Interprétation ou Description (II) : Fondements mathématiques de l'approche F+D

Pierre Bessiere, Eric Dedieu, Olivier Lebeltel, Emmanuel Mazer, Kamel Mekhnacha

Pierre Bessière
Eric Dedieu
Olivier Lebeltel
Emmanuel Mazer
Kamel Mekhnacha

Laboratoire LEIBNIZ - CNRS

FONDEMENTS MATHEMATIQUE DE L'APPROCHE F+D

I. PAL : UN FONDEMENT MATHEMATIQUE POUR L'APPROCHE F+D

La théorie des probabilités n'est rien d'autre que le sens commun fait calcul.

Laplace

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind[1].

James Clerk Maxwell[2]

By inference we mean simply: deductive reasoning whenever enough information is at hand to permit it; inductive or probabilistic reasoning when - as is almost invariably the case in real problems - all the necessary information is not available. Thus the topic of « Probability as Logic » is the optimal processing of uncertain and incomplete knowledge[3].

E.T. Jaynes

Cette article fait suite et complète l'article intitulé « *Interprétation ou Desciption : proposition pour une théorie probabiliste des systèmes cognitifs sensori-moteurs* » paru dans ce même numéro. L'objectif poursui est de présenter les principes des fondements mathématiques de l'approche F+D.

E. T. Jaynes propose dans (Jaynes95) une théorie de la cognition fondée sur les probabilités appelée « Probability as Logic » (*PAL*).

Cette théorie fournit les deux composantes fondamentales dont nous avons besoin pour notre approche F+D, d'une part, des règles formelles permettant de raisonner sur des données incertaines et incomplètes, d'autre part, le principe de maximum d'entropie, qui permet de clarifier le lien entre descriptions et expériences et donne un cadre théorique général pour l'apprentissage.

---

[1] La science logique contemporaine parle uniquement de faits certains, impossibles, ou complètement douteux - sujets sur lesquels nous n'avons (heureusement) jamais à raisonner. En conséquence, la vraie logique pour notre monde est le calcul des probabilités, qui prend en compte une quantification des plausibilités qui sont, ou devraient être, dans tout esprit rationnel.

[2] Citation empruntée à (Jaynes95)

[3] Par inférence, nous entendons le raisonnement déductif quand on dispose d'assez d'information pour le permettre; le raisonnement inductif ou probabiliste quand, comme c'est presque toujours le cas pour les problèmes réels, toute l'information nécessaire n'est pas disponible. C'est pourquoi le propos de « Probability as Logic » est le traitement optimal de l'information incertaine et incomplète.

II. RAISONNER MALGRE L'INCERTITUDE ET L'INCOMPLETUDE : L'INFERENCE PROBABILISTE.

Le raisonnement probabiliste peut être fait par un système de calcul formel. Il est fondé sur les probabilités alors que la plupart des systèmes formels sont basés sur la logique. Le raisonnement probabiliste permet de calculer pour chaque proposition une valeur réelle comprise entre 0 et 1, interprétable comme une plausibilité. Les systèmes formels courants servent, eux, à dériver ou réfuter les propositions, ce qui revient à calculer une valeur qui est soit 0, soit 1.

II.1. Le théorème de Cox

Le théorème de Cox est le résultat fondamental montrant comment la notion intuitive de plausibilité se formalise par la notion mathématique de probabilité.

Notons $\pi(A|C)$ la plausibilité d'une proposition A, jugée au vu d'un ensemble de connaissances C. Les postulats[1] suivants explicitent cette notion de plausibilité et permettent de dériver la forme mathématique exacte qu'elle doit prendre :

- Les plausibilités sont représentées par des nombres réels (i.e. $\pi(A|C)$ est un réel)[2].
- S'il existe plusieurs calculs corrects pour déterminer une plausibilité, ils doivent tous mener au même résultat (propriété dite de « consistance »).
- Si de nouvelles informations C' viennent remplacer les informations C de façon à augmenter la plausibilité de A, alors la plausibilité de la proposition contraire $\neg A$ doit diminuer. Formellement :

si $\pi(A|C') > \pi(A|C)$, alors $\pi(\neg A|C') < \pi(\neg A|C)$ [4.1]

- Si de nouvelles informations augmentent la plausibilité de A mais ne concernent en rien une autre proposition B, alors la plausibilité de la conjonction de A et B (noté AB) ne peut qu'augmenter. Formellement :

si $\begin{cases}\pi(A|C') > \pi(A|C) \\ \pi(B|AC') = \pi(B|AC)\end{cases}$, alors $\pi(AB|C') \geq \pi(AB|C)$ [4.2]

A partir de ces postulats de base R.T. Cox a montré que le raisonnement plausible devait obligatoirement suivre 2 règles à partir desquelles toute la théorie des probabilités peut être reconstruite.

La règle qui donne la probabilité d'une conjonction :

$P(AB|C) = P(A|C) \times P(B|AC) = P(B|C) \times P(A|BC)$ [R1]

Et la règle qui exprime que la somme des probabilités d'une proposition et de sa négation est égale à 1 :

$P(A|C) + P(\neg A|C) = 1$ [R2]

Ce théorème montre de plus que toute technique de calcul de plausibilité qui ne vérifierait pas ces deux règles, enfreindrait nécessairement l'un au moins des postulats posés au paragraphe

---

[1] Exprimés d'après (Jaynes95)

[2] Par commodité, nous représenterons une plus grande plausibilité par un nombre plus grand — mais ce choix est une pure convention.

précédent. Donc si l'on accepte ces postulats, le calcul des probabilités, et lui seul, permet de faire du raisonnement plausible.[1]

On déduit très aisément des règles R1 et R2 la règle qui donne la probabilité de la disjonction (A+B):

$$P(A + B|C) = P(A|C) + P(B|C) - P(AB|C) \qquad [R3]$$

II.2. Les syllogismes du raisonnement probabiliste

La déduction logique est essentiellement fondée sur les deux règles suivantes :

- Modus ponens : si « A est vrai » et si « Si A est vrai alors B est vrai » alors « B est vrai ».
- Modus tollens : si « B est faux » et si « Si A est vrai alors B est vrai » alors « A est faux ».

Cependant, nous employons constamment des critères plus « faibles » que ces deux règles :

Le premier peut être énoncé de la manière suivante : si « Si A est vrai alors B est vrai » et si « B est vrai » alors « A devient plus plausible ». La vérification que la conséquence d'une implication est vraie renforce la confiance que l'on a dans la prémisse de cette implication.

Un exemple de ceci est donné par les propositions suivantes :

- A ≡ « La valeur du proximètre avant gauche du Khépéra (V3) est supérieure à 10 » ;
- B ≡ « Il y a un obstacle proche droit devant ».

