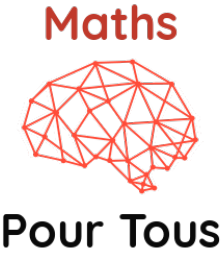
Akinator, entropie et information

Nicolas Masson

11 décembre 2017



Séminaire "Maths Pour Tous"

Table des matières

Introduction

Enoncé du problème

Mettez-vous un instant dans la peau du génie du fameux jeu Akinator. Une personne devant vous pense à un personnage quelconque, et vous devez lui poser une succession de questions auxquelles elle devra répondre par oui ou par non, et dont les réponses vous permettront de deviner à qui votre interlocuteur pense. Bien évidemment, vous n'allez pas faire une liste de tous les personnages possibles et les tester un à un, à moins d'avoir un temps infini devant vous. Vous n'avez tout de même pas envie d'y passer la nuit ! Vous voulez même mieux : vous désirez être suffisamment malin pour deviner l'identité du personnage mystère en un temps record. La question naturelle qui se pose donc est : quelle stratégie adoptez-vous ? Comment construisez-vous intelligemment votre questionnaire pour trouver la réponse le plus rapidement possible ? L'objectif de cet exposé sera de proposer des réponses à ces interrogations.

Premières remarques

Comme tout problème mathématique difficile, il est nécessaire de passer par une phase de tâtonnement, de recherche primitive, où l'on essaie de déceler ce qui nous est directement accessible. Ici, par exemple, il paraît logique dans un premier temps de se demander quelles questions on poserait naturellement au début pour aller au plus vite. On pense assez logiquement à des questions comme : "le personnage est-il réel ou fictif ?", ou "est-il mort ou vivant ?", ou encore "est-ce un homme ou une femme ?". Mais pourquoi ces questions nous semblent intuitivement comme étant de "bonnes questions" ? Comment fait-on la différence entre une bonne et une mauvaise question ? Supposons que l'on pose comme première question : "est-ce Ryan Gosling ?". Bien que le personnage ne soit pas hautement improbable, il y a tout de même globalement peu de chances *a priori* - sans information préalable - pour que ce soit lui. La réponse a donc de fortes chances d'être négative, n'apportant dans ce cas que très peu d'informations. Si on y réfléchit bien, ce qui fait que cette question est mauvaise est que l'incertitude quant à la réponse à cette question est très faible : on se doute bien que ce sera non, et on ne sera pas plus avancés qu'avant. La bonne méthode semble donc être de poser des questions dont on a aucune raison de privilégier une réponse affirmative ou

négative, comme par exemple lorsque l'on demande si le personnage est réel ou fictif. La remarque précédente semble donc suggérer un procédé pour construire un questionnaire optimal. Cependant, cela ne peut pas encore être l'objet d'une démonstration, dans la mesure où les termes sont trop flous pour qu'on puisse constituer un théorème. En effet, on ne s'est pas encore donné de cadre abstrait au sein duquel notre procédé, ainsi que les notions de questionnaire et d'optimalité, ont un sens précis. De plus, comme l'optimalité - encore une fois, en un sens à préciser - de la méthode proposée semble reposer sur le fait que l'incertitude quant aux réponses aux questions posées est totale, maximale, il paraît logique et utile de chercher également à donner un sens précis à cette notion d'incertitude.

Programme pour la suite

Les remarques faites précédemment nous suggèrent naturellement le programme suivant :

1. Donner un sens précis :
   (a) à la notion d'information
   (b) à la notion de questionnaire
   (c) à la notion d'optimalité d'un questionnaire
   (d) au procédé proposé
2. Étudier l'optimalité - au sens défini précédemment - du questionnaire construit à l'aide du procédé proposé

Ce petit cahier des charges sera notre guide tout au long de l'exposé, et notre objectif sera d'en respecter tous les termes.

1 La notion d'incertitude, a.k.a "entropie"

1.1 Eléments de théorie des probabilités

Avant de s'intéresser à la notion d'incertitude, il apparaît comme nécessaire de faire un petit point sur quelques notions élémentaires de probabilités. La théorie des probabilités est la théorie utilisée pour décrire les phénomènes aléatoires, ou en d'autres termes les phénomènes dont on ne peut prédire l'issue. Pour ce faire, on se donne un ensemble, noté classiquement Ω, représentant l'ensemble des issues possibles de notre expérience aléatoire.

Exemple 1.1. Pour un lancer de dé, l'ensemble des issues possibles est naturellement représenté par $\Omega = \{1, 2, 3, 4, 5, 6\}$

Souvent, l'étude que l'on cherche à faire ne concerne pas seulement une issue en particulier mais un groupement de plusieurs issues, d'où la définition suivante

Définition 1.1.1. On appelle événement tout sous-ensemble de Ω

Exemple 1.2. Pour le lancer de dé, on peut par exemple s'intéresser à l'événement : "le nombre obtenu est pair", représenté par l'ensemble $\{2, 4, 6\}$.

Bien évidemment, on ne veut pas se contenter de savoir que l'issue d'une expérience est incertaine. On veut une description plus précise, plus quantitative, ce pourquoi on va associer à chaque issue possible ω une probabilité p_ω, c'est-à-dire un nombre entre 0 et 1, représentant le degré de vraisemblance de la réalisation de cette issue. Le fait d'imposer que ces nombres soient compris entre 0 et 1 est arbitraire et consiste simplement à se donner une référence : 0 représente l'impossibilité et 1 la certitude qu'une issue soit réalisée.

