Université de Paris

INÉGALITÉS EN THÉORIE CLASSIQUE ET ALGORITHMIQUE DE L'INFORMATION

Moana Laurent JUBERT

Co-encadré par :

Laurent BIENVENU Andrei ROMASHCHENKO

Master Logique et Fondements de l'Informatique

Septembre 2021

Résumé

Un résultat de Hammer et al. [4] a établi au début des années 2000 une équivalence entre les inégalités linéaires en théorie classique (entropie de Shannon) et algorithmique (complexité de Kolmogorov) de l'information. Parmi ces inégalités il en existe certaines dites non-classiques, ou non-Shannon, qui ne se déduisent pas d'un certain nombre d'inégalités de base (cf. Zhang & Yeung [12]). L'objectif de ce mémoire est de présenter les outils nécessaires pour comprendre les quelques résultats connus sur les inégalités linéaires qui seront présentés ensuite. Ce travail fait une synthèse partielle d'un stage de trois mois réalisé au Laboratoire Bordelais de Recherche en Informatique (LaBRI), co-encadré par Laurent Bienvenu et Andrei Romashchenko.

Table des matières

1 Espace de Cantor

Nous serons amenés à travailler avec des suites finies ou infinies de 0 et de 1 tout au long de ce mémoire. Il semble donc raisonnable de commencer par un rappel succint de quelques notations, ainsi que des propriétés de l'espace de Cantor.

NOTATION. Les suites finies de 0 et de 1, aussi appelées des *chaînes*, sont généralement notées par des lettres grecques minuscules ou avec les lettres $x, y, z, \ldots$ La chaîne vide est notée ε, et la concaténation de deux chaînes τ et σ est simplement notée $\tau\sigma$. Enfin, la longueur d'une chaîne τ est notée $|\tau|$.

L'ensemble des suites finies de 0 et de 1 est noté $2^{<\mathbb{N}}$, ou parfois $\{ 0, 1 \}^*$.

On note $\tau \preccurlyeq A$ le fait qu'une chaîne τ est un préfixe d'une suite A quelconque (finie ou infinie). Deux chaînes qui ne sont pas un préfixe de l'une ou l'autre sont dites *incompatibles*.

L'ensemble $[\tau] \overset{\text{def}}{=} \left\{ A \in 2^{\mathbb{N}} \;\middle|\; \tau \preccurlyeq A \right\}$ des suites infinies de 0 et de 1 ayant τ pour préfixe est appelé le *cylindre* de τ.

Supposons avoir deux chaînes τ et σ telles que $\tau \preccurlyeq \sigma$. Alors dans ce cas les suites commençant par σ commencent également par τ, c'est-à-dire que $[\sigma] \subseteq [\tau]$. Réciproquement si τ et σ sont incompatibles, alors $[\tau] \cap [\sigma] = \varnothing$. En effet, une suite qui commencerait par τ ne saurait commencer par σ et inversement. Lorsque deux suites sont contenues dans un même cylindre, elles sont en quelque sorte « voisines » : plus la longueur du préfixe qu'elles ont en commun est grande, plus le cylindre qui les contient toutes les deux est petit au sens de l'inclusion. Nous pouvons formaliser cette idée en considérant l'ensemble des cylindres comme la base d'ouverts d'une topologie sur les suites.

1.1 Définition

DÉFINITION 1. L'ensemble $2^{\mathbb{N}}$ des suites infinies de 0 et de 1 muni de la topologie engendrée par la base d'ouverts $\{ [\tau] \mid \tau \text{ est une chaîne} \}$ est appelé *espace de Cantor*.

L'espace de Cantor est aussi noté Ω. Comme il n'existe qu'une quantité dénombrable de chaînes — et donc de cylindres, chaque ouvert de Ω peut s'écrire comme une union au plus dénombrable de cylindres, que l'on peut par ailleurs supposer deux à deux disjoints. Cette propriété nous garantit que l'on peut construire une mesure sur Ω sans risquer de faire n'importe quoi :

FAIT. Il existe $\square$ une mesure sur Ω telle que $\square([\tau]) = 2^{-|\tau|}$ sur les cylindres.

Il s'agit de la *mesure uniforme* sur Ω. Puisque toutes les suites commencent par la chaîne vide, on a en particulier que $\square(\Omega) = \square([\varepsilon]) = 2^0 = 1$. Une interrogation qui reviendra sans cesse est de savoir comment *effectivement* trouver des ouverts de Ω ayant une mesure donnée, c'est-à-dire avoir procédé algorithmique qui énumère les cylindres qui composent ces ouverts. Le prochain lemme permet de répondre entièrement à cette interrogation :

Lemme 1 (Kraft-Chaitin). Supposons avoir une suite *énumérable* $n_0, n_1, \dots$ d'entiers naturels tels que :

$$\sum_{k}^{\infty} 2^{-n_k} \leqslant 1$$

Alors nous pouvons calculer *uniformément* une suite $\tau_0, \tau_1, \dots$ de chaînes *incompatibles entre elles deux à deux* telle que $|\tau_k| = n_k$.

Démonstration.. Les arguments suivants sont tirés du livre de Shen et al. [10, p. 94]. Par hypothèse, il existe une machine de Turing M qui est capable d'énumérer les entiers $n_0, n_1, \dots$ au fur et à mesure. Le résultat que nous allons montrer ici est que nous pouvons utiliser cette machine M dans une procédure effective qui va construire progressivement les chaînes $\tau_0, \tau_1, \dots$ ayant les propriétés voulues. À cette fin, on peut imaginer les $n_0, n_1, \dots$ comme une suite de requêtes que nous devons traiter successivement dans l'ordre. Nous allons définir une liste L_N des chaînes « disponibles » à chaque itération N, et satisfaisant les conditions suivantes :

i) Toutes les chaînes de L_N sont incompatibles entre elles et avec les $\tau_0, \dots, \tau_{N-1}$ déjà construits
ii) $2^{-n_0} + \cdots + 2^{-n_{N-1}} + \sum_{\sigma \in L_N} 2^{-|\sigma|} = 1$
iii) L_N est une liste finie contenant des chaînes de longueurs distinctes

On procède par récurrence : remarquons que $L_0 \overset{\text{def}}{=} \{ \varepsilon \}$ satisfait bien les trois conditions précédentes. Supposons maintenant être à une itération N quelconque : on a nécessairement qu'il existe des éléments de L_N de longueur inférieure à n_N. En effet si tel n'était pas le cas, alors nous en déduirions $\sum_{\sigma \in L_N} 2^{-|\sigma|} < 2^{-n_N}$ car tous les $\sigma \in L_N$ sont de longueurs distinctes strictement supérieures à n_N, et dans ce cas :

$$2^{-n_0} + \cdots + 2^{-n_{N-1}} + \sum_{\sigma \in L_N} 2^{-|\sigma|} < 1$$

ce qui est une contradiction. Considérons la plus longue chaîne $x^* \in L_N$ telle que $|x^*| \leqslant n_N$: celle-ci est nécessairement unique car deux éléments distincts de L_N ne sauraient avoir la même longueur. Par ailleurs, il est possible d'exhiber x^* de façon *effective* : la liste L_N est en effet finie, nous pouvons donc choisir l'élément de L_N ayant les propriétés souhaitées en parcourant simplement la liste. Il

y a deux cas possibles : soit $|x^*| = n_N$, et dans ce cas on pose $\tau_N \overset{\text{def}}{=} x^*$ qui convient clairement, de même que $L_{N+1} \overset{\text{def}}{=} L_N \setminus \{ x^* \}$ satisfaint encore les trois conditions; ou soit $|x^*| < n_N$, et dans ce cas on pose $\tau_N \overset{\text{def}}{=} x^*0\ldots0$ et $L_{N+1} \overset{\text{def}}{=} L_N \setminus \{ x^* \} \cup \{ x^*1, x^*01, \ldots, x^*0\ldots01 \}$ qui satisfait là encore les trois conditions. En effet par exemple, tous les $x^*1, x^*01, \ldots, x^*0\ldots01$ sont de longueur strictement supérieure à $|x^*|$ mais inférieure à n_N, et donc s'il existait une autre chaîne de L_N ayant la même longueur que l'un des $x^*1, x^*01, \ldots, x^*0\ldots01$ alors x^* ne serait pas de longueur maximale. Toutes les opérations décrites ci-dessus sont effectives, ce qui démontre bien le résultat voulu. ■

Le lemme 1 en oubliant le procédé algorithmique est également connu sous le nom d'*inégalité de Kraft*. C'est plutôt cette version qui va nous intéresser dans un premier temps, lorsque nous allons aborder l'entropie de Shannon. En voici une conséquence directe :

PROPOSITION 1. Pour tout réel r entre 0 et 1, il existe un ouvert de mesure r.

Démonstration.. Il suffit de regarder les décimales de r dans son écriture en base 2, et d'appliquer l'inégalité de Kraft : celle-ci nous fournit des chaînes incompatibles qui induisent une base d'ouverts dont l'union donne bien ce que l'on cherche. ■

1.2 L'espace de Cantor comme univers

Nous étudierons dans la prochaine section les variables aléatoires (abrégé v.a. par la suite) qui ne prennent qu'un nombre fini de valeurs. Il est d'usage de noter également Ω l'*univers* en théorie des probabilités, et de façon bien commode il se trouve que nous pouvons supposer sans perdre en généralité que cet univers Ω est *en effet* l'espace de Cantor dans le cas particulier des v.a. à valeurs finies.

LEMME 2. Étant donné des réels positifs $p_1, \ldots, p_N$ tels que $p_1 + \cdots + p_N \leqslant 1$, il existe des ouverts *disjoints* de l'espace de Cantor $\mathcal{U}_1, \ldots, \mathcal{U}_N$ tels que $\square(\mathcal{U}_k) = p_k$.

Démonstration.. C'est la même idée que dans la démonstration de la proposition 1, mais cette fois-ci avec N valeurs au lieu d'une seule. On regarde les décimales de l'écriture en base 2 de chaque p_k :

$$p_k = \sum_i^\infty 2^{-n_{k,i}}$$

Dans ce cas, nous pouvons appliquer l'inégalité de Kraft à l'ensemble des $n_{k,i}$ réunis :

$$\sum_k^N \sum_i^\infty 2^{-n_{k,i}} \leqslant 1$$

Notons $\tau_{k,i}$ les chaînes incompatibles obtenues. En posant alors :

$$\mathcal{U}_k = \bigcup_i^{\infty} [\tau_{k,i}]$$

il est clair que les $\mathcal{U}_k$ ainsi définis conviennent. ■

PROPOSITION 2. Pour toute v.a. X à valeurs parmi $1, \ldots, N$ dont le domaine de définition est un univers quelconque, il existe une v.a. Y identiquement distribuée à X dont le domaine de définition est l'espace de Cantor.