La présence de l'obstacle (B vrai) ne garantit pas totalement que le capteur V3 réponde fortement[2] (A vrai), mais contribue certainement à nous convaincre que la valeur de V3 doit être élevée (A devient plus plausible).

Le second crirère faible s'énonce : si « Si A est vrai alors B est vrai » et si « A est faux » alors « B devient moins plausible ». Supprimer une des prémisses qui pouvait impliquer la vérité d'un fait donné, fait diminuer notre confiance dans cette vérité. Dans l'exemple précédent, si le proximètre V3 ne répond pas, nous aurons plus de mal à nous persuader qu'il peut y avoir un obstacle.

Enfin, le troisième critère est plus faible encore : si « Si A est vrai alors B devient plus plausible » et si « B est vrai » alors « A devient plus plausible ». C'est typiquement celui sous-tendant les enquêtes policières et les décisions de justice. « Si X est coupable alors il est probable que X ait laissé ses empreintes sur l'arme du crime », « X a laissé ses empreintes sur l'arme du crime » donc il devient plus plausible que « X soit coupable ». En fait, l'exemple pris pour illustrer les 2 règles précédentes est plutôt de cette nature, car la réponse élevée du proximètre avant gauche n'implique pas de manière absolue la présence d'un obstacle. Des exceptions à une telle implication sont toujours facilement imaginables[3].

La logique formelle est incapable de prendre en compte ces critères « faibles » pourtant si utiles en pratique. Le raisonnement probabiliste les intègre naturellement. De plus, et c'est très important, il apparaît que si on se limite à des propositions de probabilité 0 ou 1 (certainement vraies ou certainement fausses), les règles du raisonnement probabiliste permettent de tenir tout

---

[1] Une autre manière de voir ce résultat, consiste à se demander, pour chaque théorie non probabiliste des plausibilités, lequel des 4 postulats n'est pas verifié. Une fois répondu à cette question, on peut discuter de manière plus saine et plus sereine sur les inconvénients et avantages relatifs de la dite théorie et du calcul des probabilités.

[2] Par exemple, le capteur ne repond pas aux obstacles trop sombres absorbant les infrarouges.

[3] Par exemple, le capteur V3 est « grillé » et indique la valeur maximum de 15 en toute circonstance.

raisonnement logique souhaité - et c'est pourquoi l'approche F+I est un cas particulier de l'approche F+D.

II.3. Problèmes directs et inverses

Pour une description donnée, on peut distinguer les deux ensembles de variables suivants :
- d'une part, un ensemble Δ de variables « observables » dont les valeurs peuvent être « mesurées » par un quelconque moyen ;
- d'autre part, un ensemble Γ de variables non directement accessibles, dont les valeurs peuvent uniquement être « estimées » par le raisonnement ou données *a priori*.

Soit D un ensemble de mesures des variables de Δ appelé ensemble de données, soit C un ensemble de valeurs pour les variables de Γ appelé ensembles de paramètres ou connaissances préalables, on s'intéresse à l'étude des probabilités conjointes de D et C.

La règle [R1] nous donne :

$P(DC) = P(C) \times P(D \mid C) = P(D) \times P(C \mid D)$

Si l'on admet connaître les valeurs de certains paramètres et si l'on fixe par hypothèse celles des autres, le « problème direct » consiste à « prédire » les valeurs des données. Mathématiquement, cela signifie que l'on cherche $P(D \mid C)$.

Inversement, si l'on connaît un ensemble de données et que l'on cherche $P(C \mid D)$ on traite le « problème inverse ». Il consiste à chercher les paramètres qui rendent « au mieux » compte des expériences ou encore à choisir parmi plusieurs hypothèses envisageables laquelle est la plus probable. On voit qu'en particulier, le problème inverse recouvre les problèmes de type identification dont il a été question au paragraphe précédent.

Pour illustrer le problème direct et le problème inverse et pour donner quelques exemples élémentaires de raisonnement probabiliste, plaçons-nous dans le cas ou notre système est une urne de Bernouilli remplie de boules blanches et noires dans laquelle on effectue des tirages sans remise.

La connaissance préalable C s'exprime par les énnoncés N ≡ « L'urne contient n boules » et par B ≡ « L'urne contient b boules blanches », paramétrées par n et b.

Les données D peuvent s'exprimer, par exemple, sous la forme $T_i \equiv$ « On observe le tirage d'une boule blanche au $i^{ème}$ tirage ».

Le problème direct s'intéresse aux probabilités de D, c'est-à-dire de T_i, sachant C, c'est-à-dire n et b.

Considérons tout d'abord à la probabilité $P(T_2 \mid C)$ de tirer une boule blanche au deuxième tirage. Nous pouvons écrire que $T_2 = (T_1 + \neg T_1)T_2 = T_1T_2 + \neg T_1T_2$. En appliquant la règle [R3] puis la règle [R1] on obtient :

$$
\begin{aligned}
P(T_2 \mid C) &= P(T_1T_2 \mid C) + P(\neg T_1T_2 \mid C) \\
&= P(T_2 \mid T_1C)P(T_1 \mid C) + P(T_2 \mid \neg T_1C)P(\neg T_1 \mid C) \\
&= \left(\frac{b-1}{n-1} \times \frac{b}{n}\right) + \left(\frac{b}{n-1} \times \frac{n-b}{n}\right) \\
&= \frac{b}{n} \\
&= P(T_1 \mid C)
\end{aligned}
\qquad [4.3]
$$

On vérifie bien que la probabilité d'avoir une boule blanche au deuxième tirage (ne sachant pas le résultat du premier) est la même que celle d'avoir une boule blanche au premier tirage. Il en est d'ailleurs de même pour tous les tirages.

Intéressons-nous maintenant à $P(T_1 \mid T_2C)$ la probabilité de tirer une boule blanche au premier tirage sachant le tirage d'une boule blanche au deuxième.

La règle [R1] nous donne :

$$P(T_1T_2 \mid C) = P(T_1 \mid C)P(T_2 \mid T_1C) = P(T_2 \mid C)P(T_1 \mid T_2C) \qquad [4.4]$$

Comme on vient de démontrer en [4.3] que :

$$P(T_1 \mid C) = P(T_2 \mid C)$$

On obtient:

$$P(T_1 \mid T_2C) = P(T_2 \mid T_1C) \qquad [4.5]$$

Ce raisonnement peut se reproduire pour un i et un j quelconque et on obtient finalement :

$$P(T_i \mid T_jC) = P(T_j \mid T_iC) \qquad \forall i, \text{j} \qquad [4.6]$$

Ce résultat peut surprendre, essentiellement parce qu'il peut paraître surprenant que la probabilité du tirage d'une boule blanche au premier tirage puisse être influencée par le deuxième tirage. Une idée très profondément ancrée dans nos esprits cartésiens et newtoniens est qu'un événement ne peut être influencé que par les événements qui l'ont précédé dans le temps. En fait, contrairement aux apparences, ce résultat ne remet pas en cause ce principe de la physique. La confusion vient de ce que l'on considère que le premier tirage a été influencé par le second alors que ce que nous dit le résultat [4.5] c'est que la connaissance que l'on a de ce qui peut se passer au premier tirage est éventuellement influencée par la connaissance du résultat du deuxième tirage. Il faut bien distinguer entre la causalité physique qui ne peut que se propager vers le futur et la capacité d'inférence qui fonctionne aussi bien dans un sens que dans l'autre.