Définition 1.1.2. Soit $\Omega = \{\omega_1, ..., \omega_n\}$ un ensemble fini. Pour ω dans Ω, soit p_ω un nombre compris entre 0 et 1. La collection de nombres $(p_\omega)_{\omega\in\Omega}$ est une distribution de probabilité sur Ω si

$$p_{\omega_1} + p_{\omega_2} + ... + p_{\omega_n} = 1$$

Exemple 1.3. Toujours dans le cas du lancer de dé, si on considère que celui-ci n'est pas truqué, la distribution de probabilité la plus naturelle est celle dite "uniforme", au sens où elle ne privilégie aucune issue : $p_1 = p_2 = ... = p_6 = \frac{1}{6}$

Ayant défini la notion d'événement, on définit de manière tout aussi naturelle la probabilité qu'un événement se réalise :

Définition 1.1.3. Soit A un événement de Ω. On définit la probabilité p_A que l'événement se réalise comme la somme des probabilités des issues constituant A, soit, en écriture mathématique :

$$p_A = \sum_{\omega\in A} p_\omega$$

Exemple 1.4. Toujours dans le cas du lancé de dé, décrit par l'univers $\{1, 2, ..., 6\}$ et la distribution uniforme, la probabilité de l'événement A : "le nombre obtenu est pair" est :

$$p_A = p_2 + p_4 + p_6 = 3 \times \frac{1}{6} = \frac{1}{2}$$

Une notion importante en probabilités est la notion d'indépendance, qui traduit l'absence d'influence d'une expérience sur une autre. Soient donc deux expériences aléatoires, décrites respectivement par les univers Ω_1, Ω_2 et les distributions $(p_{w_1})_{w_1\in\Omega_1}$, $(p'_{\omega_2})_{w_2\in\Omega_2}$. Considérées simultanément, ces deux expériences peuvent être vues comme une seule expérience, avec comme issues possibles tous les couples (ω_1, ω_2) avec $\omega_1 \in \Omega_1$, $\omega_2 \in \Omega_2$. Ce nouvel univers -dit "produit"- Ω sera noté $\Omega_1 \times \Omega_2$. Mais quelle serait la bonne distribution de probabilité à mettre sur Ω ? La question revient à savoir, pour $\omega_1 \in \Omega_1$, $\omega_2 \in \Omega_2$ donnés, comment définir la probabilité que ω_1 et ω_2 se réalisent simultanément, dans le cas où le résultat d'une expérience n'influe pas sur le résultat de l'autre. Il est alors logique de traduire cette indépendance par par le fait que cette probabilité est $p_{\omega_1} \times p'_{\omega_2}$. Ainsi, on définit l'indépendance de la manière suivante :

Définition 1.1.4. Soient $(\Omega_1, (p_{w_1})_{w_1\in\Omega_1})$, $(\Omega_2, (p'_{w_2})_{w_2\in\Omega_2})$ deux expériences aléatoires. On dira ces deux expériences indépendantes si l'univers produit $\Omega = \Omega_1 \times \Omega_2$ est muni de la "loi produit"

$$p \otimes p'_{(\omega_1,\omega_2)} = p_{\omega_1} p'_{\omega_2}$$

Exemple 1.5. Supposons que l'on cherche à modéliser deux lancers de pièces équilibrés indépendants. Chaque lancer est une expérience aléatoire, avec pour univers $\Omega_1 = \Omega_2 = \{P, F\}$ - P pour "pile", F pour "face", et pour chaque issue une probabilité 1/2. L'ensemble des issues possibles pour deux lancers successifs est donc :

$$\Omega = \{(P, P), (F, P), (P, F), (F, F)\}$$

Selon la définition précédente, l'hypothèse d'indépendance des deux lancers se traduit par une distribution produit, qui ici n'est autre que la probabilité uniforme - chaque issue a une chance sur quatre de se produire.

Au-delà des définitions mathématiques, il faut bien avoir à l'esprit que la notion de probabilité est intimement liée à l'information dont on dispose sur notre expérience. Ainsi, imaginons qu'une personne devant vous lance un dé dans une boîte, et que lui que lui seul puisse voir le résultat. *A priori*, sans information supplémentaire, la distribution uniforme s'impose évidemment comme la plus pertinente. Cependant, si maintenant la personne en question vous souffle à l'oreille que le résultat est pair, la distribution uniforme devient caduque : il serait idiot de continuer à considérer que le nombre 3, étant impair, possède une chance sur six d'être tiré. Il est donc logique d'essayer de définir comment faire évoluer notre distribution de probabilité en cas d'apport d'information, ce qui peut être fait grâce à la notion de probabilité conditionnelle :

Définition 1.1.5. Soit Ω un ensemble, $(p_\omega)_{\omega\in\Omega}$ une distribution de probabilité sur Ω, et A un événement de probabilité non-nulle. On définit la distribution de probabilité "sachant A" $(p_\omega^{(A)})_{\omega\in\Omega}$ par :

$$p_\omega^{(A)} = \begin{cases} \frac{p_\omega}{p_A} & \text{si } \omega \in A \\ 0 & \text{sinon} \end{cases}$$

Exemple 1.6. Toujours dans le cas du lancer de dé, la distribution de probabilité sachant l'événement "le nombre est pair", toujours noté A, est $p_1^{(A)} = p_3^{(A)} = p_5^{(A)} = 0$, $p_2^{(A)} = p_4^{(A)} = p_6^{(A)} = \frac{1}{3}$

1.2 Petite parenthèse sur le logarithme

Nous allons devoir faire une petite parenthèse technique sur une fonction bien connue des mathématiciens : le logarithme. Cette parenthèse non motivée en apparence est malheureusement nécessaire dans la mesure où le logarithme apparaît dans la suite de l'exposé. Mais que le lecteur se rassure : il y a finalement peu de choses à savoir sur le logarithme pour que son utilisation soit naturelle lorsque nous chercherons à résoudre notre problème. Pour le lecteur érudit, le logarithme dont nous allons nous servir est le logarithme en base 2, nous nous contenterons donc de ne présenter que celui-ci, et le noterons simplement log et non $\log_2$ pour plus de lisibilité.
Le logarithme, noté log, est une fonction qui à un nombre réel strictement positif associe un nombre réel (en vocabulaire mathématique, log va de $\mathbb{R}_+^*$ dans $\mathbb{R}$). Si le lecteur est curieux de voir à quoi ressemble le logarithme, son graphe est en figure 6, en annexe. Mais pour notre exposé, il n'est pas tant nécessaire de connaître le graphe du logarithme, mais plutôt les propriétés suivantes, que nous appellerons "propriétés fondamentales du logarithme" :

Proposition 1.2.1 (Propriétés fondamentales du logarithme)**.** *Soient a, b deux réels strictement positifs. Alors :*

1. $\log(a \times b) = \log(a) + \log(b)$
2. $\log(2) = 1$
3. $\log(1) = 0$
4. $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

La propriété 3 est une conséquence de la propriété 1, et 4 est une conséquence de 1 et de 3. En somme, la seule propriété à retenir est que le logarithme **transforme les multiplications en additions**. Dans la suite, on se servira du corollaire évident de 1 et 2 :

Proposition 1.2.2. *Pour tout entier n :*

$$\log(2^n) = n$$

Dans la suite de l'exposé, on aura également besoin de l'inégalité suivante, dite de "convexité" - évidente sur la figure 7 en annexe.

