as the algorithm quickly identifies them as sub-optimal and dedicates minimal exploration to these "bad" arms.



((a)) Equal allocation



((b)) RPM

FIGURE 1 – Expected regret per time period

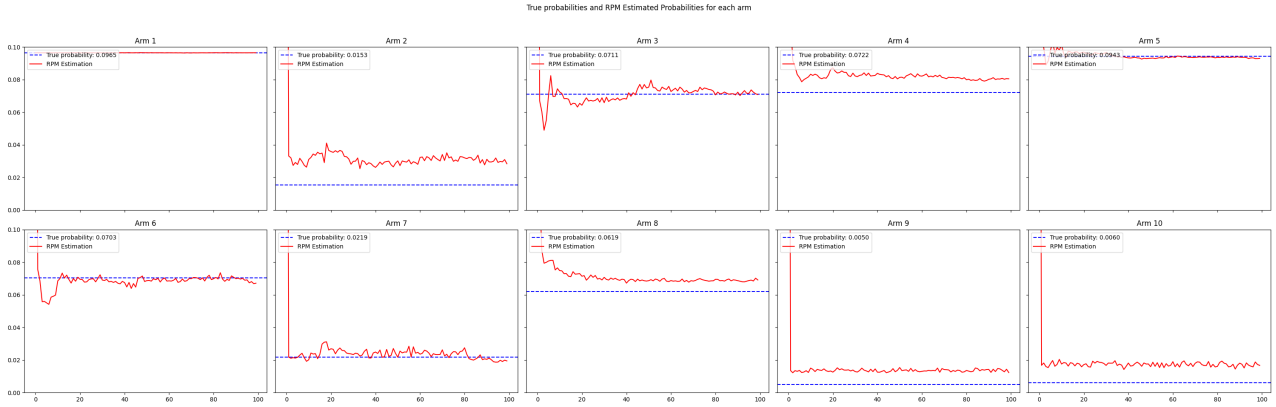


FIGURE 2 – Convergence of estimated probabilities of success

IV Experiment 2 : Linear bandits

We now focus on a different multi-armed bandit problem : **linear bandits**. Unlike the classification problem they treated in section 5 in the original paper (rewards belong to {0,1}) where they used a probit model, we address a regression problem (rewards belong to $\mathbb{R}$). While this setting shares many similarities with the fractional factorial bandit, it differs as y_t can take values beyond $\{0, 1\}$. In this section, we compute the posterior of θ in linear bandits and evaluate the performance of RPM in this new context.

IV.1 Derivation of the solution

While in the probit regression model presented in the paper, y_t was such that :

$$Pr(y_t = 1) = \Phi(\theta^T x_t)$$

We now consider a linear bandit where y_t is such that :

$$y_t = \theta^T x_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

In both case the prior on θ is Gaussian : $\theta \sim \mathcal{N}(\mathbf{0}, I)$ and the prior density is : $P(\theta) \propto \exp\left(-\frac{1}{2}\theta^\top I\theta\right)$.

$$\begin{aligned} P(y_1, \dots, y_t \mid \theta) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{s=1}^{t}(y_s - x_s^\top\theta)^2\right). \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{s=1}^{t} y_s^2 - 2\sum_{s=1}^{t} y_s x_s^\top\theta + \theta^\top\left(\sum_{s=1}^{t} x_s x_s^\top\right)\theta\right]\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\left[\theta^\top(\mathbf{X}_t^T\mathbf{X}_t)\theta - 2(\mathbf{X}_t^T\mathbf{y_t})^\top\theta\right]\right) \end{aligned}$$

With $\mathbf{X}_t$ as defined in the paper i.e. with x_t^T in row t, so $(\mathbf{X}_t^T\mathbf{X}_t) = \sum_{s=1}^{t} x_s x_s^\top$ and $\mathbf{X}_t^T\mathbf{y_t} = (\sum_{s=1}^{n} x_s y_s)^T = \sum_{s=1}^{n} y_s x_s^T$. Using Bayes' rule :

$$\begin{aligned} P(\theta \mid \{x_s, y_s\}_{s=1}^{t}) &\propto P(y_1, \dots, y_t \mid \theta)P(\theta) \\ &\propto \exp\left(-\frac{1}{2}\left[\theta^\top\left(\frac{1}{\sigma^2}(\mathbf{X}_t^T\mathbf{X}_t) + I\right)\theta - 2\frac{1}{\sigma^2}(\mathbf{X}_t^T\mathbf{y_t})^\top\theta\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\theta^\top\mathbf{\Omega}^{-1}\theta - 2\frac{1}{\sigma^2}(\mathbf{X}_t^T\mathbf{y_t})^\top\theta\right]\right) \text{ with } \mathbf{\Omega}^{-1} = \frac{1}{\sigma^2}(\mathbf{X}_t^T\mathbf{X}_t) + I, \\ &\propto \exp\left(-\frac{1}{2}\left[(\theta - \tilde{\theta}_t)^\top\mathbf{\Omega}^{-1}(\theta - \tilde{\theta}_t) - \tilde{\theta}_t^\top\mathbf{\Omega}^{-1}\tilde{\theta}_t\right]\right) \text{ where } \tilde{\theta}_t = \frac{1}{\sigma^2}\mathbf{\Omega}(\mathbf{X}_t^T\mathbf{y_t}) \end{aligned}$$