Le calcul qui a conduit à [4.5] est particulièrement simple et résulte d'un usage tout à fait élémentaire des règles des probabilités. Pourtant, même ce résultat trivial n'est compréhensible que si l'on considère une distribution de probabilité comme l'état de connaissance d'un individu sur un certain phénomène et non pas comme une description de ce phénomène indépendamment de l'observateur. Ce seul exemple pose déjà un sérieux problème pour toute épistémologie objectiviste des probabilités.

Plaçons-nous maintenant, pour illustrer le problème inverse, dans le cas où n le nombre total de boules est connu mais où b le nombre de boules blanches n'est pas connu. On va chercher à cerner b grâce à un ensemble de données D constitué avec des tirages issus de cette urne.

Alors que pour le problème direct, nous nous intéressions à la probabilité de D sachant C=NB, nous nous intéressons, maintenant à la probabilité de B sachant D et N, d'où le nom de problème inverse.

La règle [R1] nous donne :

$$P(BD \mid N) = P(B \mid N) \times P(D \mid BN) = P(D \mid N) \times P(B \mid DN) \quad [4.7]$$

d'où on tire :

$$P(B \mid DN) = P(B \mid N)\frac{P(D \mid BN)}{P(D \mid N)} \quad [4.8]$$

On constate que le résultat dépend à la fois de P(D | BN) probabilité directe, et de P(B | N)[1] : cette distribution, dite « *a priori* » depuis la fin du XVIII[e] siècle, doit être connue - donnée ou calculable - pour qu'on puisse effectuer le calcul. Quant à P(D | N), elle peut toujours être obtenue par normalisation en sommant P(B | DN) sur toutes les valeurs de B possibles. Ainsi le résultat d'une série d'expériences (D) n'est pas interprétable en soi, mais uniquement au jour de connaissances préalables apportées ici par la distribution *a priori*. Voici un deuxième exemple de ce qui vient juste d'être dit, à savoir que le raisonnement probabiliste ne peut pas être la description d'une quelconque réalité indépendante de l'observateur, mais nécessite pour interpréter toute donnée la présence d'informations préalables sur l'état de connaissance de cet observateur.

Il serait trop long de développer ici les calculs pour les différentes distributions *a priori* envisageables. Nous invitons donc le lecteur à se reporter au chapitre 6 de (Jaynes95) pour satisfaire sa curiosité concernant P(B|DN).

II.4. Moteur d'inférence probabiliste

Un moteur d'inférence et d'apprentissage bayésien a été développé. Ce moteur a pour ambition d'être capable d'automatiser les raisonnements probabilistes. Il peut être, en cela, comparé aux nombreux moteurs d'inférences logiques qui existent et permettent d'automatiser le raisonnement symbolique (voir, par exemple, Prolog (Colmerauer86), LogLisp (Robinson83a) ou Lolita (Bessière87)).

Une première version du moteur d'inférence a été développé, fonctionne et est utilisée. Son principe, très simplifié, est le suivant :

Si l'on connait :

- $P(X_1, X_2 ..., X_n)$ la distribution de probabilité conjointe sur les variables X1, X2, ... Xn sous la forme d'une décomposition de cette distribution conjointe en un produit de distributions plus simples, par exemple
  $P(X_1, X_2 ..., X_n) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)...P(X_n \mid X_{n-1}X_2)$;
- et, les formes paramétriques identifiées de toutes les distributions de ce produit ;

le moteur d'inférence est alors capable de répondre à toute question concernant la distribution conjointe.

Par toute question, on entend n'importe quelle expression de la forme $P(X_i ... X_j \mid X_k ... X_l)$ avec un nombre arbitraire de variables à gauche, un nombre arbitraire de variables à droite et un nombre arbitraire de variables inconnues (abscentes).

---

[1] Distribution de probabilité traduisant la connaissance a priori que l'on a du nombre b de boules blanches dans l'urne sachant le nombre n total de boules.

La solution générale a une telle question est donnée par :

$$P\left(X_i \cdots X_j \mid X_k \cdots X_l\right) = \frac{1}{Z} \sum_{\text{variables inconnues}} P(X_1) P(X_2 \mid X_1) P(X_3 \mid X_2) \ldots P\left(X_n \mid X_{n-1} X_2\right)$$

avec Z une constante de normalisation.

Pour une question donnée, cette expression est tout d'abord simplifiée symboliquement par le moteur. Sa valeur numérique est ensuite estimée. La complexité du calcul numérique vient évidemment des sommes. Nous avons développé dans le moteur une méthode d'estimation de ces sommes très performante qui fait notamment appel à un algorithme d'optimisation de type « génétique » et à une représentation des distributions sous forme très compacte et efficace, inspirée des Octree.

Il est essentiel de remarquer que dans l'approche préconisée ici toutes les variables jouent des rôles mathématiques parfaitement équivalents. Le moteur d'inférence étant capable de répondre à n'importe quelle question, il n'y a pas à proprement parler de problèmes directs et de problèmes inverses. Ceci est sans doute l'une des qualités essentielles de l'inférence probabiliste.

III. ANCRER LES DESCRIPTIONS : LE PRINCIPE DE MAXIMUM D'ENTROPIE

La deuxième composante nécessaire à notre approche F+D est de pouvoir disposer d'un moyen générique de construire les descriptions.

Etant donné un ensemble de connaissances préalables C et de données expérimentales D, il existe en général une infinité de distributions de probabilités (descriptions) compatibles avec ces connaissances. Toutefois, toutes ces descriptions ne sont pas équivalentes. Certaines semblent plus « cohérentes », plus « probables », plus « intéressantes » que d'autres. Cet « intérêt » peut être jugée dans un sens mathématique précis par la fonction entropie H. Pour une distribution discrète P qui assigne, à q propositions, les probabilités $\{p_1, ..., p_q\}$, l'entropie est définie par :

$$H(P) = -\sum_{i=1}^{q} p_i \log\left(p_i\right) \qquad [4.11]$$

Le principe de maximum d'entropie affirme que parmi les distributions compatibles avec un ensemble de connaissances préalables C et de données expérimentales D, la meilleure possible est celle qui maximise l'entropie H. Ce principe donne donc le moyen théorique et générique recherché pour construire les descriptions. Avant d'en justifier l'emploi présentons d'abord un exemple très simple afin de bien comprendre à quoi il sert et comment il peut être utilisé.