Proposition 1.2.3 (Inégalité de convexité)**.** *Il existe $C > 0$ tel que, pour tout $x > 0$, on a :*

$$\log(x) \leq C(x-1)$$

1.3 L'entropie

Revenons à notre problème de départ. Comme remarqué précédemment, il serait souhaitable d'avoir à disposition un outil donnant un sens quantitatif à la notion d'incertitude, de manque d'information. On considère une expérience aléatoire, décrite par un univers Ω et une distribution de probabilité $(p_\omega)_{\omega\in\Omega}$ sur Ω. Notre objectif est donc de définir une fonction - que l'on notera H par la suite - qui à une distribution de probabilité associe l'incertitude quant au résultat de l'expérience. L'entreprise n'a pas l'air aisée, du moins au premier abord. Il y a en effet beaucoup de distributions de probabilité que l'on peut définir pour un même univers, on commencera donc par se demander comment doit se comporter H vis-à-vis des distributions les plus simples : les distributions uniformes. Une distribution uniforme n'étant déterminée que par le nombre d'issues possibles de l'expérience, on est en droit de se poser la question suivante : y a-t-il plus d'incertitude à lancer un dé ou une pièce de monnaie ? En posant cette question, on se rend compte que plus le nombre d'issues possibles est grand, plus l'incertitude doit être grande, ce pourquoi on imposera comme premier axiome :

$$\text{Pour tout entier } n, \quad H\left(\frac{1}{n+1},...,\frac{1}{n+1}\right) \geq H\left(\frac{1}{n},...,\frac{1}{n}\right) \tag{$\mathcal{A}_1$}$$

Considérons désormais deux expériences aléatoires indépendantes, décrites respectivement par les univers Ω, Ω', munis respectivement des distributions p et p'. Essayons de traduire l'indépendance de ces deux expériences en terme d'incertitude. Sachant que la connaissance du résultat de la première expérience ne renseigne en rien quant au résultat de la seconde, il paraît naturel de dire que l'incertitude de l'expérience produit, ôtée de l'incertitude de la première expérience, est égale à l'incertitude de la seconde. On imposera donc en deuxième axiome :

$$\text{pour toutes distibutions } p, p'; \quad H(p\otimes p') - H(p) = H(p') \tag{$\mathcal{A}_2$}$$

Désormais, essayons de voir comment doit se comporter notre incertitude lors d'un apport d'information. En d'autres termes, conformément aux définitions données dans la partie précédente, l'objectif est donner une relation entre l'incertitude d'une distribution conditionnelle et celle de sa distribution originelle. Soit donc un univers Ω, p une distribution sur Ω, et A un événement de probabilité différente de 0 et de 1. Notons $\overline{A}$ l'événement complémentaire de A, c'est-à-dire l'événement "A n'est pas réalisé". Il découle directement des définitions 1.1.2 et 1.1.3 que $p_{\overline{A}} = 1 - p_A > 0$ car $p_A < 1$, ce qui nous permettra de définir les distributions conditionnelles sachant A et sachant $\overline{A}$. Il est assez logique de considérer que l'incertitude quant au résultat de l'expérience se décompose en la somme de l'incertitude relative à la réalisation de A plus la moyenne entre

les incertitudes conditionnelles sachant A et sachant $\overline{A}$. En des termes plus mathématiques, on imposera donc :

$$H(p) = H\left(p_A, 1-p_A\right) + p_A H\left(p^{(A)}\right) + \left(1-p_A\right) H\left(p^{(\overline{A})}\right) \qquad (\mathcal{A}_3)$$

Ces trois axiomes ne suffisent malheureusement pas à définir un objet satisfaisant, dans la mesure où ils n'imposent aucune "régularité" sur H. En des termes plus simples, des fonctions "monstrueuses", impossibles à calculer, vérifient ces trois conditions. On imposera donc une dernière condition, un peu moins naturelle et plus technique, mais tout de même assez compréhensible :

$$q \mapsto H(q, 1-q) \quad \text{continue} \qquad (\mathcal{A}_4)$$

ce qui pourrait se traduire - pour le lecteur ne connaissant pas la notion de continuité - par : deux distributions proches ont des incertitudes proches.

Il s'avère que ces quatre conditions ne laissent finalement plus beaucoup de possibilités pour H :

Théorème 1.3.1. *Les seules fonctions vérifiant* $(\mathcal{A}_1)$, $(\mathcal{A}_2)$, $(\mathcal{A}_3)$ *et* $(\mathcal{A}_4)$ *sont les :*

$$p = (p_1, ..., p_n) \longmapsto -C(p_1 \log p_1 + p_2 \log p_2 + ... + p_n \log p_n) = -C \sum_{i=1}^{n} p_i \log p_i$$

où C est une constante positive.

Remarque. Dans le théorème précédent, on prend pour convention que $0 \log 0 = 0$. Par la suite, on n'étudiera que des distributions (p_i) censées modéliser le degré de vraisemblance d'apparition d'un personnage. Sachant qu'il n'est pas très pertinent de prendre en compte des personnages auxquels il est impossible de penser, on ne considèrera par la suite que des distributions de probabilité dont tous les termes sont strictement positifs.

Toutes ces fonctions ne diffèrent que d'une constante multiplicative, elles ont donc des comportements identiques. Par la suite, on ne se servira plus que de celle correspondant à $C = 1$, pour des raisons qui s'éclairciront par la suite.

Définition 1.3.1. On appellera entropie la fonction :

$$p = (p_1, ..., p_n) \longmapsto -\sum_{i=1}^{n} p_i \log p_i$$

que l'on notera H_2

Une démonstration de ce résultat peut être trouvée dans [2].