Démonstration.. Considérons n'importe quelle v.a. X à valeurs parmi $1, \ldots, N$. Alors dans ce cas nous avons $\mathbf{Pr}(X = 1) + \cdots + \mathbf{Pr}(X = N) \leqslant 1$, et par conséquent en appliquant le lemme 2 nous pouvons définir presque partout une application Y sur l'espace de Cantor Ω de telle sorte que $Y(\mathcal{U}_k) = \{ k \}$ avec $\square(\mathcal{U}_k) = \mathbf{Pr}(X = k)$. Par construction, Y est identiquement distribuée à X. ■

La notation Ω fera ainsi toujours référence à l'espace de Cantor dans ce mémoire. Il est désormais temps de regarder de plus près les propriétés des v.a. à valeurs finies, et en particulier ce que l'on appelle leur entropie.

2 Entropie de Shannon

L'ingrédient de base de la théorie classique de l'information sont donc les v.a. à valeurs finies. Pour alléger la notation, on écrira p_k au lieu de $\mathbf{Pr}(X = k)$ lorsque cela ne pose pas de problèmes.

2.1 Définition

Définition 2. Soit X une v.a. à valeurs parmi $1, \ldots, N$. Alors l'*entropie* de X, que l'on note $H(X)$, est définie de la façon suivante :

$$H(X) \stackrel{\text{def}}{=} \sum_{k}^{N} p_k \log \frac{1}{p_k}$$

Supposons avoir une deuxième v.a. Y à valeurs parmi $1, \ldots, M$. Alors l'entropie de la paire $\langle X, Y \rangle$ se calcule directement :

$$H(X, Y) = \sum_{k}^{N} \sum_{\ell}^{M} p_{k,\ell} \log \frac{1}{p_{k,\ell}} \quad \text{avec} \quad p_{k,\ell} = \mathbf{Pr}(X = k \wedge Y = \ell)$$

Que se passe-t-il si l'on sait, par exemple, que $X = 3$? De manière générale, nous serions tentés de calculer l'entropie de Y *sachant* l'issue de la v.a. X au préalable. Il est toujours possible de définir une nouvelle v.a. « $Y \mid X = k$ » en remplaçant les probabilités $\mathbf{Pr}(Y = \ell)$ par $\mathbf{Pr}(Y = \ell \mid X = k)$ dans la distribution de Y. La définition suivante a donc un sens :

Définition 3. L'entropie *conditionnelle* de Y sachant X, notée $H(Y \mid X)$, est définie comme l'espérance de l'entropie de Y sachant l'issue de la v.a. X :

$$H(Y \mid X) \stackrel{\text{def}}{=} \sum_{k}^{N} \mathbf{Pr}(X = k) \cdot H(Y \mid X = k)$$

Nous verrons dans un instant pourquoi l'intuition voudrait que plus l'entropie de Y est grande, plus il est difficile de « prédire » quelle sera l'issue de celle-ci. Si les entropies $H(Y)$ et $H(Y \mid X)$ quantifient en quelque sorte le niveau de « surprise » que l'on aurait à connaître l'issue de Y, alors $H(Y) - H(Y \mid X)$ contient la quantité de « surprise » *effacée* par la connaissance de l'issue de X, autrement dit d'une certaine manière le niveau de dépendance de Y par rapport à X. C'est ce que l'on appelle l'information mutuelle :

Définition 4. L'*information mutuelle* de X et Y, notée $I(X : Y)$, est définie par :

$$I(X : Y) \stackrel{\text{def}}{=} H(Y) - H(Y \mid X)$$

Cette définition a un sens : supposons que les v.a. X et Y sont indépendantes, dans ce cas $H(Y \mid X) = H(Y)$ et par conséquent $I(X : Y) = 0$, c'est-à-dire que X et Y ne partagent aucune information. Il se trouve que la réciproque est également vraie.

2.2 Propriétés de base

Fait. Si $I(X : Y) = 0$, alors X et Y sont indépendantes.

Par conséquent, l'information mutuelle observe exactement la dépendance de Y par rapport à X. La propriété est symétrique, c'est-à-dire que $I(Y : X) = 0$ implique que X et Y sont indépendantes. Cela peut sembler curieux puisque la définition de l'information mutuelle ne l'est pas. En fait, nous avons effectivement $I(X : Y) = I(Y : X)$ en vertu du théorème fondamental qui va suivre.

Théorème 1 (Symétrie de l'information). $H(X, Y) = H(X) + H(Y \mid X)$

Démonstration.. Il suffit de faire le calcul :

$$\begin{aligned}
H(X, Y) &= \sum_{k}^{N} \sum_{\ell}^{M} p_{k,\ell} \log \frac{1}{p_{k,\ell}} \\
&= \sum_{k}^{N} p_k \left[\sum_{\ell}^{M} \frac{p_{k,\ell}}{p_k} \left(\log \frac{1}{p_k} + \log \frac{p_k}{p_{k,\ell}} \right) \right] \\
&= \underbrace{\sum_{k}^{N} p_k \overbrace{\left(\sum_{\ell}^{M} \frac{p_{k,\ell}}{p_k} \right)}^{=1} \log \frac{1}{p_k}}_{= H(X)} + \underbrace{\sum_{k}^{N} p_k \overbrace{\left(\sum_{\ell}^{M} \frac{p_{k,\ell}}{p_k} \log \frac{p_k}{p_{k,\ell}} \right)}^{= H(Y \mid X = k)}}_{= H(Y \mid X)}
\end{aligned}$$

■

Corollaire. Les égalités suivantes sont vraies :

i) $H(Y \mid X) = H(X, Y) - H(X)$

ii) $I(X : Y) = H(X) + H(Y) - H(X, Y)$

iii) $H(X, Y) = H(X \mid Y) + H(Y \mid X) + I(X : Y)$

L'égalité ii) nous donne bien $I(X : Y) = I(Y : X)$. En particulier, l'égalité iii) nous autorise à penser l'entropie de la paire sous la forme d'un diagramme de Venn (cf. Figure 1).

Les trois notions que nous avons définies peuvent ainsi chacune se réécrire comme somme des deux autres. Cela va nous permettre de regarder uniquement les propriétés de l'entropie *simple* (i.e. non conditionnelle) afin de pouvoir directement en déduire des propriétés sur l'entropie conditionnelle et sur l'information mutuelle.

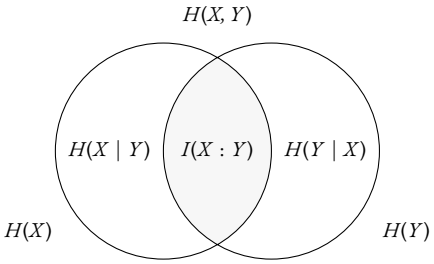

FIGURE 1 – Décomposition de $H(X, Y)$.

FAIT. L'entropie de X est positive. Elle est nulle si X prend une valeur presque sûrement.

Cela confirme un sens de notre intuition : lorsque $H(X)$ est proche de zéro, alors l'une des valeurs de X porte toute la masse de probabilité. Autrement dit, nous avons peu de chances de nous tromper en « prédisant » que X va effectivement prendre cette valeur. Reste à voir le sens réciproque, qui nécessite un résultat classique d'analyse :

LEMME 3 (Inégalité de Jensen). Si f est une fonction réelle *convexe*, alors pour tous réels $x_1, \dots, x_N$ et toute pondération $w_1, \dots, w_N$ l'inégalité suivante est vraie :

$$f\left(\frac{\sum_k^N w_k \cdot x_k}{\sum_k^N w_k}\right) \leqslant \frac{\sum_k^N w_k \cdot f(x_k)}{\sum_k^N w_k}$$

L'inégalité dans l'autre sens est vraie si f est *concave* au lieu d'être convexe. En particulier, si f est *strictement* convexe (resp. concave), alors le cas d'égalité n'est atteint que lorsque tous les $x_1, \dots, x_N$ sont égaux.

PROPOSITION 3. L'entropie de X est maximale si, et seulement si X suit une loi uniforme.

Démonstration.. D'une part, un simple calcul permet de déterminer que l'entropie d'une v.a. uniforme sur N valeurs est exactement égale à $\log N$. D'autre part, nous avons par définition :

$$H(X) = \sum_k^N p_k \log \frac{1}{p_k}$$

Le logarithme est concave, nous pouvons donc appliquer l'inégalité de Jensen :

$$\sum_k^N p_k \log \frac{1}{p_k} \leqslant \log \left(\sum_k^N p_k \frac{1}{p_k} \right) = \log N$$

Comme $\log N$ est atteint par les v.a. uniformes, cette inégalité doit être une égalité par hypothèse de maximalité sur $H(X)$. Or, le logarithme est même *strictement* concave, par conséquent le cas d'égalité ne peut être atteint que si tous les $\frac{1}{p_k}$ sont égaux, ce qu'il fallait démontrer. ■

Corollaire. L'entropie de X est inférieure à $\log N$.

La proposition 3 achève ainsi de compléter le sens réciproque de notre intuition : la difficulté de « prédire » l'issue de X est maximale exactement lorsqu'aucune de ses valeurs n'est plus probable que les autres.

Proposition 4. Pour tout réel positif r, il existe une v.a. dont l'entropie est exactement r.

Démonstration.. Choisissons un entier N tel que $r \leqslant \log N$. Alors l'ensemble :

$$\Delta_{\text{Top}}^{N-1} \overset{\text{def}}{=} \left\{ (x_1, \ldots, x_N) \in (\mathbb{R}^+)^N \;\middle|\; \sum_k^N x_k = 1 \right\}$$

est homéomorphe à la boule fermée de $\mathbb{R}^N$, en considérant par exemple la norme 1. En particulier, $\Delta_{\text{Top}}^{N-1}$ est connexe par arcs. Mais alors, l'application :

$$\begin{array}{ccc} \Delta_{\text{Top}}^{N-1} & \xrightarrow{H_N} & \mathbb{R}^+ \\ (x_1, \ldots, x_N) & \longmapsto & \displaystyle\sum_k^N x_k \log \frac{1}{x_k} \end{array}$$

est continue, et coïncide avec H sur les distributions des v.a. à valeurs parmi $1, \ldots, N$. Puisque H_N atteint les valeurs 0 et $\log N$ d'après les propriétés vues précédemment, la connexité par arcs implique nécessairement l'existence d'une distribution — et donc *a fortiori* d'une v.a. — pour laquelle l'entropie est exactement égale à r. ■

2.3 Codage d'un message de longueur ℓ

Shannon [8] montre dans son article fondateur que le nombre moyen de bits nécessaires pour encoder un message est déterminé par l'entropie. Cette section résume quelques résultats importants à ce sujet.