III.1. Vingt mille chiffres, pour quoi faire?

Supposons que nous ayons à notre disposition un ensemble de données expérimentales brutes sous la forme de 20 000 chiffres (voir figure 16). Un tel jeu de données est manifestement inexploitable tel quel. Sans connaissances préalables, ce n'est qu'une suite de signes cabalistiques sans aucune espèce de signification possible.

Une connaissance préalable minimale absolument requise consiste à préciser comment ces 20 000 chiffres se repartissent pour former des nombres. Supposons qu'ici chaque chiffre soit un nombre : on a donc 20 000 nombres correspondant à 20 000 valeurs d'un certain nombre, disons n, de variables.

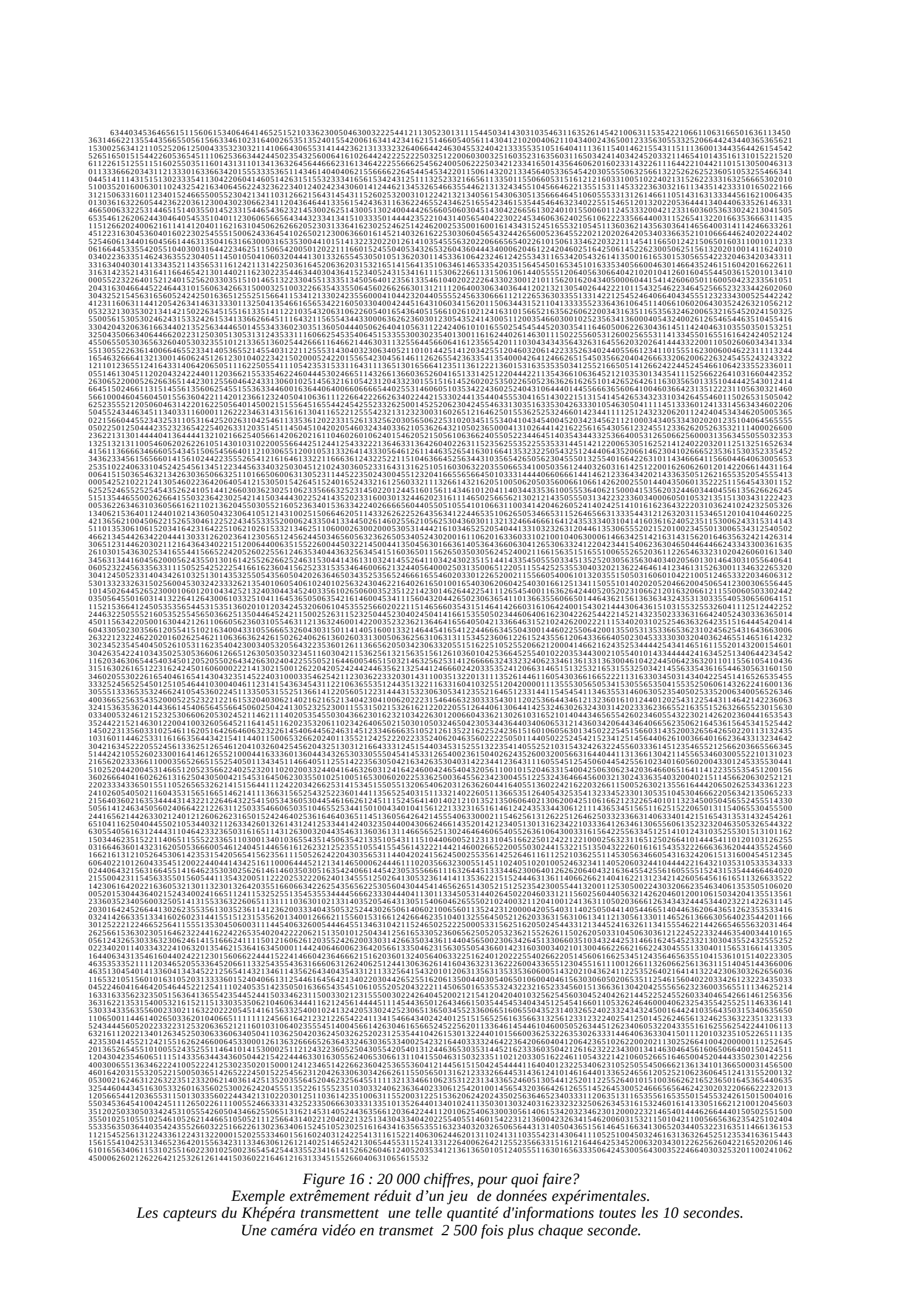

Figure 16 : 20 000 chiffres, pour quoi faire?
Exemple extrêmement réduit d'un jeu de données expérimentales.
Les capteurs du Khépéra transmettent une telle quantité d'informations toutes les 10 secondes.
Une caméra vidéo en transmet 2 500 fois plus chaque seconde.

Une deuxième connaissance préalable nécessaire est de garantir que ces 20 000 données forment un tout cohérent, correspondant bien à l'observation d'un seul et même phénomène et qu'il convient d'essayer de les analyser ensemble. [1]

Une troisième connaissance préalable consiste à préciser le nombre n des variables et à répartir entre elles les 20 000 valeurs. Supposons qu'ici une seule variable V ait été retenue. Nous avons donc 20 000 expériences concernant V.

Enfin, une quatrième connaissance préalable tout à fait déterminante suppose que l'ordre (temporel et/ou spatial) dans lequel ces 20 000 expériences se présentent n'est pas pertinent.[2]

Par ces choix, nous avons défini l'espace des phases et placé les points expérimentaux. Les connaissances préalables de ce type ont été qualifiées de structurelles dans l'article « *Interprétation ou Description* ».

Un premier travail sur les données devient maintenant possible. On peut en construire l'histogramme en dénombrant pour chaque valeur observée de la variable le nombre de fois ou elle est apparue (voir la ligne ni du tableau 1). Un histogramme n'est pas une distribution de probabilité (et donc pas une description) mais il permet une réduction, éventuellement considérable, de l'information.

Pour pouvoir produire notre première description, il faut encore faire une hypothèse supplémentaire, à savoir le nombre de valeurs possibles que peut prendre la variable observée. On peut alors démontrer (Jaynes95) que la meilleure description possible est la loi de succession de Laplace :

$$P([V=i]) = \frac{1+n_i}{\langle\Omega\rangle + n} \qquad [4.12]$$

avec <Ω> le nombre de valeurs possibles pour V, n_i le nombre d'expériences où la valeur i a été mesurée et n le nombre d'expériences totales.