2 Formalisation du problème

2.1 Qu'est-ce qu'un questionnaire ?

Poursuivons le programme établi dans l'introduction. Après avoir donné un sens à la notion d'incertitude, revenons au problème initial et demandons nous comment formaliser ce qu'est un questionnaire. Tout d'abord, on remarque que l'ensemble des personnages auxquels peut pense un

humain est fini. Notons donc abstraitement N le nombre de personnages possibles, et numérotons lesdits personnages de 1 à N. Ainsi, l'ensemble des personnages auxquels on peut penser s'identifie naturellement à l'ensemble $\{1, ..., N\}$. En première approche, on peut se représenter un questionnaire par un arbre binaire, ainsi qu'un ensemble de N noeuds marquant la fin du questionnaire, comme sur la figure 4 :

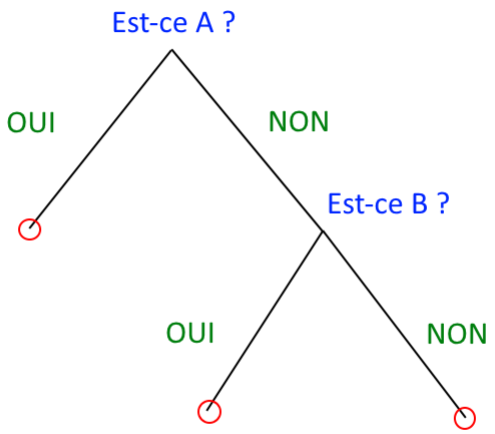


FIGURE 1 – Exemple de questionnaire abstrait

Mais avec la seule donnée d'un arbre binaire et de N noeuds, on n'est pas forcément en mesure de construire un questionnaire correspondant ; comme le montre l'exemple ci-dessous en figures 2 et 3, on aurait un problème si l'un des noeuds est parent d'un autre. En effet, si le noeud n_j est parent du noeud n_i, on aboutirait à un erreur, car si le personnage i est choisi, le questionnaire aboutirait à la réponse j. Il est donc logique de modéliser un questionnaire par la donnée d'un arbre binaire et d'un ensemble de N noeuds terminaux dont aucun n'est parent d'un autre.

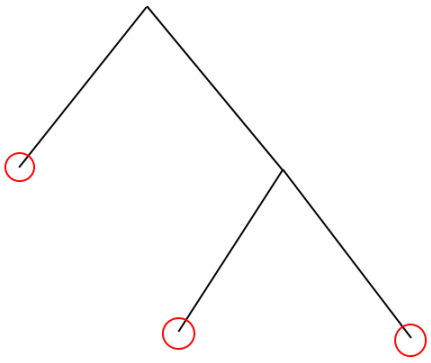


FIGURE 2 – Bon choix de noeuds

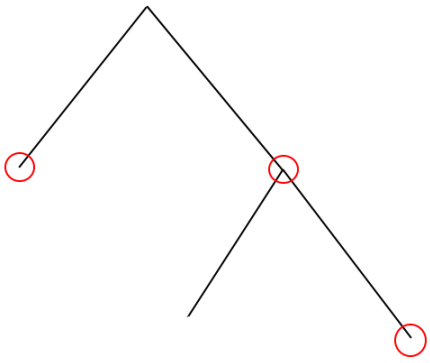


FIGURE 3 – Mauvais choix de noeuds

Cependant, comme le vocabulaire n'est pas forcément familier du lecteur, et que cette définition n'est pas très pratique, nous allons procéder légèrement différemment. Pour un questionnaire donné,

on va associer à chaque personnage un nom de code constitué de 0 et de 1, correspondant à la suite de réponses données pour aboutir au personnage considéré - où 0 représente une réponse négative, 1 une réponse positive. Par exemple, dans la figure 4, B aurait le code 01. On sent que ce point de vue sera plus fécond, dans la mesure où a notion de code paraît plus maniable. Mais avant de continuer, précisons ce que l'on entend par code :

Définition 2.1.1. Si $\mathcal{A}$ est un ensemble, on note $\mathcal{A}^*$ l'ensemble des mots que l'on peut former avec pour caractères les éléments de $\mathcal{A}$. On appelle alors $\mathcal{A}$ un alphabet.

Exemple 2.1. Si $\mathcal{A} = \{a, b, c, ..., z\}$, $\mathcal{A}^*$ est l'ensemble des mots, entendu au sens courant - en dehors de toute notion sémantique ; les mots au sens ici considéré peuvent ne rien vouloir dire.

Définition 2.1.2. Si $\mathcal{A}$ est un alphabet, χ un ensemble, on appelle un code de χ sur l'alphabet $\mathcal{A}$ toute fonction $c : \chi \mapsto \mathcal{A}^*$

Ainsi, conformément au vocabulaire introduit, on peut coder chacun des personnages par un code de $\{1, ..., N\}$ sur l'alphabet $\{0, 1\}$ pour un questionnaire donné. Il nous reste à traduire ce que signifie : "aucun noeud terminal n'est parent d'un autre" en terme de code. En regardant la figure ci-dessous, c'est assez clair : on voit que cette propriété correspond au fait qu'aucun nom de code n'est préfixe d'un autre. Précisons ce que l'on n'entend par là :

Définition 2.1.3. Soit c un code de χ sur l'alphabet $\mathcal{A}$. On dit que c possède la propriété du préfixe - ou plus rapidement que c est préfixe - si pour tous $x \neq y$ dans χ, m dans $\mathcal{A}^*$, on a :

$$c(x) \neq c(y)m$$

Ainsi, on a vu que si on se donnait un arbre binaire et N noeuds terminaux dont aucun n'est le parent d'un autre, le code associé, construit selon le procédé décrit précédemment, possède la propriété du préfixe. Réciproquement, si on se donne c un code ayant la propriété du préfixe, il n'est pas difficile de voir que l'on peut construire un questionnaire, dont le code associé est c. Dorénavant, nous parlerons donc indifféremment de code préfixe sur $\{0, 1\}$ ou de questionnaire par la suite :

Définition 2.1.4. On appellera questionnaire tout code de $\{1, ..., N\}$ sur l'alphabet $\{0, 1\}$ ayant la propriété du préfixe.