Définition 5. Étant donné un alphabet $\Sigma \overset{\text{def}}{=} \{ a_1, \ldots, a_N \}$, une application de Σ dans l'ensemble des suites finies $2^{<\mathbb{N}}$ est appelée un *code*. Si toutes les images sont incompatibles entre elles, alors on dit que le code est *préfixe*.

Exemple. L'application $a_k \longmapsto 0^k1$ est un code préfixe. La notation 0^k1 signifie k fois le symbole 0, suivi du symbole 1.

Les codes préfixes permettent d'encoder des messages sans introduire une ambiguïté : il n'y a exactement qu'une seule manière de décoder un message étant donné son code. Une conversation hypothétique entre Alice et Bob via un canal de transmission binaire peut être schématisée par la Figure 2 ci-dessous :

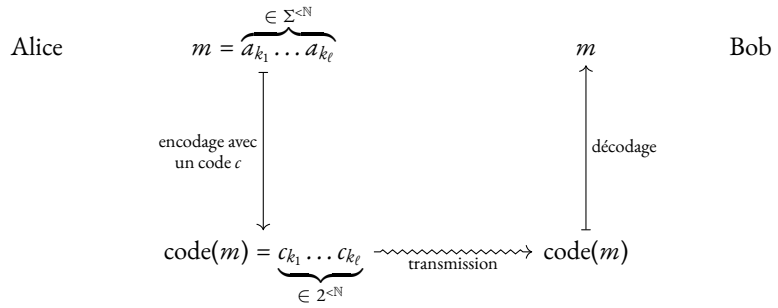

Figure 2 – Alice envoie un message de longueur ℓ à Bob.

Dans un souci d'économie, Alice et Bob souhaitent minimiser la longueur moyenne d'une transmission. Pour cela, nous allons supposer connaître par avance la v.a. X qui modélise le fait de regarder une lettre à n'importe quelle position fixée, c'est-à-dire que pour chaque position dans le message, $\mathbf{Pr}(X = k)$ est la probabilité d'observer la lettre a_k.

2.3.1 $\ell = 1$

Si un message ne consiste qu'en une seule lettre, cela revient simplement à vouloir trouver un code préfixe c qui minimise l'espérance $\mathbb{E}^{|c|}$ des longueurs de ses images :

$$\mathbb{E}^{|c|} \overset{\text{def}}{=} \sum_{k}^{N} \mathbf{Pr}(X = k) \cdot |c_k|$$

Proposition 5. Étant donné un code préfixe c quelconque, nous avons $H(X) \leqslant \mathbb{E}^{|c|}$.

Démonstration.. Toujours en notant p_k la probabilité $\mathbf{Pr}(X = k)$, il suffit de voir que :

$$H(X) - \mathbb{E}^{|c|} = \sum_k^N p_k \left(\log \frac{1}{p_k} - |c_k| \right) = \sum_k^N p_k \log \frac{2^{-|c_k|}}{p_k}$$

En appliquant l'inégalité de Jensen, il vient dans ce cas :

$$H(X) - \mathbb{E}^{|c|} \leqslant \log \left(\sum_k^N 2^{-|c_k|} \right)$$

Puisque c est un code préfixe, les cylindres $[c_k]$ sont tous disjoints et par conséquent la somme de leurs mesures est inférieure à la mesure de Ω, c'est-à-dire 1 :

$$\sum_k^N 2^{-|c_k|} \leqslant 1$$

On conclut dans ce cas par croissance du logarithme que $H(X) - \mathbb{E}^{|c|} \leqslant \log 1 = 0$, ce qu'il fallait démontrer. ■

Alice doit donc transmettre *au moins* $H(X)$ bits en moyenne afin que Bob soit capable de décoder son message. En réalité, il n'y a pas besoin de beaucoup plus :

PROPOSITION 6. Il existe un code préfixe c tel que $\mathbb{E}^{|c|} \leqslant H(X) + 1$.

Démonstration.. Considérons les entiers naturels $n_1, \ldots, n_N$ définis de telle sorte que :

$$\log \frac{1}{p_k} \leqslant n_k \overset{\text{def}}{=} \left\lceil \log \frac{1}{p_k} \right\rceil \leqslant \log \frac{1}{p_k} + 1$$

Ceci implique d'une part :

$$\sum_k^N 2^{-n_k} \leqslant \sum_k^N 2^{-\log \frac{1}{p_k}} = \sum_k^N p_k = 1$$

En appliquant l'inégalité de Kraft (voir le lemme 1), nous pouvons trouver un code préfixe c de telle sorte que $|c_k| = n_k$ pour tout k. Mais alors, d'autre part :

$$\mathbb{E}^{|c|} = \sum_k^N p_k |c_k| \leqslant \sum_k^N p_k \left(\log \frac{1}{p_k} + 1 \right) = H(X) + 1$$

■

2.3.2 $\ell > 1$

Pour les messages d'une longueur quelconque — et de surcroît en s'autorisant une erreur de décodage $\varepsilon_\ell > 0$ — la théorie nous emmènerait sans doute trop loin par rapport au cadre de

ce mémoire. Néanmoins, la proposition principale sera présentée ci-après sans démonstration, à titre indicatif.

Supposons avoir $X_1, \ldots, X_\ell$ indépendantes identiquement distribuées à X (abrégé i.i.d. par la suite), ainsi qu'une fonction d'encodage E_ℓ d'une part, et une fonction de décodage D_ℓ d'autre part. Notons m_X la v.a. $\langle X_1, \ldots, X_\ell \rangle$ qui prend ses valeurs parmi les messages de longueur ℓ. Pour simplifier, on suppose qu'il existe un « coefficient d'encodage » $\lambda > 0$ tel que la longueur d'un message codé est constante :

$$|E_\ell(m_X)| = \lceil \lambda \cdot \ell \rceil$$

L'*erreur de décodage* est dans ce cas définie par :

$$\varepsilon_\ell \overset{\text{def}}{=} \mathbf{Pr}(m_X \neq D_\ell E_\ell(m_X))$$

Fait (Shannon [8]).

i) Si $\lambda > H(X)$, alors pour ℓ suffisamment grand il existe E_ℓ et D_ℓ tels que l'erreur de décodage ε_ℓ est arbitrairement petite.

ii) Si $\lambda < H(X)$, alors l'erreur de décodage ε_ℓ tend vers 1 quelque soient les fonction d'encodage et de décodage considérées.

3 Complexité de Kolmogorov

Certaines chaînes semblent plus faciles à « décrire » que d'autres. Par exemple nous pourrions simplement dire « *douze fois zéro* » pour parler de la chaîne 000000000000 à quelqu'un. Au contraire, la chaîne 110100100101 semble à première vue difficile à communiquer oralement, autrement qu'en énonçant péniblement chacun des douze bits un par un. La théorie algorithmique de l'information étend cette idée aux descriptions *effectives*, c'est-à-dire aux machines de Turing capables de reconstruire une chaîne fixée. Le résultat spectaculaire de cette théorie est que la notion d'information qui en découle *ne dépend pas du modèle de calcul considéré*, autrement dit qu'il existe une bonne notion de complexité « intrinsèque » d'une chaîne. Malheureusement pour nous, cette complexité n'est en général pas calculable.

Remarque. Ce mémoire ne saurait évidemment pas être exhaustif sur toutes les propriétés intéressantes de toutes les variantes de la complexité de Kolmogorov. Pour une introduction plus détaillée (et sans doute plus limpide), voir le livre de Shen et al. [10].

3.1 Complexité pleine

Supposons avoir une chaîne x, ainsi qu'une machine de Turing D. Nous pouvons imaginer D comme étant un *décompresseur*, qui prend en entrée un fichier compressé et produit le fichier décompressé sur sa sortie. Dans ce cas nous serions peut-être tentés de proposer la définition suivante :

$$C_D(x) \stackrel{\text{def}}{=} \min \{\, |\sigma| \mid D(\sigma) = x \,\}$$

La *complexité* de x *relative* à la machine D est la longueur de la plus petite entrée σ sur laquelle D produit x. Bien entendu il se peut tout-à-fait que $C_D(x)$ n'est pas toujours définie, et dans ce cas on considère par convention que $C_D(x) = \infty$.

Notation. On note $\langle x, y \rangle$ n'importe quelle manière raisonnable d'encoder la paire des chaînes x et y, c'est-à-dire que l'application $\langle -, - \rangle$ est une bijection *calculable*.

La notation $\phi_e(\sigma)$ désigne l'exécution sur une entrée σ d'un programme ayant comme « code source » la chaîne e.

La théorie de la calculabilité nous garantit l'existence d'une machine *universelle*, notée U, capable de simuler *uniformément* n'importe quelle autre machine de Turing sur n'importe quelle entrée :

$$U(e, \sigma) = \phi_e(\sigma)$$

Puisqu'il existe toujours une machine de Turing qui ne fait rien d'autre qu'afficher x tout entier, l'ensemble $\{\, |\sigma| \mid U(\sigma) = x \,\}$ est toujours non vide, et par conséquent $C_U(x)$ est toujours

définie. Supposons maintenant qu'il existe une machine de Turing D « meilleure » que U pour une certaine chaîne x, c'est-à-dire que $C_D(x) \leqslant C_U(x)$. Formulé encore autrement, la machine D est capable de « mieux » compresser x que la machine universelle U. La machine D possède un « code source » e_D, de telle sorte que :

$$\phi_{e_D} = D \quad \text{et donc} \quad U(e_D, \sigma) = D(\sigma)$$

Mais alors, si x^* désigne *une* plus petite entrée sur laquelle D produit x, nous pouvons par exemple construire $\langle e_D, x^* \rangle$ en concaténant e_D et x^*, puis en réservant $O(\log |e_D|)$ bits de préfixe pour indiquer à la machine universelle où « couper » dans leur concaténation. C'est ce qu'illustre la Figure 3 ci-dessous :

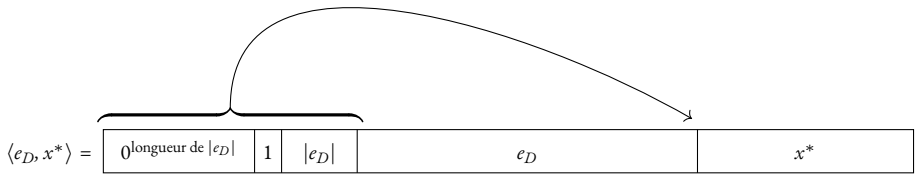

Figure 3 – Exemple de construction de $\langle e_D, x^* \rangle$.