Cette loi de succession de Laplace se rapproche de l'histogramme dès que le nombre d'expériences n devient grand devant le nombre de cas possible <Ω>. Elle présente, cependant, deux avantages sur l'histogramme : elle se comporte bien quand le nombre d'expériences n est faible (notamment, on retrouve la distribution uniforme pour $n=n_i=0$) et, surtout, elle ne fixe la probabilité d'aucun événement à 0, évitant par là un choix définitif devant reposer sur une certitude (une probabilité de 0 ne peut pas évoluer par apprentissage et rend impossible tout événement dépendant de l'événement de probabilité nulle). Cette loi de succession de Laplace est la forme la plus rudimentaire de description de données expérimentales. C'est elle qui suppose le moins de connaissances préalables et, donc, « colle » au plus près des données. C'est elle, aussi qui nécessite le plus de paramètres étant donné que pour coder n données il faut <Ω> paramètres.

---

[1] Cette exigence peut paraître évidente et triviale. Pourtant les physiciens et les biologistes savent bien l'extrême difficulté qu'il peut y avoir à garantir l'intégrité d'un jeu de données vis-à-vis d'un phénomène observé.
Le choix des données à traiter ensemble est, en fait, fondamental car il définit implicitement le phénomène qu'on veut étudier.
Si, par exemple, on sépare le jeu de 20 000 données en deux parties, cela signifie qu'on supose qu'une variable « cachée » (non mesurée) a changé d'état et qu'on préfére étudier ce qu'on a observé comme 2 phénomènes séparés.
C'est aussi souvent à ce stade que certaines données sont écartées car jugées « aberrantes». En rejetant ces données comme « extérieures » au phénomène, on définit subjectivement ses limites.

[2] Ce choix est très « compromettant», car il suppose que les dépendances qui existent entre ces 20 000 valeurs ne sont d'origine ni spatiale ni temporelle.

Supposons que nous apprenions maintenant que la variable V peut prendre une quelconque des 6 valeurs entre 1 et 6 (<Ω>=6). Nous allons alors pouvoir calculer effectivement les probabilités pour chaque valeur de V (voir ligne L du tableau 1). Notre collection de données expérimentales commence à « prendre du sens » et à permettre le raisonnement. Il devient notamment possible de faire des prédictions sur les résultats à venir.

La loi de succession de Laplace s'avère rapidement complètement inadaptée dès que l'on a des connaissances préalables plus riches. De ce fait, elle a souvent été critiquée dans l'histoire des probabilités. Par exemple, la loi de succession de Laplace attribue une probabilité de 11/12 à un enfant de dix ans de vivre une année de plus (<Ω>=2, n_i=10, n=10) alors qu'elle attribue pour son grand-père de 70 ans une probabilité de 71/72 au même événement (<Ω>=2, n_i=70, n=70). Ce résultat nous choque par son inexactitude évidente. Cependant, ce n'est pas la loi de succession de Laplace qui est en cause : c'est simplement que pour juger de la validité de ce résultat, nous disposons d'informations préalables qui n'ont pas été prises en compte pour établir ces probabilités.

Curieusement, face à des problèmes de ce genre, de nombreux auteurs ont jugé que les calculs utilisant la loi de succession de Laplace devaient être faux, ou ont invoqué d'obscurs principes pour bannir l'utilisation de cette distribution, alors qu'il leur suffisait de constater que son emploi n'était tout simplement pas en accord avec leurs prémisses.

Mathématiquement, rechercher la distribution maximisant l'entropie H revient à résoudre un problème d'optimisation sous contraintes. Bien qu'en théorie le principe de maximum d'entropie soit applicable à n'importe quel type de connaissances préalables, il s'avère qu'en pratique, la plupart du temps, on ne lui connaît pas de solution analytique.

Il existe cependant un type de contraintes (connaissances préalables) très souvent rencontré dans la pratique (notamment en physique) pour lequel on connaît la forme analytique de la distribution qui correspond au maximum d'entropie. Il s'agit du cas où les connaissances préalables C sont données sous forme de m fonctions réelles f_j, nommées *observables*, et dont on peut calculer les moyennes F_j sur les données expérimentales. On assume alors que les descriptions adéquates du phénomène sont données par les distributions P pour lesquelles les espérances des fonctions f_j sont égales aux moyennes expérimentales F_j.

$$\forall j, j \in \{1,\ldots,m\};\ \sum_{v \in D} P(v) f_j(v) = F_j \qquad [4.13]$$

La distribution de maximum d'entropie correspondant à ces contraintes est alors P* donnée par :

$$P^*(v) = \frac{1}{Z(\lambda_1,\ldots,\lambda_m)} e^{-\sum_{i=1}^{m} \lambda_i f_i(v)} \qquad [4.14]$$

où Z désigne la fonction de partition, donnée par :

$$Z(\lambda_1,\ldots,\lambda_m) = \sum_{v \in \Omega} e^{-\lambda_1 f_1(v) - \ldots - \lambda_m f_m(v)} \qquad [4.15]$$

les m coefficiants λ_i pouvant être déterminés par les m équations traduisant les contraintes [4.13]:

$$\frac{\partial}{\partial \lambda_j} ln\Big(Z\big(\lambda_1,...,\lambda_j,...,\lambda m\big)\Big) + F_j = 0 \qquad [4.16]$$

Ce type de distributions « exponentielles » permet de coder n données avec m paramètres qui sont, au choix, ou bien les F_i appelés « niveaux de contraintes » des observables, ou bien les λ_i appelés « multiplicateurs de Lagrange ».

Il existe en physique un très grand nombre de telles distributions de maximum d'entropie correspondant au choix de tel ou tel ensemble d'observables, à commencer, bien sûr, par les distributions habituelles de la physique statistique. La plus connue de ces distributions de maximum d'entropie exponentielle est la Gaussienne. Elle correspond au cas à une variable V, avec les 2 observables $f_1(V)=V$ et $f_2(V)=V^2$. La Gaussienne est donc la distribution de maximum d'entropie lorsque ce que l'on a retenu des expériences c'est leur moyenne expérimentale et celle de leur carré.

Appliquons ce type de raisonnement à nos 20 000 expériences.

Supposons que nous apprenions que nos 20 000 expériences sont les résultats de lancers consécutifs d'un dé. Force est de constater d'après les résultats expérimentaux que ce dé semble largement pipé (la loi de succession de Laplace obtenue est très différente de la distribution uniforme escomptée).