2.2 L'optimalité

Nous avons désormais rempli les deux premiers termes de notre contrat, charge à nous de continuer notre entreprise en donnant un sens à la notion d'optimalité. Au premier abord, on aurait envie de dire qu'un questionnaire optimal est un questionnaire dont la longueur est minimale. Mais comment définir la longueur d'un questionnaire ? Selon le cas de figure, la série de questions que l'on pose peut être plus ou moins longue. L'idée la plus naturelle serait donc d'envisager systématiquement le pire des cas, et de définir la longueur d'un questionnaire comme celle de la plus longue série de questions possible, c'est-à-dire, conformément aux définitions de la partie précédente, comme la longueur du plus long nom de code. Le code optimal serait donc le code "le moins pire", parmi les pires cas de figures. Mais définir l'optimalité de la sorte serait nier complètement le caractère fondamentalement probabiliste de l'expérience. En effet, votre interlocuteur a plus de chances *a priori* de penser à Bach qu'à Penderecki (sauf si vous le connaissez bien et que vous le savez passionné de musique sérielle polonaise). Imaginons la situation suivante : on restreint

l'ensemble des possibles à trois personnes : une *superstar*, et deux inconnus. Comparons les deux questionnaires suivants : le premier où l'on demande d'abord si c'est la *superstar*, et ensuite, en cas de réponse négative, si c'est l'inconnu numéro 1 ; le deuxième où l'on demande d'abord si c'est l'inconnu, puis si c'est la *superstar*.

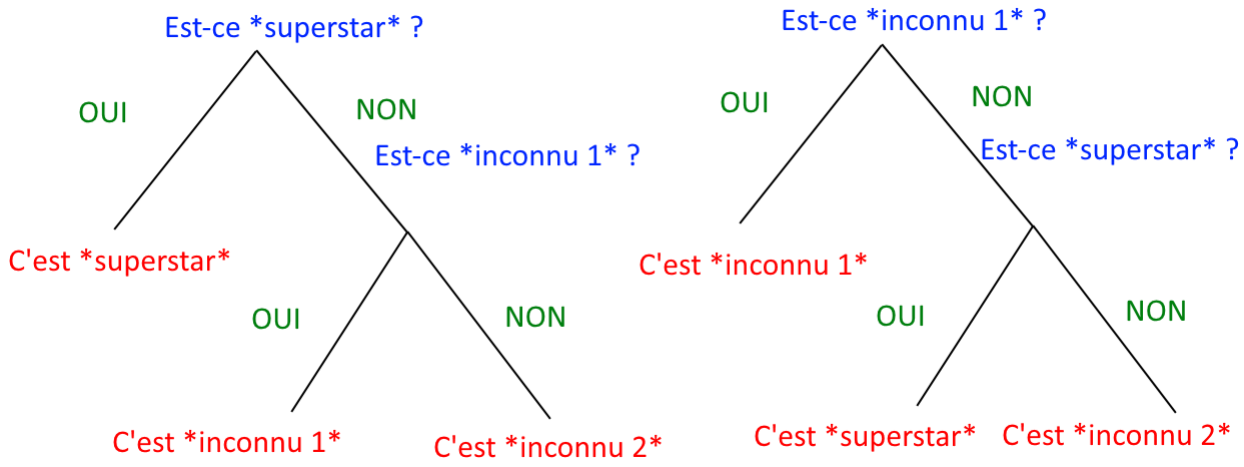


FIGURE 4 – Comparaison de deux méthodes

Selon notre définition de l'optimalité, nos deux questionnaires sont équivalents (deux questions au maximum). Pourtant, le premier semble intuitivement meilleur dans la mesure où il y a de fortes chances pour qu'on n'ait à poser qu'une seule question, contrairement au deuxième qui, selon toute vraisemblance, devra nécessiter deux questions.
On préférera donc une notion d'optimalité dite "en moyenne", au sens où si on fait un grand nombre de fois l'expérience, le nombre moyen de questions posées est minimal. Cette notion de "longueur moyenne" du questionnaire peut s'exprimer facilement à l'aide des probabilités. Notons donc p_i la probabilité que le personnage i soit choisi, et notons l_i le nombre de questions à poser pour trouver i. Si on fait un grand nombre de fois l'expérience, il est raisonnable d'interpréter p_i comme la proportion de fois où le personnage i a été choisi. Ainsi, on définira naturellement la longueur moyenne d'un questionnaire comme suit :

Définition 2.2.1. Soit c un questionnaire. Notons l_i la longueur du mot $c(i)$. On définit la longueur moyenne de c, notée $\mathbb{E}(c)$ (comme "espérance"), par :

$$\mathbb{E}(c) = p_1 l_1 + p_2 l_2 + ... + p_N l_N = \sum_{i=1}^{N} p_i l_i$$

Définition 2.2.2. On dira que c est optimal si pour tout autre questionnaire c',

$$\mathbb{E}(c) \leq \mathbb{E}(c')$$

3 Résolution du problème

Nous avons rempli notre objectif de formalisation : chaque terme du problème possède désormais un sens mathématique clair, permettant ainsi la formulation précise dudit problème.

Problème 3.1. Trouver un code préfixe de $\{1, ..., N\}$ sur l'alphabet $\{0, 1\}$ de longueur moyenne minimale

3.1 Borne inférieure de la longueur moyenne

Le problème énoncé n'a rien d'évident *a priori*, on tâchera donc tout d'abord de trouver une borne inférieure de la longueur moyenne, c'est à dire une constante K strictement positive telle que pour tout code c, $\mathbb{E}(c) \geq K$. On se doute bien qu'une telle constante doit exister. En effet, la propriété du préfixe semble être assez contraignante, et paraît forcer le code à être suffisamment long en moyenne. Essayons de voir plus précisément ce que contraint la propriété du préfixe sur la longueur des noms de code. Soit donc c un questionnaire. Supposons, quitte à réordonner, que les longueurs $l_1, ..., l_N$ sont rangées dans l'ordre croissant. Dessinons un arbre binaire de profondeur l_N, de sorte à pouvoir placer n'importe quel personnage sur cet arbre.

Remarque. Un arbre de profondeur n possède 2^n noeuds terminaux

Plaçons l'objet i sur l'arbre, et disons par exemple qu'il est au niveau du point bleu (cf figure 5). La propriété du préfixe interdit qu'un autre objet se situe sur un des noeuds de la zone rouge, et donc entre autres, sur un des noeuds terminaux dans cette zone (points rouges). Or l'arbre étant défini comme de longueur l_N, l'objet N se situe sur un des noeuds terminaux de l'arbre. Par la remarque précédente, il lui est interdit de se trouver sur un des $2^{l_N - l_1}$ noeuds rouges.