Nous pouvons dans ce cas borner la longueur de la paire $\langle e_D, x^* \rangle$:

$$|\langle e_D, x^* \rangle| \leqslant \underbrace{O(\log |e_D|) + |e_D|}_{\text{ne dépend pas de } x} + |x^*|$$

En particulier, le terme à gauche *ne dépend que de D*. Ainsi, nous avons finalement l'inégalité :

$$C_U \leqslant C_D + O(1)$$

La machine U n'est par conséquent pas seulement *universelle*, elle est également *optimale* : elle admet des descriptions au moins aussi courtes que n'importe quelle autre machine de Turing *à une constante près*. C'est à ce titre que la définition de complexité de Kolmogorov qui va suivre ne dépend finalement pas du modèle de calcul considéré : nous allons pouvoir en particulier parler de *la* complexité de Kolmogorov.

3.1.1 Définition

Définition 6. La *complexité de Kolmogorov pleine* d'une chaîne x est définie par :

$$C(x) \stackrel{\text{def}}{=} \min \{ |\sigma| \mid U(\sigma) = x \}$$

À l'instar de l'entropie, la définition se relativise à la connaissance d'une autre chaîne :

Définition 7. La complexité pleine de y sachant x est définie par :

$$C(y \mid x) \stackrel{\text{def}}{=} \min \{ |\sigma| \mid U^x(\sigma) = y \}$$

où U^x désigne la machine universelle U étant donné l'oracle x.

Remarque. Disposer d'un *oracle* signifie en quelque sorte qu'une machine de Turing se donne une deuxième bande de lecture en entrée contenant tous les bits d'une suite finie ou infinie fixée à l'avance. Les oracles joueront un rôle important dans une prochaine section sur les suites aléatoires.

3.1.2 Présentation axiomatique

La complexité pleine possède également une caractérisation axiomatique que nous allons voir là encore sans démonstration, simplement à titre indicatif.

Définition 8. Une fonction f définie sur les entiers est *upper semi-computable* (abrégé u.s.-c. par la suite) si elle est « effectivement approchable par le dessus », c'est-à-dire qu'il existe une fonction totale calculable u à deux paramètres telle que pour tout entier x, d'une part $u(x, -)$ est décroissante, et d'autre part :

$$\lim_n u(x, n) = f(x)$$

Inversement, la fonction f est *lower semi-computable* (l.s.-c.) si elle est « effectivement approchable par le dessous ».

Exemple. La complexité de Kolmogorov est u.s.-c. puisqu'en exécutant tous les programmes possibles en parallèle, nous découvrons au fur et à mesure des descriptions de plus en plus courtes à mesure que les exécutions se terminent.

Bien sûr, nous aurions pu choisir f comme étant définie sur les chaînes et le résultat aurait été le même, puisqu'il est toujours possible de représenter les entiers par des chaînes et

inversement. Être u.s.-c., accompagné de deux autres axiomes, suffit dans ce cas à déterminer complètement la complexité de Kolmogorov pleine à constante près :

Fait (Shen [9], [10, p. 19-20]). Supposons avoir une fonction k définie sur les *chaînes* ayant les trois propriétés suivantes :

i) k est u.s.-c.

ii) Pour toute fonction partielle calculable f, il existe une constante c_f telle que :

$$k(f(x)) \leqslant k(x) + c_f$$

iii) Il existe deux constantes c_1 et c_2 telles que pour tout N :

$$2^{N-c_1} \leqslant |\{\, x \mid k(x) < N \,\}| \leqslant 2^{N+c_2}$$

Alors $k = C + O(1)$.

3.1.3 Vers la complexité préfixe

Pour décrire une paire de chaînes, il semble bien suffisant de se donner une description pour chacune des chaînes et de les assembler ensemble. Nous avons vu précédemment que l'on pouvait se contenter d'ajouter $O(\log |x|)$ bits de préfixe pour construire la paire $\langle x, y \rangle$ sans ambiguïté : il est malheureusement impossible de se débarrasser de ce facteur supplémentaire dans le cas de la complexité *pleine*, ce que nous allons voir maintenant. Une première observation triviale, mais non moins importante, est que la complexité d'une chaîne n'excède pas sa longueur :

Lemme 4. $C(x) \leqslant |x| + O(1)$

Une deuxième observation porte sur le *nombre* de chaînes dont la complexité est bornée par une constante. Par un argument combinatoire ordinaire, on peut voir assez facilement que ce nombre est majoré :

Lemme 5. $|\{\, x \mid C(x) < N \,\}| < 2^N$

Démonstration.. Il y a au plus $1 + 2 + 2^2 + \cdots + 2^{N-1} = 2^N - 1$ descriptions de longueur strictement inférieure à N, il ne peut donc y avoir que strictement moins de 2^N chaînes ayant une description de longueur strictement inférieure à N. ■

Ces deux observations nous permettent de faire la conclusion suivante :

Proposition 7. Il n'existe pas de constante d telle que pour toutes chaînes x et y nous ayons $C(x, y) \leqslant C(x) + C(y) + d$.

Démonstration.. Supposons par l'absurde que la proposition est fausse et qu'il existe une telle constante d. En appliquant notre première observation, il existe une autre constante c telle que :

$$C(x,y) \leqslant C(x) + C(y) + d < |x| + |y| + 2 \cdot c + d + 1 \tag{1}$$

Faisons l'hypothèse supplémentaire que $|x| + |y| = N$ est fixé. Alors, d'après notre deuxième observation :

$$|A| \stackrel{\text{def}}{=} |\{\, \langle x,y \rangle \mid C(x,y) < N + 2 \cdot c + d + 1 \,\}| < 2^{N+O(1)}$$

Un simple argument combinatoire nous permet de voir qu'il existe $N \cdot 2^{N+O(1)}$ paires $\langle x,y \rangle$ telles que $|x| + |y| = N$. Or toutes ces paires doivent appartenir à l'ensemble A en vertu de l'inégalité (1). On conclut dans ce cas :

$$N \cdot 2^{N+O(1)} \leqslant |A| < 2^{N+O(1)}$$

ce qui est une contradiction, pour N assez grand. ■

PROPOSITION 8. $C(x,y) \leqslant C(x) + C(y) + O(\log |x|)$

Démonstration.. Comme nous l'avons vu, $O(\log C(x))$ bits supplémentaires suffisent pour encoder la paire des plus petites descriptions de x et y, et donc en particulier $O(\log |x|)$ bits suffisent amplement. ■

La complexité de Kolmogorov *préfixe*, que nous allons voir juste après, va nous permettre de faire disparaître ce facteur supplémentaire. Avant toute chose, remarquons que la proposition 8 peut être raffinée de la manière suivante :

FAIT. $C(x,y) \leqslant C(x) + C(y \mid x) + O(\log |x|)$

En effet, il suffit de connaître une plus petite description de y sachant x et de la combiner avec une description de x pour obtenir une description de $\langle x,y \rangle$. Sous cette forme, l'inégalité pourrait nous rappeler le théorème de symétrie de l'information en théorie classique, et c'est bien normal : la réciproque de cette inégalité est vraie.

THÉORÈME 2 (Kolmogorov-Levin). $C(x,y) = C(x) + C(y \mid x) + O(\log |x| + |y|)$

Démonstration.. Nous allons reproduire ici la preuve détaillée dans [10, p. 37-39]. En posant $N \stackrel{\text{def}}{=} |x| + |y|$ il nous suffit de montrer que :

$$C(x) + C(y \mid x) \leqslant C(x,y) + O(\log N)$$

Notons A_τ l'ensemble des chaînes σ telles que la complexité de la paire $\langle \tau, \sigma \rangle$ est bornée par $C(x, y)$. La réunion de tous ces ensembles — notée A — ne peut pas avoir un cardinal excessivement grand d'après ce que nous avons vu au lemme 5 :

$$|A| \stackrel{\text{def}}{=} \left| \bigcup_{\tau} A_\tau \right| < 2^{C(x,y)+O(1)}$$

Il n'y a pas besoin de plus de $\log N$ bits pour décrire la chaîne « $C(x, y)$ », qui *n'est pas* le plus petit programme qui décrit la paire $\langle x, y \rangle$ mais bien *l'information de la longueur* d'un tel plus petit programme. Dans ce cas, nous pouvons énumérer les éléments de A en connaissant $O(\log N)$ bits : il suffit d'exécuter tous les programmes en parallèle, et d'afficher chaque paire générée par un programme de longueur inférieure à $C(x, y)$ à mesure que les exécutions se terminent. *A fortiori* nous pouvons donc énumérer les éléments de A_x, et par conséquent nous en déduisons d'une part :

$$C(y \mid x) \leqslant \log |A_x| + O(\log N) \tag{2}$$

En effet, il suffit de donner la position de y dans l'énumération de A_x pour obtenir une description de y : cela ne nécessite au plus que $\log |A_x| + O(\log N)$ bits.

D'autre part nous avons que $\log |A_x| < C(x, y) + O(1)$, et donc il n'y a pas besoin de plus que $O(\log N)$ bits pour décrire la chaîne « $\log |A_x|$ ». Nous pouvons alors énumérer toutes les chaînes τ telles que $\log |A_\tau| \geqslant \log |A_x|$ avec seulement $O(\log N)$ bits : il suffit de compter le nombre d'éléments dans chaque A_τ énuméré en parallèle, et de regarder lorsque le cardinal dépasse $|A_x|$. Il ne peut pas exister plus de $2^{C(x,y)-\log |A_x|+O(1)}$ telles chaînes par un argument analogue à l'inégalité de Markov, et donc nous pouvons décrire x simplement en donnant sa position dans cette énumération, c'est-à-dire finalement :

$$C(x) \leqslant C(x, y) - \log |A_x| + O(\log N) \tag{3}$$

En additionnant les inégalités (2) et (3), nous avons bien le résultat voulu. ■

3.2 Complexité préfixe

3.2.1 Machine de Turing préfixe

Si une chaîne τ et un préfixe de σ, on dit que σ est une *extension* de τ. Lorsque τ et σ ne sont pas égales, on dit de plus que τ est un préfix *strict* (resp. σ est une extension *stricte*).

Définition 9. Une machine de Turing M est *préfixe* si pour toute chaîne σ sur laquelle la machine M s'arrête, M ne s'arrête sur aucun préfixe ni aucune extension stricte de σ.

Exemple. Étant donné une machine M, nous pouvons toujours construire une machine préfixe P^M ayant la propriété suivante :

$$P^M(0^{|\sigma|}1\sigma) = M(\sigma)$$

Dans ce cas il est clair que pour tous τ, σ distincts, les chaînes $0^{|\tau|}1\tau$ et $0^{|\sigma|}1\sigma$ ne sont ni préfixes, ni extensions de l'une ou l'autre.