Un dé est un cube dans lequel de légères excavations ont été faites pour graver les chiffres sur les faces. Chacune de ces excavations provoque un léger déplacement du centre de gravité du dé en direction opposée. Une première hypothèse possible (C1) pour justifier le biais constaté consiste alors à dire que globalement le centre de gravité se déplace de manière à favoriser l'apparition des faces les plus légères, c'est-à-dire celles portant les chiffres les plus élevés. Pour tenter d'observer cet effet escompté, nous choisissons l'observable $f_1(V)=V$. La moyenne expérimentale de $f_1(V)=$ vaut F1=3,5983. L'application du principe de maximum d'entropie nous conduit alors à une description du phénomène (une distribution de maximum d'entropie) ayant la forme :

$$P^*(v) = \frac{1}{Z} e^{-\lambda_1 v} \qquad [4.17]$$

avec $\lambda_1 = 3{,}372\ 10^{-2}$ et Z une constante de normalisation. Les probabilités correspondantes sont données dans la ligne C1 du tableau 2. Ces valeurs semblent encore très mal rendre compte des expériences.

Il est possible d'usiner 5 faces d'un dé sans avoir à le manipuler. Par contre, l'usinage de la dernière face suppose obligatoirement une manipulation et donc une perte de précision. On peut donc supposer qu'un dé a de bonne chance d'avoir un côté légèrement plus long (ou plus court) que les 2 autres. L'apparition des deux faces correspondantes sera alors légèrement moins probable (ou plus probable) que celle des 4 autres. On a donc 3 nouvelles hypothèses à tester :

- le côté reliant les faces 1 et 6 est de longueur différente (C2)
- le côté reliant les faces 2 et 5 est de longueur différente (C3)
- le côté reliant les faces 3 et 4 est de longueur différente (C4)

Prenons l'exemple de l'hypothèse C2, on peut essayer d'en quantifier l'observation avec l'observable $f_2(V)=0$ si V vaut 2, 3, 4 ou 5 et $f_2(V)=-1$ si V vaut 1 ou 6. La moyenne expérimentale de $f_2(V)$ vaut donc :

	V=1	V=2	V=3	V=4	V=5	V=6	Total
n_i	3246	3449	2897	2841	3635	3932	20000
U Uniforme	0.16666	0.16667	0.16667	0.16667	0.16666	0.16667	1
L Laplace	0,16230	0,17245	0,14486	0,14206	0,18174	0,19659	1
C1	0,15294	0,15818	0,16361	0,16922	0,17502	0,18103	1
C2	0,16497	0,15259	0,15760	0,16278	0,16813	0,19393	1
C3	0,14803	0,16808	0,15843	0,16390	0,18612	0,17543	1
C4	0,16433	0,16963	0,14117	0,14573	0,18656	0,19258	1

Tableau 1

$$\frac{1}{20000}\sum_{i=1}^{6} n_i f_2(i) = -0,3589 \qquad [4.18]$$

L'application du principe de maximum d'entropie à cette hypothèse C2, nous conduit alors à une description du phénomène (une distribution de maximum d'entropie) ayant la forme:

$$P^*(V) = \frac{1}{Z} e^{-\lambda_1 V - \lambda_2 f_2(V)} \text{ avec } \lambda_1 = 3,234\ 10^{-2}, \lambda_2 = -1.104\ 10^{-1} \qquad [4.19]$$

Les probabilités correspondantes sont données dans la ligne C2 du tableau 1. Les mêmes calculs peuvent être effectués pour les hypothèses C3 et C4. Les résultats correspondants sont présentés dans le tableau 1 aux lignes C3 et C4.

A la vue du tableau 1, l'hypothèse C4 semble rendre compte des données expérimentales beaucoup mieux que les autres. Cette intuition peut se vérifier mathématiquement en appliquant les techniques déjà rencontrées précédemment sous le nom de « problème inverse ». La règle [R1] nous donne:

$$P(C_i \mid D) = P(C_i)\frac{P(D \mid C_i)}{P(D)} \qquad [4.20]$$

Si nous supposons, *a priori*, que les 6 hypothèses en présence (U, L, C1, C2, C3 et C4) sont équiprobables, nous obtenons:

$$P(C_i \mid D) = \frac{P(D \mid C_i)}{\sum_{j=1}^{6} P(D \mid C_j)} \qquad [4.21]$$

P(D|C_i) est donné par :

$$P(D|C_i) = W p_{i1}^{3246} p_{i2}^{3449} p_{i3}^{2897} p_{i4}^{2841} p_{i5}^{3635} p_{i6}^{3932} \qquad [4.22]$$

W (indépendant de Ci) est le nombre de permutations sur les 20000 expériences avec 3246 fois le résultat 1, 3449 fois le résultat 2, 2897 fois le résultat 3, 2841 fois le résultat 4, 3635 fois le résultat 5 et 3932 fois le résultat 6. P_{ij} est la probabilité d'obtenir le résultat j pour l'hypothèse Ci. En effectuant les calculs numériques on obtient :

Hypothèses	Laplace	C4	C2	C3	C1	U
P(Ci | D)	0.99	0.01	10^{-32}	10^{-35}	10^{-45}	10^{-59}

Tableau 2

L'hypothèse C4 est donc énormément plus probable que toutes les autres - sauf Laplace, qui se contente de coller aux résultats expérimentaux sans chercher à les « expliquer » .

L'hypothèse C4 ayant une interprétation et une justification physique, on est libre de considérer qu'il vaut mieux l'adopter comme description que Laplace[1]. C'est un choix subjectif, qui peut se traduire mathématiquement en lui donnant une probabilité *a priori* P(C4) beaucoup plus grande qu'à la loi de succession de Laplace.

En fait, les données utilisées dans cet exemple sont des données réelles issues d'une expérience menée par Wolf il y a plus d'un siècle. Six nombres (les nombres de fois où chacun des 6 résultats possibles est apparu), quelques connaissances préalables traduisant un peu de bon sens physique, et le principe de maximum d'entropie : voilà qui nous permet de penser avec une très grande confiance que ce dé que nous n'avons évidemment jamais vu, était légèrement oblong suivant l'axe « 3-4 ».

III.2. Justification du principe de maximum d'entropie

La justification la plus intuitive du principe de maximum d'entropie repose sur un argument combinatoire issu directement de son origine venant de la mécanique statistique. Cet argument fut proposé originellement par Boltzmann.

Supposons que nous ayons un ensemble de n particules identiques et que chacune de ces particules puisse être dans q différents états microscopiques équiprobables.

Définissons ν_k, un état macroscopique, comme un ensemble $\{n_1,...,n_q\}$ de q nombres tel que n_i soit le nombre de particules dans l'état microscopique i.

Nous devons, bien sûr, avoir :

$$\sum_{i=1}^{q} n_i = n \qquad [4.23]$$

et de plus, par exemple, des contraintes telles que :

---

[1]Ce genre de connaissance préalable ayant un sens physique permet en général une meilleure généralisation que la loi de succession de Laplace.
De plus, cela présente l'immense avantage pratique de conduire à des représentations beaucoup plus compactes : m (nombre d'observables) paramètres sont suffisants, au lieu des <Ω> nécessaires avec Laplace.

$$\sum_{i=1}^{q} n_i e_i = e \qquad [4.24]$$

où, e désigne l'énergie globale du système et e_i l'énergie de l'état i.