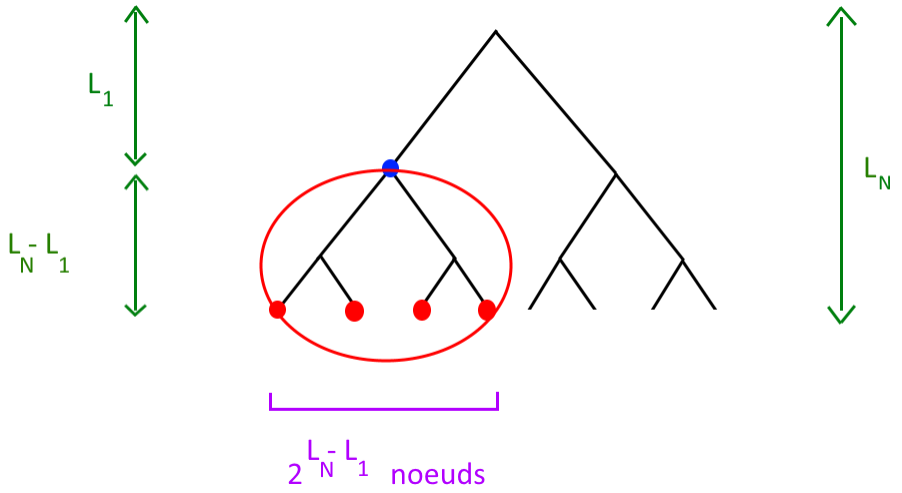


FIGURE 5 – Inégalité de Kraft

Mais en appliquant le même raisonnement, la position de l'objet 2, $2^{l_N - l_2}$ noeuds supplémentaires sont également interdits à N. Donc en prenant en compte toutes les contraintes imposées par la position des objets 1 à $N-1$ et par la propriété du préfixe, ce sont $\sum_{i=1}^{N-1} 2^{l_N - l_i}$ noeuds "interdits" à l'objet N. Mais il faut bien qu'il reste de la place pour pouvoir placer N. Or il y a exactement 2^{l_N} noeuds terminaux. Il faut donc que

$$\sum_{i=1}^{N} 2^{l_N - l_i} \leq 2^{l_N} - 1$$

En ajoutant 1 et en divisant par 2^{l_N} de chaque côté, on arrive finalement à l'inégalité suivante, appelée inégalité de Kraft :

Proposition 3.1.1. *Si c est un questionnaire, alors, en notant l_i la longueur de $c(i)$:*

$$\sum_{i=1}^{N} \frac{1}{2^{l_i}} \leq 1 \qquad \text{(Kr)}$$

En fait, notre étude précédente nous permet même d'avoir mieux : si on se donne une collection d'entiers $l_1, ..., l_N$ vérifiant (Kr), elle nous suggère un procédé pour construire un questionnaire c dont la longueur de $c(i)$ est précisément l_i. En effet, il suffit de se donner un arbre binaire de profondeur l_N (quitte à réordonner, on peut toujours supposer que $l_1 \leq ... \leq l_N$), et partitionner les noeuds terminaux en un premier groupe G_1 contenant les $2^{l_N - l_1}$ premiers noeuds, un deuxième G_2 contenant les $2^{l_N - l_1}$ suivants, etc. Cette construction est rendue possible grâce à l'inégalité de Kraft. Il suffit alors de placer l'objet i au niveau de la racine commune du groupe G_i, et il est alors assez clair -à condition de faire un dessin pour bien se représenter les choses - que notre méthode construit un code préfixe (pour plus de détails, voir [1]. Ainsi :

Proposition 3.1.2. *Si $l_1, ..., l_N$ sont des entiers vérifiant (Kr), alors il existe un questionnaire c tel que la longueur de $c(i)$ soit égale à l_i.*

On sent donc que l'on est sur la bonne voie : on a réussi à trouver ce qu'impliquait la propriété du préfixe sur les longueurs l_i. Reste à traduire ce que l'inégalité de Kraft implique sur la longueur moyenne d'un code préfixe, c'est-à-dire $\sum p_i l_i$. Mais l'inégalité de Kraft fournit une inégalité sur la somme des 2^{-l_i}, et on aimerait bien passer de 2^{l_i} à l_i directement. Or, si on se souvient bien de l'appendice en début d'exposé, un outil est tout trouvé : le logarithme, qui transforme les multiplications en additions (*cf* propriété 1.2.1). On peut donc écrire, si c est un questionnaire, l_i la longueur de $c(i)$:

$$\mathbb{E}(c) = \sum_{i=1}^{N} p_i l_i = -\sum_{i=1}^{N} p_i \log\left(2^{-l_i}\right) \qquad (*)$$

On sent bien, au vu de cette dernière expression, que l'entropie aura un rôle à jouer. Et si l'entropie, que l'on a construite pour représenter un manque d'information, était précisément la borne que l'on cherche ? Ce serait logique : y aurait-il de meilleure interprétation pour le manque d'information que le nombre minimal moyen de questions à poser pour découvrir le personnage mystère ? Comparons donc, pour c questionnaire, $\mathbb{E}(c)$ et $H_2(p)$, où $p = (p_1, ..., p_n)$. On a :

$$\begin{aligned}
\mathbb{E}(c) - H_2(p) &= \sum_{i=1}^{N} p_i l_i - \left(-\sum_{i=1}^{N} p_i \log p_i\right) \\
&= \sum_{i=1}^{N} p_i l_i - \sum_{i=1}^{N} \left(-p_i \log p_i\right)
\end{aligned}$$

par (*)

$$\begin{aligned}
&= -\left(\sum_{i=1}^{N} p_i \log\left(2^{-l_i}\right) + \sum_{i=1}^{N} \left(-p_i \log p_i\right)\right) \\
&= -\sum_{i=1}^{N} \left(p_i \log\left(2^{-l_i}\right) - p_i \log p_i\right) \\
&= -\sum_{i=1}^{N} \left[p_i \left(\log\left(2^{-l_i}\right) - \log p_i\right)\right]
\end{aligned}$$

par les propriétés fondamentales du logarithme

$$= -\sum_{i=1}^{N} \left[p_i \left(\log \frac{1}{p_i} + \log \left(2^{-l_i} \right) \right) \right]$$

toujours par les propriétés fondamentales du logarithme

$$= -\sum_{i=1}^{N} p_i \log \left(\frac{2^{-l_i}}{p_i} \right)$$

Ainsi, par l'inégalité de convexité :

$$\begin{aligned} &\leq \sum_{i=1}^{N} p_i \left(\frac{2^{-l_i}}{p_i} - 1 \right) \\ &\leq \sum_{i=1}^{N} \left(2^{-l_i} - p_i \right) \\ &\leq \sum_{i=1}^{N} 2^{-l_i} - \sum_{i=1}^{N} p_i \end{aligned}$$

p étant une distribution de probabilité :

$$\begin{aligned} &\leq \sum_{i=1}^{N} 2^{-l_i} - 1 \\ &\leq 0 \quad \text{car } (l_i) \text{ satisfait l'inégalité de Kraft} \end{aligned}$$