La chaîne $0^{|\sigma|}1\sigma$ est également notée $\overline{\sigma}$. Nous avons la propriété suivante :

Fait. $\overline{\tau}$ est un préfixe de $\overline{\sigma}$ si, et seulement si $\tau = \sigma$.

La construction précédente n'est évidemment pas la seule. En voici une autre qui nous intéressera tout particulièrement :

Définition 10. Étant donné une machine de Turing M, la machine $\widetilde{M}$ est définie de la manière suivante :

i) Exécuter M sur toutes les entrées σ possibles
ii) Soit σ une chaîne. Si parmi tous les préfixes et et toutes les extensions de σ, la machine M termine son exécution sur σ *en premier*, alors $\widetilde{M}(\sigma) = M(\sigma)$
iii) Sinon, $\widetilde{M}(\sigma)$ ne s'arrête jamais

Il n'est pas difficile de se convaincre que cette définition donne bien des machines de Turing préfixes. En particulier si M est déjà préfixe, alors $\widetilde{M} = M$. Il nous reste à voir le cas des machines universelles :

Fait. Il existe une machine universelle pour les machines préfixes, notée $\mathbf{U}$, définie de telle sorte que :

$$\mathbf{U}(0^e1\sigma) = \widetilde{\phi_e}(\sigma)$$

Les définitions de la complexité pleine peuvent ainsi s'étendre naturellement aux machines de Turing préfixes.

3.2.2 Définition

Définition 11. La *complexité de Kolmogorov préfixe* d'une chaîne x est définie par :

$$K(x) \stackrel{\text{def}}{=} \min \{\, |\sigma| \mid \mathbf{U}(\sigma) = x \,\}$$

Là encore, la complexité préfixe se relativise :

Définition 12. La complexité préfixe de y sachant x est définie par :

$$K(y \mid x) \stackrel{\text{def}}{=} \min \{ |\sigma| \mid \mathbf{U}^x(\sigma) = y \}$$

où $\mathbf{U}^x$ désigne la machine universelle $\mathbf{U}$ étant donné l'oracle x.

Remarque. Pour tous x et y distincts les machines $\mathbf{U}^x$ et $\mathbf{U}^y$ sont différentes, et par conséquent il peut tout-à-faix exister $\tau \preccurlyeq \sigma$ distincts tels que $\mathbf{U}^x(\tau)$ et $\mathbf{U}^y(\sigma)$ s'arrêtent.

3.2.3 Lien avec la complexité pleine

Nous allons montrer que la complexité pleine et la complexité préfixe sont égales à un terme logarithmique près.

Théorème 3. Pour toute chaîne x, nous avons $K(x) = C(x) + O(\log |x|)$. De plus, cette borne est *optimale*.

Nous savons déjà que $C(x) \leqslant K(x) + O(1)$ par optimalité de C sur *toutes* les machines de Turing, et donc en particulier sur celles qui sont préfixes. Le reste de la démonstration du théorème 3 est divisé en trois lemmes :

Lemme 6. $K(x) \leqslant C(x) + O(\log |x|)$

Démonstration.. Nous pouvons construire une machine de Turing préfixe M telle que :

$$M(\overline{|\sigma|}\sigma) = U(\sigma)$$

En notant e_M un code de M et x^* une plus petite description de x pour la machine universelle « non préfixe » U, on a alors :

$$\mathbf{U}(0^{e_M}1\overline{C(x)}x^*) = \widetilde{M}(\overline{C(x)}x^*) = M(\overline{C(x)}x^*) = U(x^*) = x$$

où $C(x) = |x^*|$ par définition. Mais dans ce cas $0^{e_M}1\overline{C(x)}x^*$ est une description de x de longueur $C(x) + O(\log |x|)$, ce que l'on cherchait. ∎

Nous en déduisons bien que $K(x) = C(x) + O(\log |x|)$. Il nous reste à voir l'optimalité du logarithme : pour cela, nous avons besoin de montrer qu'au contraire de la complexité pleine, la complexité préfixe se comporte « bien » avec les paires.

Lemme 7. $K(x, y) \leqslant K(x) + K(y) + O(1)$

Démonstration.. Il existe une machine de Turing préfixe M qui effectue les opérations suivantes sur une entrée σ :

i) Énumérer toutes les chaînes s, t telles que $st = \sigma$

ii) Exécuter en parallèle $\mathbf{U}(s)$ et $\mathbf{U}(t)$

iii) S'il existe deux telles chaînes s, t sur lesquelles $\mathbf{U}$ termine sur chacune d'entre elles, alors ces deux chaînes sont uniques par propriété de $\mathbf{U}$ d'être préfixe : dans ce cas, on affiche la chaîne $\langle \mathbf{U}(s), \mathbf{U}(t) \rangle$

Notons e_M un code de M, ainsi que x^* et y^* des plus petits programmes *préfixes* produisant respectivement x et y. Alors dans ce cas :

$$\mathbf{U}(0^{e_M}1x^*y^*) = \widetilde{M}(x^*y^*) = M(x^*y^*) = \langle x, y \rangle$$

ce qui permet de conclure. ∎

Lemme 8. Il n'existe pas de fonction ε telle que $K(x) = C(x) + O(\varepsilon(|x|))$ et qui serait négligeable devant le logarithme.

Démonstration.. On va montrer un raffinement de la proposition 7. Par l'absurde, s'il existait une telle fonction ε, alors en réinjectant le résultat du lemme précédent, nous en déduirions :

$$C(x, y) \leqslant C(x) + C(y) + O(\varepsilon(|x| + |y|))$$

En supposans $|x| + |y| = N$ fixé, nous pouvons dérouler un raisonnement analogue à la démonstration de la proposition 7 et ainsi obtenir l'inégalité :

$$N \cdot 2^{N+O(1)} \leqslant \cdots < 2^{\varepsilon(N)} \cdot 2^{N+O(1)}$$

Mais alors, en prenant la racine $\log N$-ième au voisinage de l'infini, l'inégalité devient :

$$2 \cdot 2^{\frac{N}{\log N}} \leqslant \cdots < 2^{\frac{\varepsilon(N)}{\log N}} \cdot 2^{\frac{N}{\log N}} \approx 1 \cdot 2^{\frac{N}{\log N}}$$

ce qui est une contradiction. ∎

Le théorème 3 est ainsi démontré dans sa totalité.

3.3 Correspondance des inégalités

Corollaire. Les inégalités linéaires entre les complexités pleines et préfixes qui sont vraies *à un facteur logarithmique près* sont les mêmes.

Puisqu'une égalité n'est jamais qu'une double inégalité, nous avons par exemple automatiquement le théorème de Kolmogorov-Levin pour la complexité préfixe :

$$K(x, y) = K(x) + K(y \mid x) + O(\log |x| + |y|)$$

Remarque. Il est même possible d'avoir plus finement les (in)égalités suivantes :

i) $K(x, y) \leqslant K(x) + K(y \mid x) + O(1)$

ii) $K(x, y) = K(x) + K(y \mid x, K(x)) + O(1)$

La démonstration est disponible dans le livre de Shen et al. [10, p. 105-108] mais fait appel aux *semi-mesures*, notion que nous ne verrons que dans une section ultérieure consacrée aux suites aléatoires.

La situation est différente pour les inégalités vraies *à constante près*. Nous avons vu par exemple que l'inégalité $K(x, y) \leqslant K(x) + K(y) + O(1)$ est vraie (cf. lemme 7), alors que ce n'est pas le cas de la complexité pleine (cf. proposition 7). Dans l'autre sens, il y a entre autres l'inégalité $C(x) \leqslant |x| + O(1)$ qui est trivialement vraie, alors que ce n'est plus le cas de la complexité préfixe :

Proposition 9. Il n'existe pas de constante d telle que $K(x) \leqslant |x| + d$.

Démonstration.. Car sinon nous aurions :

$$K(x, y) \leqslant K(x) + K(y) + O(1) \leqslant |x| + |y| + O(1)$$

et toujours par un raisonnement analogue à la démonstration de la proposition 7 nous obtiendrions une contradiction. ■

En fait, cette correspondance se prolonge à l'entropie de Shannon : les inégalités linéaires entre la complexité de Kolmogorov (à un facteur logarithmique près) et l'entropie de Shannon sont les mêmes. Ce résultat central est abordé dans la prochaine section.

4 Théorème d'équivalence

Nous allons redémontrer le théorème principal de Hammer et al. [4], qui établit donc une équivalence entre les inégalites linéaires en théorie classique et en théorie algorithmique de l'information à un facteur logarithmique près : nous pourrons ainsi à l'issue de cette section discuter des inégalités d'information en toute généralité.

Fait. Les égalités suivantes sont vraies :

$$H(X, X) = H(X) \qquad K(x, x) = K(x) + O(1)$$
$$H(X, Y) = H(Y, X) \qquad K(x, y) = K(y, x) + O(1)$$

Supposons avoir un entier L fixé. Alors nous pouvons dans ce cas nous ramener à étudier les L-uplets de v.a. ou de chaînes qui n'ont pas de doublons, et ce à permutation près : les objets mathématiques par excellence qui ont ces propriétés sont les ensembles de cardinal L. La notation ci-dessous capture cette idée, et sera incontournable pour la suite :

Notation. On se donne L objets $T_1, \ldots, T_L$ qui sont des v.a. ou des chaînes. Alors pour tout sous-ensemble $E \overset{\text{def}}{=} \{ i_1, \ldots, i_\ell \}$ parmi les indices $1, \ldots, L$, on définit T_E comme le ℓ-uplet des T_k correspondant aux éléments de E :

$$T_E \overset{\text{def}}{=} \langle T_{i_1}, \ldots, T_{i_\ell} \rangle$$

En particulier, cette notation n'induit aucune ambiguïté d'après le fait précédent.

Il est désormais possible d'énoncer précisément ce que sont les inégalités d'information :

Définition 13. Une inégalité d'information pour l'entropie de Shannon est la donnée d'une famille (a_E) de réels indicée par les parties E de $\{ 1, \ldots, L \}$, de telle sorte que pour toutes v.a. $X_1, \ldots, X_L$ on a :

$$\sum_{E \subseteq \{ 1, \ldots, L \}} a_E \cdot H(X_E) \geqslant 0$$

Similairement, une inégalité d'information pour la complexité de Kolmogorov est définie comme la donnée d'une famille (a_E) de réels telle que :

$$\sum_{E \subseteq \{ 1, \ldots, L \}} a_E \cdot K(x_E) \geqslant -O(\log N)$$

pour toutes chaînes $x_1, \ldots, x_L$ avec $N = |x_1| + \cdots + |x_L|$.