Appelons $W(\nu_k)$ le nombre de manière de réaliser l'état macroscopique ν_k. Nous avons:

$$W(\upsilon_k) = \frac{n!}{n_1 ! \times n_2 ! \times ... \times n_q !} \qquad [4.25]$$

Pour Boltzmann, l'état macroscopique le plus probable est alors celui qui peut être réalisé du plus grand nombre de manières microscopiques possible, c'est-à-dire celui qui maximise $W(\nu_k)$ tout en respectant les contraintes imposées [4.23] et [4.24].

En utilisant la formule de Stirling :

$$\log(n!) = n\log(n) - n + \sqrt{2\pi n} + \frac{1}{12n} + o\left(\frac{1}{n^2}\right) \qquad [4.26]$$

on voit que l'état le plus probable est celui qui maximise (toujours en respectant les contraintes [4.23] et [4.24]) :

$$log\left(W(\upsilon_k)\right) \approx -n\sum_{i=1}^{q} \frac{n_i}{n} log\left(\frac{n_i}{n}\right) \qquad [4.27]$$

On retrouve ainsi la forme de la fonction H.

Dans notre terminologie PaL, le principe de maximum d'entropie peut être vu comme l'exact équivalent du raisonnement de Boltzmann.

Supposons que dans notre système formel probabiliste, nous ayons q propositions mutuellement exclusives possibles (l'analogue des états microscopiques).

Supposons que nous ayons n expériences à notre disposition, une expérience consistant à constater que l'une des q propositions est vérifiée (l'analogue d'une particule étant dans l'état q).

Définissons δ_k, une description, comme une distribution de probabilité :

$$\left\{p_1,...,p_q\right\} = \left\{\frac{n_1}{n},...,\frac{n_q}{n}\right\} \qquad [4.28]$$

sur les q propositions mutuellement exclusives du système formel (l'analogue des états macroscopiques).

Nous devons, bien sur, avoir :

$$\sum_{i=1}^{q} p_i = 1 \qquad [4.29]$$

Nous pouvons de plus avoir des contraintes traduisant les connaissances préalables d'observations que nous avons sur le phénomène. Ces contraintes peuvent prendre des formes très diverses, le principe de maximum entropie étant toujours applicable. Par exemple, les connaissances préalables peuvent avoir la forme de m observables :

$$\forall j, j \in \{1,\ldots,m\}; \sum_{i=1}^{q} p_i f_j(i) = F_j \qquad [4.30]$$

Appelons $W(\delta_k)$ le nombre de manières de permuter les expériences en conservant les fréquences n_i/n correspondant à la description δ_k. Nous avons:

$$W(\delta_k) = \frac{n!}{n_1! \times n_2! \times \ldots \times n_q!} \qquad [4.31]$$

En utilisant la formule de Stirling, on obtient comme précédemment que la description la meilleure est celle qui correspond au plus grand nombre possible de permutations des expériences, c'est-à-dire celle qui maximise (en respectant les m+1 contraintes issues de [4.29] et [4.30]):

$$log\left(W(\delta_k)\right) \approx -n \sum_{i=1}^{q} p_i \, log(p_i) \qquad [4.32]$$

Reprenons l'exemple de l'hypothèse C1. La variable V peut prendre 6 valeurs entières entre 1 et 6. Nous avons fait 20 000 expériences dont nous avons choisi de ne mémoriser que la moyenne (observable $f_1(V)=V$) valant 3,5983.

Parmi les 6^{20000} séries d'expériences possibles, très peu, bien sûr, ont cette moyenne. Ces dernières peuvent être regroupées en « classes » (les descriptions), chaque classe étant caractérisée par la donné de 6 nombres (les p_i ou n_i) correspondant aux nombres d'apparitions de chacune des 6 valeurs possibles. A une classe donnée correspond $W(\delta_k)$ séries d'expériences possibles.

Numériquement, soit les trois classes :
- $\delta 1$: (3058, 3165, 3272, 3384, 3500, 3621)
- $\delta 2$: (3058, 3265, 3172, 3284, 3600, 3621)
- $\delta 3$: (0, 0, 8034, 11966, 0, 0)

ayant toutes les trois une moyenne de 3,5983, on calcule aisément (à partir de la formule [4.32]) que la première peut être réalisée par 10^{15549} séries d'expériences, la deuxième par 10^{15546} séries d'expériences et la troisième par 10^{5869} séries d'expériences. *Donc en tirant au hasard parmi les séries de 20000 expériences ayant 3,5983 comme moyenne (ce que nous pensons avoir fait en observant nos 20000 données) nous avons 1000 fois plus de chance de tirer une série d'expériences de la classe $\delta 1$ que de la classe $\delta 2$ et 10^{9680} fois plus de chance de tirer une série d'expériences de la classe $\delta 1$ que $\delta 3$. Comment, dès lors, ne pas considérer que la classe $\delta 1$ est une meilleure description que les 2 autres ?*

Les connaissances préalables définissent « l'instrument d'optique », le point de vue sous lequel on décide d'observer le phénomène étudié. Différentes connaissances préalables correspondent à différents points de vue et résultent en des distributions de maximum d'entropie, des descriptions différentes. Changer de connaissances préalables, c'est changer la forme de la paroi de la caverne de Platon sur laquelle se projettent les ombres portées de la « réalité », c'est changer inexorablement la perception intime du phénomène observé.

Il existe, bien entendu, des manières plus rigoureuses de justifier l'emploi du principe de maximum d'entropie, notamment les théorèmes, dits de « concentration d'entropie », tels que démontrés par Jaynes et par Robert. On se référera aux notes bibliographiques pour plus de précisions.

IV. NOTES BIBLIOGRAPHIQUES

La référence incontournable concernant ce paragraphe est le livre de Jaynes intitulé « Probability Theory: The Logic of Science » (Jaynes95). Cet ouvrage inachevé n'a pas encore été publié mais est accessible électroniquement depuis déjà plus de deux ans et s'enrichit régulièrement de nouvelles parties. Ce livre est le « chef-d'œuvre » de Jaynes qui fait la synthèse des travaux de toute une vie concernant le raisonnement probabiliste, les techniques de maximum d'entropie et leurs applications à la physique.

Un historique passionnant de cette approche peut être trouvé dans (Jaynes79).

Une analyse épistémologique des probabilités est donnée dans (Matalon67) précisant notamment la nature du débat entre « fréquentistes » et « subjectivistes ».

Concernant plus spécifiquement le théorème de Cox on se référera à (Cox46) et (Cox61). Une démonstration de ce théorème peut aussi être trouvée au chapitre 2 de (Jaynes95).