Ainsi, on a montré :

Théorème 3.1.3. *Pour tout questionnaire c :*

$$\mathbb{E}(c) \geq H_2(p)$$

Remarque. Si on prend $l_i = -\log p_i$, on a que $\sum 2^{-l_i} = \sum p_i = 1$, donc les l_i vérifient toujours l'inégalité de Kraft, donc si les l_i ainsi définis étaient entiers, on pourrait leur associer un questionnaire c dont la longueur moyenne serait précisément $H_2(p)$. Mais $-\log p_i$ est loin d'être toujours entier. La borne que l'on a donné n'est donc pas forcément toujours atteignable. En revanche, si on prend l_i égal au plus petit entier supérieur ou égal à $-\log p_i$, les l_i ainsi définis sont entiers, et vérifient toujours l'inégalité de Kraft. On peut donc leur associer un questionnaire c. Ils vérifient également, par définition :

$$-\log p_i \leq l_i < -\log p_i + 1$$

Ainsi :

$$-p_i \log p_i \leq p_i l_i < -p_i \log p_i + p_i$$

d'où

$$-\sum_{i=1}^{N} p_i \log p_i \leq \sum_{i=1}^{N} p_i l_i < -\sum_{i=1}^{N} p_i \log p_i + \sum_{i=1}^{N} p_i$$

c'est-à-dire :

$$H_2(p) \leq \mathbb{E}(c) < H_2(p) + 1$$

En résumé, on a réussi dans cette partie à exhiber une borne inférieure de la longueur moyenne d'un questionnaire, à savoir $H_2(p)$. De plus, cette borne paraît assez pertinente dans la mesure où on a vu grâce à la remarque précédente qu'il existait des questionnaires dont la longueur moyenne était "proche" de $H_2(p)$. On sait donc maintenant que quoiqu'on fasse, il nous sera impossible de faire un questionnaire de moins de $H_2(p)$ questions.

3.2 Formalisation du procédé proposé

Maintenant que l'on a vu qu'un questionnaire optimal ne pouvait pas être de longueur moyenne inférieure à l'entropie, il nous reste à savoir comment construire un questionnaire optimal. Les remarques faites en début d'exposé nous avaient suggéré le procédé consistant à poser des questions dont on a aucune raison de privilégier une réponse affirmative ou négative, c'est-à-dire, en utilisant le vocabulaire introduit précédemment, à séparer les personnages en deux groupes dont la probabilité est à peu près la même. Essayons de formaliser un peu la chose : reprenons les notations précédentes, avec $\Omega = \{1, ..., N\}$ l'ensemble des personnages possibles, muni de la distribution de probabilité (p_i). Quitte à réordonner les personnages, on peut supposer $p_1 \geq ... \geq p_N$. On va, pour définir les questions de notre questionnaire, appliquer à chaque étape un principe dit de "dichotomie" : soit u le plus petit entier tel que :

$$p_1 + ... + p_u \geq \frac{1}{2}$$

On sépare Ω en deux parties M_0 et M_1, où $M_0 = \{1, ..., u\}$, et $M_1 = \{u+1, ..., N\}$. On définit notre première question par : "l'objet est-il dans M_1 ?", c'est-à-dire, pour le reformuler en terme de code, que l'on attribue 0 comme première lettre de nom de code aux éléments de M_0, 1 à ceux de M_1. Puis, on continue et on applique le principe de dichotomie à M_0, muni de la distribution conditionnelle "sachant M_0", et de même pour M_1 muni de la distribution conditionnelle "sachant M_1". Pour être plus précis, c'est-à-dire que l'on sépare M_0 en $M_{00} = \{1, ..., v\}$ et $M_{01} = \{v+1, .., u\}$, où v est le plus petit entier tel que :

$$\frac{p_1}{p_{M_0}} + ... + \frac{p_v}{p_{M_0}} \geq \frac{1}{2}$$

et de même pour M_1. On continue jusqu'à ne plus pouvoir faire la dichotomie, et on obtient un questionnaire que l'on appellera C_{Fano}, du nom de son inventeur Robert Fano. Le paragraphe précédent pouvant paraître aride, nous conseillons au lecteur, pour bien se représenter la méthode, de regarder la vidéo de l'exposé ou de taper "algorithme de Fano" sur Youtube, dans la mesure où la vidéo est un *medium* plus adapté que l'écrit pour illustrer la méthode - qui est en substance extrêmement simple.

Malheureusement, ce code n'est pas toujours optimal. En effet, il existe des cas un peu "pathologiques" pour lesquels le procédé ne sépare pas l'ensemble en deux sous-ensembles de probabilités à peu près égales : reprenons les notations précédentes, et imaginons que $p_1 + ... + p_{u-1}$ est très proche de $1/2$, mais strictement inférieur, et que $p_1 + ... + p_u$ soit sensiblement plus grand que $1/2$. Si on sépare selon le principe de dichotomie, on obtient tout de même des ensembles de probabilités sensiblement différentes. Cependant, on peut tout de même montrer (mais cela rallongerait trop l'exposé) que le code de Fano est "presque optimal", c'est-à-dire qu'il est proche de la borne inférieure :

Théorème 3.2.1. *Si $p = (p_1, ..., p_N)$:*

$$H_2(p) \leq \mathbb{E}\left(C_{Fano}\right) < H_2(p) + 1$$

3.3 Un code véritablement optimal : le code de Huffmann

Pour le lecteur curieux, il existe un procédé similaire connu sous le nom de "méthode de Huffmann", qui présente l'avantage d'être véritablement optimale au sens donné dans l'exposé.