Remarque. Si nous pensons aux $H(X_E)$ ou aux $K(x_E)$ comme des coordonnées de vecteurs dans le $\mathbb{R}$-espace vectoriel de dimension $2^L - 1$ (la coordonnée $H(X_\varnothing)$ ou $K(x_\varnothing)$

est toujours égale à zéro), alors les inégalités d'information sont exactement les équations d'hyperplans qui délimitent la région contenant tous ces vecteurs. Cette région est usuellement notée Γ_L^* et les élément de cette région sont appelés les vecteurs *entropiques*. Une étude approfondie des propriétés de Γ_L^* peut être trouvée dans le livre de Yeung [11, p. 325]. Nous reviendrons sur cette vision lorsque nous étudierons les inégalités d'information en tant que telles dans la prochaine section.

Théorème 4 (Hammer et al. [4]). Les inégalités linéaires qui sont vraies pour la complexité de Kolmogorov sont également vraies pour l'entropie de Shannon, et vice versa.

La démonstration du théorème est divisée en deux parties.

4.1 Kolmogorov implique Shannon

Supposons avoir l'inégalité ci-dessous vraie pour n'importe quelles chaînes $x_1, \dots, x_L$:

$$\sum_{E \subseteq \{\,1,\dots,L\,\}} a_E \cdot K(x_E) \geqslant -O(\log N)$$

Considérons des v.a. $X_1, \dots, X_L$ quelconques, et notons $X \stackrel{\text{def}}{=} \langle X_1, \dots, X_L \rangle$. Sans perdre en généralité, nous pouvons supposer que les X_k prennent leurs valeurs parmi des chaînes d'une longueur c fixée à l'avance : cela signifie entre autres que l'on peut écrire $K(X_k)$ pour la complexité de Kolmogorov d'une valeur prise par la v.a. X_k. La prochaine étape consiste à regarder M copies i.i.d. de X, notées $X^1, \dots, X^M$, et que l'on peut décomposer comme suit :

$$\begin{aligned} X^1 &= \langle X_1^1, \dots, X_L^1 \rangle \\ X^2 &= \langle X_1^2, \dots, X_L^2 \rangle \\ &\vdots \\ X^M &= \langle X_1^M, \dots, X_L^M \rangle \end{aligned}$$

En particulier, les $X_k^1, \dots, X_k^M$ sont toutes des copies i.i.d. de chaque X_k. Définissons maintenant pour chaque k une v.a. $Y_k \stackrel{\text{def}}{=} \langle X_k^1, \dots, X_k^M \rangle$: cela nous donne une famille $Y_1, \dots, Y_L$ des « M-puissances i.i.d. » de chaque X_k, où les Y_k prennent leurs valeurs cette fois-ci parmi des chaînes de longueur $M \cdot c$. Pour rappel, les notations X_E^k et Y_E signifient les ℓ-uplets suivants :

$$X_E^k = \langle X_{i_1}^k, \dots, X_{i_\ell}^k \rangle \quad \text{et} \quad Y_E = \langle Y_{i_1}, \dots, Y_{i_\ell} \rangle \quad \text{avec} \quad E = \{\, i_1, \dots, i_\ell \,\}$$

Lemme 9. $K(Y_E) = K(X_E^1, \dots, X_E^M) + O(\log N)$

Démonstration.. Le lemme dit essentiellement que l'on peut réordonner les issues des X_i^k à un coût moindre. En effet, nous pouvons supposer avoir connaissance de c et E au préalable, qui sont des constantes qui ne dépendent pas des $X_1, \dots, X_L$ considérés initialement. Dans ce cas, il existe clairement un algorithme qui réordonne les éléments de $X_E^1, \dots, X_E^M$ pour donner Y_E en ne nécessitant que les bits pour décrire c et E. Puisque $N = |X_1| + \cdots + |X_L| = L \cdot c$, il suffit bien d'un facteur $O(\log N)$. ∎

Lemme 10. $K(X_E^1, \dots, X_E^M) = K(X_E^1 \dots X_E^M) + O(\log N)$

Démonstration.. Il faut remarquer que $K(X_E^1 \dots X_E^M)$ est la complexité de la *concaténation* des $X_E^1, \dots, X_E^M$ et non du *M-uplet* $\langle X_E^1, \dots, X_E^M \rangle$: le lemme n'est donc pas une tautologie. Observons déjà que chaque X_E^k prend pour valeur une chaîne de longueur $|E| \cdot c$ constante. En supposant c et E connus, nous pouvons retrouver chaque X_E^k individuellement de la concaténation $X_E^1 \dots X_E^M$ en découpant par tranches de longueur $|E| \cdot c$: c'est ici que nous sert le fait que toutes les v.a. $X_1, \dots, X_L$ prennent pour valeur des chaînes d'une longueur c fixée. Là encore, il suffit de se donner $O(\log N)$ bits pour connaître c et E comme dans le lemme précédent. ∎

Nous avons besoin d'un dernier lemme pour conclure, dont le lecteur peut aller consulter la démonstration dans l'article référencé :

Lemme 11 (Zvonkin & Levin [13, équation 5.18]). La complexité moyenne des X_E^k converge presque sûrement vers l'entropie de X_E :

$$\lim_M \frac{K(X_E^1 \dots X_E^M)}{M} = H(X_E)$$

Passons à la conclusion. L'inégalité suivante est vraie par hypothèse :

$$\sum_E a_E \cdot K(Y_E) \geqslant -O(\log M + \log N)$$

D'après les deux premiers lemmes, cela est équivalent à :

$$\sum_E a_E \cdot K(X_E^1 \dots X_E^M) \geqslant -O(\log M + \log N)$$

puisqu'en effet les facteurs $O(\log N)$ ont été « absorbés » par $O(\log M + \log N)$. Mais alors comme N ne dépend pas de M, en divisant par M de part et d'autre et en passant à la limite nous déduisons finalement du lemme de Zvonkin & Levin le résultat voulu :

$$\begin{aligned} \lim_M \sum_E a_E \cdot \frac{K(X_E^1 \dots X_E^M)}{M} &= \sum_E a_E \cdot H(X_E) \\ &\geqslant \lim_M -\frac{O(\log M + \log N)}{M} = 0 \end{aligned}$$

∎

4.2 Shannon implique Kolmogorov

Supposons maintenant que n'importe quelles v.a. $X_1, \ldots, X_L$ vérifient l'inégalité :

$$\sum_{E \subseteq \{\, 1,\ldots,L \,\}} a_E \cdot H(X_E) \geqslant 0$$

Étant donné L chaînes $x_1, \ldots, x_L$, l'idée est de construire des v.a. dont l'entropie est suffisamment proche (c'est-à-dire à un facteur logarithmique près) de la complexité de Kolmogorov des $x_1, \ldots, x_L$. Pour cela, nous allons soigneusement construire l'espace dans lequel ces v.a. vont prendre leurs valeurs.

Définition 14. On définit A comme l'ensemble des L-uplets de chaînes dont les complexités conditionnelles sont *toutes* bornées par les complexités conditionnelles des chaînes $x_1, \ldots, x_L$ que nous nous sommes données :

$$A \stackrel{\text{def}}{=} \{\, \langle \sigma_1, \ldots, \sigma_L \rangle \mid K(\sigma_F \mid \sigma_E) \leqslant K(x_F \mid x_E) \text{ pour tous } E, F \subseteq \{\, 1, \ldots, L \,\} \,\}$$

La notation A_E désigne dans ce cas l'ensemble des projections $\{\, \sigma_E \mid \langle \sigma_1, \ldots, \sigma_L \rangle \in A \,\}$.

L'ensemble A est non vide puisqu'il contient au moins le L-uplet $\langle x_1, \ldots, x_L \rangle$. En posant maintenant $N = |x_1| + \cdots + |x_L|$ nous avons la propriété suivante :

Lemme 12. Il n'y a besoin que de $O(\log N)$ bits pour qu'un algorithme soit capable d'énumérer les éléments de A.

Démonstration.. Nous avons $K(x_F \mid x_E) \leqslant \log N + O(1)$ d'une part, et d'autre part L n'est qu'une constante fixée à l'avance qui ne dépend pas de N. Dans ce cas, il n'y a besoin que de $2^L \cdot \log N$ bits — c'est-à-dire $O(\log N)$ bits — pour décrire tous les $K(x_F \mid x_E)$ pour tous les sous-ensembles possibles E et F. Cela suffit bien à déterminer les L-uplets appartenant à A. ∎

Le principe est le même que dans la démonstration du théorème de Kolmogorov-Levin : nous pouvons décrire un élément de A en donnant sa position dans l'énumération de A. Il en va de même pour les A_E, et par conséquent nous avons une borne inférieure sur leurs cardinaux :

Lemme 13. $\log |A_E| \geqslant K(x_E) - O(\log N)$

Démonstration.. Ceci est une conséquence immédiate du fait que :

$$K(x_E) \leqslant \log |A_E| + O(\log N)$$

En effet, nous pouvons énumérer A avec seulement $O(\log N)$ bits et projeter ses éléments sur A_E, pour ensuite sélectionner un élément de A_E dans l'énumération avec seulement $\log |A_E|$ bits. Cela nous fournit une description de tous les éléments de A_E, or l'un de ces éléments est x_E par construction, d'où le résultat. ■

Si $\langle \sigma_1, \ldots, \sigma_L \rangle \in A$ alors en particulier $K(\sigma_E \mid \sigma_\varnothing) \leqslant K(x_E \mid x_\varnothing)$, ce qui peut se réécrire simplement $K(\sigma_E) \leqslant K(x_E)$. Tous les éléments de A_E satisfont donc cette propriété, or nous savons d'autre part qu'il n'y a pas plus que $2^{K(x_E)+1}$ chaînes de complexité inférieure ou égale à $K(x_E)$ par des arguments similaires au lemme 5. Nous en déduisons finalement un encadrement complet du cardinal de A_E :

$$2^{K(x_E)-O(\log N)} \leqslant |A_E| \leqslant 2^{K(x_E)+1} \tag{4}$$

Considérons une v.a. X suivant une loi uniforme sur les L-uplets appartenant à A, et notons $X_1, \ldots, X_L$ ses projections. Les $X_E = \langle X_{i_1}, \ldots, X_{i_\ell} \rangle$ sont dans ce cas des v.a. qui prennent leurs valeurs parmi les éléments des A_E par définition. On a trivialement :

$$H(X_E) \leqslant \log |A_E| \leqslant K(x_E) + 1 \quad \text{(cf. corollaire de la proposition 3)}$$

Il nous reste à voir une borne inférieure qui nous permettra d'établir une égalité à un facteur logarithmique près. Nous avons besoin d'un dernier lemme :

Lemme 14. Si Y est une v.a. telle que $\mathbf{Pr}(Y = \ell) \leqslant r$ pour tout ℓ, alors $H(Y) \geqslant \log \frac{1}{r}$.