La logique, comme cas particulier des probabilités, est un sujet abordé par plusieurs auteurs, notamment, bien sur, par Jaynes dans (Jaynes95). (Cox79) est un article entièrement consacré à ce sujet.

Les problèmes directs et inverses sont largement développés dans les chapitres 3, 4, 6, 9 et 10 de (Jaynes95).

Pour une introduction à la notion d'entropie, on se référera bien évidemment à l'article fondateur de Shannon (Shannon49) qui introduisit le premier la notion d'entropie en dehors de la thermodynamique et de la mécanique statistique, ainsi qu'au passionnant livre de Campbell (Campbell82) qui présente bien les différents aspects de cette notion.

Les justifications théoriques de l'approche probabiliste et du principe de maximum d'entropie peuvent être trouvées, notamment, dans (Cox46), (Cox61), (Cox79), (deFinetti72), (Shore80), (Shore81), (VanCampenhout81), (Jaynes82), (Cheeseman85), (Shore86), (Hunter86) et (Robert91). (Jaynes82) et (Robert91) méritant une mention particulière puisqu'ils proposent des démonstrations des théorèmes de concentration d'entropie, respectivement dans les cas discret et continu.

Les outils mathématiques correspondant à ces techniques et les applications en physique sont amplement décrits dans (Levine79), (Smith85), (Tarentola87), (Erickson88a), (Erickson88b), (Bretthorst88), (Kapur92), (Mohammad-Djafari92) et évidemment (Jaynes95).

L'exemple du dé de Wolf est présenté et discuté en détail dans (Fougere88).

Des exemples d'applications à des problèmes de contrôle sont donnés par (MacKay92) ou (Neal93).

La description des « réseaux bayésiens », intéressante implantation informatique du raisonnement probabiliste, peut être trouvée dans (Pearl91) même si la dimension apprentissage en est presque totalement absente.

V. BIBLIOGRAPHIE

Bessière, P.; *LOLITA : Un langage de programmation logique intégré à LISP*; Cahier technique I.T.M.I., 1987.

Campbell Jerremy ; *Gramatical Man* ; Simon & Schuster, 1982

Cheeseman P. ; *In defense of Probability* ; Proceedings of AAAI85, 1985

Colmerauer A. ; *Theoretical model of PROLOG II* ; Logic Programming and its Applications; Ablex Publishing Corporation; 1986

Cox R. T. ; *Probability, Frequency, and Reasonable Expectation* ; American Journal of Physic, N°17 ; 1946

Cox R. T. ; *The algebra of probable inference* ; The John Hopkins Press, Baltimore, 1961

Cox R. T. ; *Of inference and inquiry, an essay in inductive logic ;* in *The maximum entropy formalism*, edited by Raphael D. Levine & Myron Tribus ; M.I.T. Press, 1979

de Finetti B. ; *Probability, induction and statistics* ; John Willey & sons, 1972

Erickson Gary J. & Smith C. Ray ; *Maximum-Entropy and Bayesian methods in science and engineering* ; *Volume 1 : Foundations* ; Kluwer Academic Publishers ; 1988

Erickson Gary J. & Smith C. Ray ; *Maximum-Entropy and Bayesian methods in science and engineering* ; *Volume 2 : Applications* ; Kluwer Academic Publishers ; 1988

Fougere P. F. ; *Maximum entropy calculations on a discrete probaility space* ; in *Maximum entropy and bayesian methods in science and engineering*, Vol. 1, edited by G.J. Erickson & C.R. Smith ; Kluewer Academic Publishers ; 1988

Hunter Daniel ; *Uncertain reasoning using Maximum entropy Inference* in *Uncertainty in Artificial Intelligence* ; edited by L. N. Kanal & J. F. Lemmer ; Elsevier Science Publishers, 1986

Jaynes E.T. ; *Where do we Stand on Maximum Entropy?* ; in *The maximum entropy formalism*, edited by Raphael D. Levine & Myron Tribus ; M.I.T. Press, 1979

Jaynes E. T. ; *On the rationale of maximum-entropy methods ;* Proceedings of the IEEE, 1982

Jaynes E.T. ; *Probability theory - The logic of science* ; à paraître. Version provisoire disponible à http://bayes.wustl.edu. ; 1995

Kapur J.N. & Kesavan H.K. ; *Entropy optimization principles with applications* ; Academic Press ; 1992

Levine Raphael D. & Tribus Myron ; *The maximum entropy formalism* ; MIT Press ; 1979

MacKay David J.C. ; *A practical bayesian framework for backprop networks* ; Neural computation ; 1992

Matalon Benjamin ; *Epistémologie des probabilités* ; dans *Logique et connaissance scientifique* sous la direction de Jean Piaget ; Encyclopédie de la Pléiade ; Editions Gallimard ; Paris, France ; 1967

Mohammad-Djafari A. & Demoment G. ; *Maximum entropy and bayesian methods* ; Kluwer Academic Publishers ; 1992

Neal Radford M. ; *Probabilistic inference using Markov chain Monte-Carlo Methods* ; Technical Report university of Toronto ; 1993

Pearl Judea ; *Probabilistic reasoning in intelligent systems : Networks of plausible inference* ; Morgan Kaufmann Publishers, San Mateo, California, USA ; 1991

Robert Claudine ; *An entropy concentration theorem: applications in artificial intelligence and descriptive statistics* ; Journal of Applied Probabilities ; september 1990

Robert Claudine ; *Modèles statistiques pour l'intelligence artificielle : l'exemple du diagnostic médical* ; Masson, Paris, France ; 1991

Robinson J.A. & Sibert E.E. ; *LOGLISP : an alternative to PROLOG*; Machine Intelligence; Vol. 10; 1983

Shannon Claude E. ; *The mathematical theory of communication* ; University of illinois, 1949

Shore J. E. & Johnson R. W. ; *Axiomatic derivation of the principle of Maximum Entropy and the Principle of Minimum Cross-Entropy ;* IEEE Transactions on Information Theory, 1980

Shore J. E. & Johnson R. W. ; *Properties of cross-entropy minimization ;* IEEE Transactions on Information Theory, 1981

Shore John E. ; *Relative entropy, probabilistic inference and A.I.* in *Uncertainty in Artificial Intelligence* ; edited by L. N. Kanal & J. F. Lemmer ; Elsevier Science Publishers, North-Holland, 1986

Smith C. Ray & Grandy, W. T. Jr. ; *Maximum-Entropy and bayesian methods in inverse problems* ; D. Reidel Publishing Company ; 1985

Tarentola Albert ; *Inverse Problem Theory ; Methods for data fitting and model parameters estimation* ; Elsevier ; New York, USA ; 1987

Van Campenhout J. M. & Cover T.M. ; *Maximum entropy and conditional probability ;* IEEE Transactions on Information Theory, 1981