Conclusion

Notre cahier des charges a donc quasiment été rempli : nous avons donné un sens précis aux notions d'incertitude, de questionnaire, et d'optimalité d'un questionnaire, ainsi qu'au procédé intuitif évoqué en introduction. Cependant, ce dernier s'avère être "presque optimal", mais pas optimal au sens strict que nous lui avons donné, à une question près. Notre stratégie était donc une bonne stratégie, mais il en existe certaines qui ont des résultats très légèrement meilleurs. Mais au-delà du problème en lui-même, l'objectif principal était de sensibiliser le lecteur à la démarche utilisée, qui n'est pas celle de n'importe quel mathématicien. En effet, les mathématiques sont généralement séparées - un peu artificiellement - en deux branches : les mathématiques fondamentales et les mathématiques appliquées. Dans le cas des mathématiques fondamentales, le lien avec le concret importe peu : les objets mathématiques sont étudiés pour eux-mêmes, en complète autonomie vis à vis du réel. L'approche des mathématiques appliquées au contraire, ne cherche pas à se suffire à elle-même ; si les objets d'étude peuvent être très abstraits, l'objectif reste toujours de résoudre des problèmes en lien direct avec la réalité. Ainsi, à la différence des mathématiques fondamentales, le mathématicien appliqué se retrouve confronté au problème de la modélisation, ce qui veut dire qu'il doit essayer de traduire le réel dans un langage propre à la démonstration. Et c'est justement dans cet esprit que s'inscrit notre démarche : nous nous sommes posés un problème concret, mais dont les termes étaient trop flous pour produire un énoncé pouvant faire l'objet d'une démonstration. Nous nous sommes donc donnés un cadre formel, abstrait, au sein duquel tous les notions qui nous intéressaient prenaient un sens clair et précis. En d'autres termes, nous avons modélisé le problème. Puis, à l'aide des notions mathématiques nouvellement introduites, nous avons proposé une résolution du problème.
Mais le plus important est de ne pas oublier ce que "résoudre" veut dire : nous avons résolu notre problème *au sens* des termes de notre modèle. Par exemple, dans notre cas, nous avons considéré - non sans raison, bien sûr - que "découvrir au plus vite l'identité du personnage mystère" se traduisait par une longueur moyenne minimale. Il est donc nécessaire d'avoir à l'esprit que l'étape de formalisation du problème repose sur des partis pris, des biais de pensée, et qu'un modèle se définit aussi par ses limites. Bien entendu, cela ne veut pas dire qu'un modèle est inutile ; il faut simplement se rappeler qu'il s'agit d'un point de vue sur le monde, et un point de vue qui a beaucoup de choses à nous apprendre, mais qu'il ne faut en aucun cas sacraliser et lui faire dire plus qu'il ne peut nous dire.

Annexe mathématique

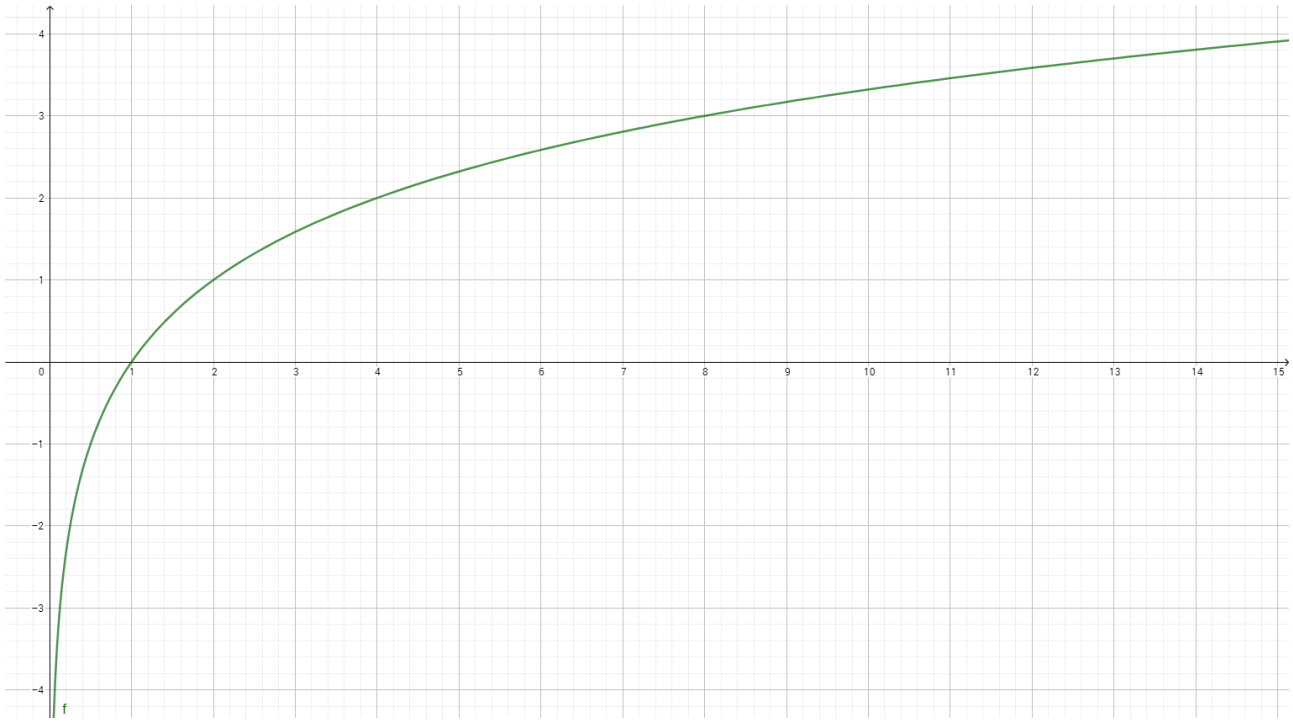


FIGURE 6 – Graphe du logarithme

FIGURE 7 – Inégalité de convexité

Références

[1] Pierre Brémaud. *Initiation aux Probabilités et aux chaînes de Markov*. Springer.

[2] Michel Zinsmeister. *Formalisme thermodynamique et systèmes dynamiques holomorphes*. SMF.