Démonstration.. L'hypothèse sur Y équivaut à dire $\log \frac{1}{\mathbf{Pr}(Y = \ell)} \geqslant \log \frac{1}{r}$ ce qui permet de déduire le résultat voulu en passant à la somme. ■

Voyons comment majorer $\mathbf{Pr}(X_E = \tau_E)$. Puisque X est une v.a. uniforme, alors la probabilité que X_E prenne la valeur τ_E est par définition la proportion de L-uplets $\langle \sigma_1, \ldots, \sigma_L \rangle$ appartenant à A dont la projection sur A_E est égale à τ_E :

$$\mathbf{Pr}(X_E = \tau_E) = \frac{|\{ \langle \sigma_1, \ldots, \sigma_L \rangle \in A \mid \sigma_E = \tau_E \}|}{|A|}$$

On sait déjà d'une part minorer le dénominateur $|A|$ en utilisant l'inégalité (4). D'autre part, si $\langle \sigma_1, \ldots, \sigma_L \rangle$ et $\langle \sigma'_1, \ldots, \sigma'_L \rangle$ sont deux L-uplets appartenant à A tels que $\sigma_E = \sigma'_E = \tau_E$ alors ils ne diffèrent au mieux que sur les coordonnées qui n'appartiennent pas à E, c'est-à-dire sur les coordonnées qui appartiennent au complémentaire $E^{\mathrm{c}} \stackrel{\text{def}}{=} \{ 1, \ldots, L \} \setminus E$. Les éléments du numérateurs peuvent par conséquent être entièrement décrits par une description de τ_E ainsi qu'une description d'un $\sigma_{E^{\mathrm{c}}}$ sachant τ_E. Par hypothèse sur la construction de A on a pour tout L-uplet $\langle \sigma_1, \ldots, \sigma_L \rangle$ appartenant au numérateur :

$$K(\sigma_{E^{\mathrm{c}}} \mid \tau_E) = K(\sigma_{E^{\mathrm{c}}} \mid \sigma_E) \leqslant K(x_{E^{\mathrm{c}}} \mid x_E)$$

Avec toujours le même argument sur le nombre de chaînes d'une complexitée bornée, le cardinal du numérateur ne saurait excéder $2^{K(x_{E^c}|x_E)+1}$. En combinant les deux résultats, on en déduit que pour tout τ_E :

$$\mathbf{Pr}(X_E = \tau_E) \leqslant 2^{K(x_{E^c}|x_E)+1} \cdot 2^{-K(x_1,\ldots,x_L)+O(\log N)}$$

d'où $H(X_E) \geqslant K(x_1, \ldots, x_L) - K(x_{E^c} \mid x_E) + 1 - O(\log N)$ en vertu du dernier lemme, ce qui peut se réécrire de la sorte en utilisant le théorème de Kolmogorov-Levin :

$$H(X_E) \geqslant K(x_E) - O(\log N)$$

C'est bien ce que l'on cherchait. ■

4.3 Conséquences

La manière dont nous avons construit la démonstration du théorème 4 nous permet de faire les deux conclusions suivantes :

Proposition 10. Si une inégalité linéaire est vraie pour la complexité de Kolmogorov à $o(N)$ près, alors elle est vraie à $O(\log N)$ près.

Démonstration.. En effet tous les termes négligeables devant N disparaissent dans la démonstration lorsque l'on utilise le lemme de Zvonkin & Levin. Ainsi nous en déduisons que l'inégalité linéaire est *exactement* vraie pour l'entropie de Shannon, et en appliquant le théorème 4 une seconde fois l'inégalité est nécessairement vraie pour la complexité de Kolmogorov à $O(\log N)$ près. ■

Proposition 11. Étant donné L chaînes $x_1, \ldots, x_L$ avec $N = |x_1| + \cdots + |x_L|$, alors le vecteur dont les coordonnées correspondent aux $K(x_E)$ est entropique à $O(\log N)$ près.

Démonstration.. Pour rappel, un vecteur est *entropique* s'il correspond aux entropies d'une famille de variables aléatoires. La proposition ne fait qu'énoncer ce que nous avons exactement fait dans la seconde partie de la démonstration en construisant des v.a. dont l'entropie est égale à la complexité de Kolmogorov à $O(\log N)$ près. ■

5 Inégalités classiques

Regardons de plus près les inégalités d'information en tant que telles, puisque l'on sait maintenant qu'elles sont équivalentes entre les deux théories de l'information. Il en existe deux sortes :

i) les inégalités *classique* — ou *de type Shannon* — que nous allons étudier maintenant ;

ii) et les inégalités *non-classiques* — ou *de type non-Shannon* — plus subtiles, qui feront l'objet d'une prochaine section.

5.1 Motivation et définition

Un certain nombre d'inégalités triviales peuvent être observées :

Fait. Soient L v.a. $X_1, \dots, X_L$. Alors pour n'importe quels $E, F \subseteq \{ 1, \dots, L \}$ les inégalités suivantes sont vraies :

i) $H(X_E) \geqslant 0$

ii) Si $F \subseteq E$ alors $H(X_F) \leqslant H(X_E)$

iii) $H(X_{E \cup F}) \leqslant H(X_E) + H(X_F)$

Leur démonstration ne pose aucune difficulté. Bien évidemment, ces inégalités sont vraies pour la complexité de Kolmogorov préfixe à $O(\log N)$ près, mais là encore il n'y aurait aucune difficulté à montrer que ces inégalités sont même vraies à $O(1)$ près. Il semble émaner une structure commune de ces trois inégalités : l'inégalité i) est un cas particulier de l'inégalité ii), et l'inégalité ii) semble *presque* être une conséquence de l'inégalité iii). Nous allons préciser ces « airs de famille ». D'après le corollaire du théorème de symétrie de l'information (p. 6), l'information mutuelle peut s'écrire de la manière suivante :

$$I(X : Y) = H(X) + H(Y) - H(X, Y)$$

Dans ce cas, il est facile de *relativiser* la définition de l'information mutuelle.

Définition 15. On appelle *information mutuelle conditionnelle* de X et Y sachant Z la quantité définie par :

$$I(X : Y \mid Z) \stackrel{\text{def}}{=} H(X \mid Z) + H(Y \mid Z) - H(X, Y \mid Z)$$

Théorème 5. L'information mutuelle conditionnelle est toujours positive.

Démonstration.. Cela revient à montrer que :

$$H(X, Y \mid Z) \leqslant H(X \mid Z) + H(Y \mid Z) \tag{5}$$

ce qui est clairement le cas en relativisant l'inégalité iii) vue précédemment. ■

Remarque. L'inégalité (5) ci-dessus peut se réécrire grâce au théorème de symétrie de l'information sous la forme suivante :

$$H(Z) + H(X, Y, Z) \leqslant H(X, Y) + H(Y, Z) \tag{6}$$

Ces deux inégalités sont strictement équivalentes dans le cas de l'entropie de Shannon, alors que pour la complexité de Kolmogorov préfixe seule l'inégalité (5) est vraie à $O(1)$ près tandis que l'inégalité (6) ne l'est qu'à un facteur logarithmique près. Cette asymétrie implique entre autres qu'il existe au moins deux manières incompatibles de définir l'information mutuelle conditionnelle en théorie algorithmique de l'information :

$$I(x : y \mid z) \stackrel{\text{def}}{=} K(x \mid z) + K(y \mid z) - K(x, y \mid z)$$

ou

$$I(x : y \mid z) \stackrel{\text{def}}{=} K(x, z) + K(y, z) - K(x, y, z) - K(x)$$

La première définition est toujours positive à $O(1)$ près, alors que ce n'est pas le cas de la seconde définition. Ce genre de distinction sera importante prochainement, lorsque nous verrons un exemple d'inégalité non triviale qui vraie à $O(1)$ près pour la complexité de Kolmogorov.

On obtient immédiatement de l'inégalité (6) un corollaire fondamental :

Corollaire. Pour n'importe quels sous-ensembles $E, F, G \subseteq \{ 1, \ldots, L \}$ l'inégalité suivante est vraie :

$$H(X_G) + H(X_{E \cup F \cup G}) \leqslant H(X_{E \cup G}) + H(X_{F \cup G}) \tag{7}$$

Définition 16. Les inégalités qui peuvent s'écrire sous la forme (7) sont appelées des *inégalités basiques*. Une inégalité est *classique* si elle peut s'écrire comme une somme d'inégalités basiques.

Exemple. Toutes les inégalités triviales que nous avons vues au début de cette section peuvent en effet s'écrire comme des cas particuliers de l'inégalité (7). En conséquence, ce sont des inégalités classiques.

Nous avions remarqué au début de la section précédente (p. 22) que les inégalités d'information correspondaient exactement aux équations des hyperplans qui délimitent la région des vecteurs entropique dans le $\mathbb{R}$-espace vectoriel de dimension $2^L - 1$. Les coefficients (a_E) de ces hyperplans forment ensemble un cône de vecteurs $(\lambda \cdot a_E)$ où $\lambda \geqslant 0$, et dans ce cas les inégalités classiques correspondent aux sous-cônes engendrés par les inégalités basiques. Cette vision nous permet de résoudre « algorithmiquement » le test d'une inégalité :

Lemme 15. Il existe un algorithme qui décide si une inégalité est classique.

Démonstration.. Il suffit d'appliquer l'algorithme du pivot de Gauss à l'inégalité que l'on veut tester avec toutes les familles libres d'inégalités basiques. ■

Le lemme 15 va nous servir pour affirmer un certain nombre de fois que des inégalités sont classiques ou non-classiques sans le démontrer : le lecteur peut donc utiliser son logiciel d'algèbre linéaire préféré afin de vérifier les dires de l'auteur s'il le souhaite. Si l'algorithme détermine qu'une inégalité est effectivement classique, alors elle est automatiquement vraie. La question réciproque de savoir si toutes les inégalités vraies étaient classiques a connu sa résolution complète à la fin des années 1990 par une réponse négative de Zhang & Yeung [12]. Il est cependant nécessaire de monter jusqu'à au moins quatre v.a. pour trouver des contre-exemples : en effet, il se trouve que toutes les inégalités avec au plus trois v.a. sont classiques.