Hence, the posterior distribution at time t is :

$$\boxed{\theta \mid \{x_s, y_s\}_{s=1}^{t} \sim \mathcal{N}(\tilde{\theta}_t, \mathbf{\Omega})}$$

where $\tilde{\theta}_t = \frac{1}{\sigma^2}\mathbf{\Omega}(\mathbf{X}_t^T \mathbf{y_t})$, $\mathbf{\Omega}^{-1} = \frac{1}{\sigma^2}(I + \mathbf{X}_t^T \mathbf{X}_t)$.

Note the similarity with the result for the probit posterior of the paper with $\Sigma = I$ and $b = 0$ (Appendix, section A.1). Here we don't need the variable z as we work directly with a linear bandit, which simplifies the algorithm which is now :

In each round $t + 1$,

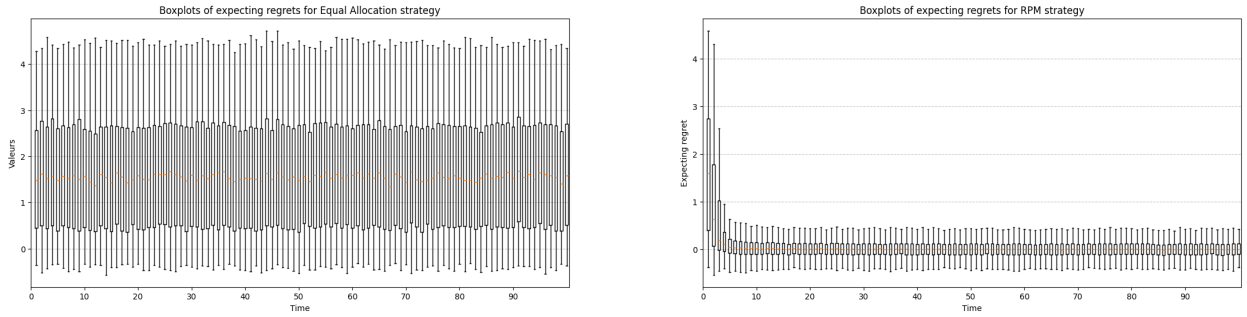
$$\text{Sample } \hat{\theta}_t \sim \mathcal{N}\left(\tilde{\theta}_t, \mathbf{\Omega}\right)$$

$$x_{t+1} = \arg \max_{x \in \mathcal{X}_{t+1}} x^\top \hat{\theta}_t$$

IV.2 Results

The code and results for this section can be found in the notebook `Linear_Bandits.ipynb`, available here.

In figure 3, we consider a linear bandit environment with $K = 7$ (number of possible choices/actions), $\theta = [1.22, 1.25, 1.3]$ (but other values of θ work well), and a noise standard deviation $\sigma = 0.4$. Each boxplot in the figure represents the distribution of L_t. At $t = 1$, the regret distributions are identical for both strategies, as both explore all actions equally at the start. However, with Randomized Probability Matching (RPM), the expected regret rapidly decreases as sub-optimal actions are identified and excluded. By leveraging the linear structure of the environment and sampling from the posterior distribution of θ, RPM focuses more effectively on the optimal action. In contrast, the equal allocation strategy continues to assign observations uniformly across all actions, including sub-optimal ones, throughout the experiment. The effect of the noise ($\sigma = 0.5$) is also mitigated more effectively by RPM. Note that RPM works well in general but the convergence can be sensitive to the parameters (σ, θ). Intuitively, the algorithm achieves better convergence when the values of the vector θ are far apart and when the variances of the rewards are low. That said, RPM performs well in large action space, balancing exploration and exploitation effectively with adaptive posterior sampling while remaining computationally scalable.



((a)) Equal allocation ((b)) RPM

FIGURE 3 – Expected regret per time period

V Conclusion

This study examines the performance of Bayesian methods for solving multi-armed bandit problems, focusing in particular on the Randomized Probability Matching (RPM) algorithm. This algorithm, now better known as Thompson Sampling, is characterized by its ability to efficiently handle the exploration-exploitation tradeoff and has established itself as a powerful and widely used Bayesian solution. Notabaly, it has good convergence results, accurately aligning success probability estimates with true values over time, with faster convergence for higher-performing arms and efficient (non)exploration of suboptimal ones. RPM ability to combine fast learning, regret minimization, and convergence to true parameters solidifies its position as one of the most prominent and effective approaches in the multi-armed bandit field.