5.2 Situation avec trois variables

Fait (Hammer et al. [4, théorème 3]). Les inégalités linéaires vraies avec au plus trois v.a. sont classiques.

La démonstration originale nécessiterait de faire un détour pour parler des inégalités de rang de sous-espaces vectoriels, néanmoins une idée que nous allons reprendre ici est la suivante : le théorème de symétrie de l'information nous premet de réécrire l'entropie simple comme une somme d'entropies conditionnelles et d'informations mutuelles. Étant donné un vecteur entropique, nous pourrions dans ce cas « changer de base » en exprimant ses coordonnées précisément avec des entropies conditionnelles et des informations mutuelles. En analysant sous cette nouvelle forme la structure des vecteurs entropiques, on espère ainsi pouvoir mieux comprendre la structure des inégalités d'information par dualité. Pour illustrer cette idée, la Figure 4 ci-après étend à trois v.a. la Figure 1 que nous avions vue avec le théorème de symétrie de l'information.

5.2.1 Information mutuelle « triple »

Fait. La quantité suivante permet de compléter les « ??? » de la Figure 4 de telle manière à ce que toutes les sommes possibles sur le diagramme correspondent bien :

$$\begin{aligned} I(X : Y : Z) &\stackrel{\text{def}}{=} H(X) + H(Y) + H(Z) \\ &\quad - H(X, Y) - H(Y, Z) - H(Z, X) \\ &\quad + H(X, Y, Z) \end{aligned}$$

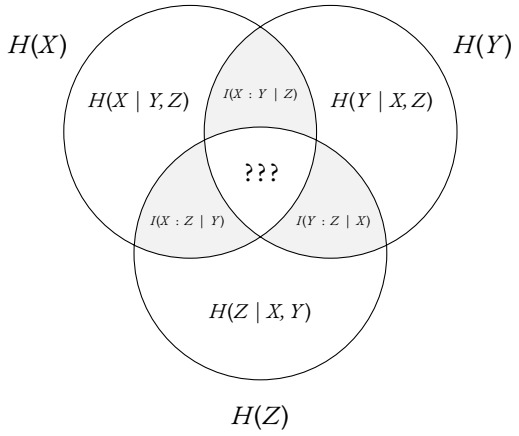

Figure 4 – Décomposition de $H(X, Y, Z)$.

C'est une simple application de la *formule d'inclusion-exclusion* (aussi connue sous le nom de *crible de Poincaré*) avec le théorème de symétrie de l'information. Nous avons par conséquent l'égalité suivante :

$$\underbrace{I(X : Y)}_{\geqslant 0} = \underbrace{I(X : Y \mid Z)}_{\geqslant 0} + I(X : Y : Z)$$

Si au lieu de v.a. X, Y, Z nous nous étions intéressés à des ensembles finis A, B, C, alors nous aurions eu dans ce cas une égalité similaire portant sur les cardinaux :

$$\underbrace{|A \cap B|}_{\geqslant 0} = \underbrace{|A \cap B \setminus C|}_{\geqslant 0} + \underbrace{|A \cap B \cap C|}_{\geqslant 0}$$

Cela représente exactement l'assemblage des « morceaux partagés » par A et B sur le diagramme de Venn correspondant aux trois ensembles A, B, C. En particulier, le cardinal $|A \cap B \cap C|$ est évidemment toujours positif, ce qui ne choque évidemment pas l'intuition que l'intersection des trois ensembles au centre du diagramme de Venn « contient de la matière ». Ce n'est cependant pas le cas des variables aléatoires.

Proposition 12. Il existe trois v.a. X, Y, Z telles que $I(X : Y : Z) < 0$.

Démonstration.. Prenons X et Y deux v.a. i.i.d. suivant une loi uniforme sur $\{ 0, 1 \}$. Alors en posant $Z \overset{\text{def}}{=} X$ xor Y — c'est-à-dire Z est est définie comme le « ou exclusif » de X et de Y — les propriétés du xor nous permettent de déduire les égalités suivantes :

$$H(Z) = 1 \qquad H(Z \mid X, Y) = 0$$
$$I(X : Z) = 0 \qquad I(Y : Z) = 0$$

Mais alors dans ce cas, nous avons d'une part :

$$\begin{aligned} H(Z) &= \overbrace{H(Z \mid X, Y)}^{=0} + \overbrace{I(X : Z \mid Y) + I(X : Y : Z)}^{=I(X:Z)=0} + I(Y : Z \mid X) \\ &= I(Y : Z \mid X) \\ &= 1 \end{aligned}$$

et d'autre part :

$$I(Y : Z) = I(Y : Z \mid X) + I(X : Y : Z) = 0$$

Finalement, on en déduit $I(X : Y : Z) = -1$. ■

5.2.2 Réalisation de diagrammes

Maintenant que nous savons qu'un vecteur entropique pour trois v.a. peut se décomposer à la manière de la FIGURE 4 nous pouvons nous demander si, étant donné un diagramme, nous sommes capables de trouver des v.a. qui *réalisent* celui-ci, c'est-à-dire telles que leurs entropies correspondent effectivement au diagramme :

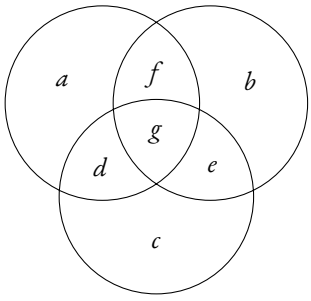

FIGURE 5 – Existe-t-il des v.a. qui réalisent le diagramme ci-dessus ?

On dira dans ce cas qu'un diagramme est *réalisable* ou *entropique* s'il existe de telles variables aléatoires. L'entropie et l'information mutuelle devant toujours être positives, nous

pouvons d'ores et déjà énumérer des conditions *nécessaires* à leur existence :

$$a \geqslant 0 \qquad d \geqslant 0 \qquad d+g \geqslant 0$$
$$b \geqslant 0 \qquad e \geqslant 0 \qquad e+g \geqslant 0$$
$$c \geqslant 0 \qquad f \geqslant 0 \qquad f+g \geqslant 0$$

Étant donné deux vecteurs entropiques $(H(X_E))$ et $(H(Y_E))$, la somme des deux vecteurs est elle-même entropique en supposant que les X_E et les Y_E sont indépendantes deux à deux : c'est une supposition que l'on peut toujours faire quitte à prendre des copies i.i.d. des v.a. initiales. En effet nous avons dans ce cas immédiatement $H(X_E, Y_E) = H(X_E) + H(Y_E)$ pour tout E, ce que l'on voulait. Cette observation s'étend naturellement aux diagrammes réalisables : nous pouvons « sommer » (ou « superposer ») deux diagrammes en recombinant des copies i.i.d. des v.a. qui les réalisent. On pourrait alors essayer de diviser en deux diagrammes les conditions nécessaires que nous avons formulées précédemment, de telle sorte à isoler l'information mutuelle « triple » qui semble plus difficile à réaliser que les autres :

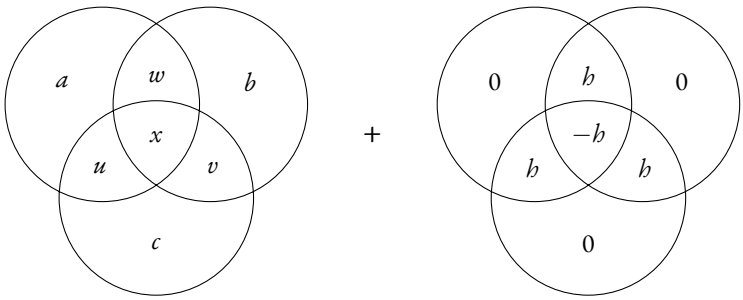

FIGURE 6 – On isole la réalisation de l'information mutuelle « triple » du reste.

En supposant que les nouvelles quantités u, v, w, x et h sont toutes positives, la superposition des deux diagrammes de la FIGURE 6 satisfait toujours les conditions nécessaires :

$$d = u + h \qquad e = v + h \qquad f = w + h \qquad g = x - h$$

Réciproquement, en choisissant h tel que :

$$\min(d, e, f \max(0, -g)) \leqslant h \leqslant \min(d, e, f) \tag{8}$$

Nous pouvons dans ce cas transformer les coordonnées du diagramme de la FIGURE 5 en deux diagrammes comme sur la FIGURE 6 :

$$u = d - h \geqslant 0 \qquad v = e - h \geqslant 0 \qquad w = f - h \geqslant 0 \qquad x = g + h \geqslant 0$$

Remarque. Il faut faire attention ici : il semblerait que la réalisation de la Figure 5 et la réalisation de la Figure 6 sont deux problèmes équivalents, puisqu'il est possible d'écrire des coordonnées de l'un dans les coordonnées de l'autre. Cependant, il pourrait tout-à-fait s'avérer que les diagrammes de la forme :

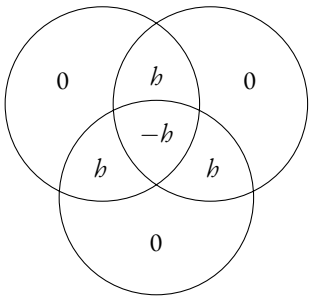

Figure 7 – Diagramme de l'information mutuelle « triple ».

ne soient réalisables que pour certaines valeurs de h seulement, et par conséquent l'inégalité (8) qui permet la transformation de la Figure 5 en la Figure 6 pourrait ne jamais être satisfaite alors même que le diagramme de la Figure 5 est bel est bien réalisable. Sans apporter une réponse complète à cette difficulté, nous allons voir que h ne peut effectivement pas prendre n'importe quelle valeur.

Essayons maintenant de voir comment réaliser les deux diagrammes de la Figure 6. Pour celui de gauche, c'est facile : grâce à la proposition 4 nous pouvons trouver des v.a. indépendantes $X_a, \dots, X_x$ telles que leurs entropies correspondent à chaque « morceau » du diagramme, c'est-à-dire $H(X_a) = a, \dots, H(X_x) = x$. En combinant ces v.a. ensemble, nous pouvons ainsi construire sans difficulté trois v.a. qui réalisent bien le diagramme. Pour celui de droite, c'est une autre affaire.

Proposition 13. Le diagramme de la Figure 7 est réalisable si, et seulement s'il existe un entier N tel que $h = \log N$.

Démonstration.. Commençons par le sens réciproque : on suppose qu'il existe N tel que $h = \log N$. En fait, nous avons déjà traité le cas $N = 2$: il s'agit de la construction réalisée dans la démonstration de la proposition 12. Nous pouvons généraliser cette construction en se donnant X et Y deux v.a. i.i.d. suivant une loi uniforme sur le groupe additif $\mathbb{Z}/N\mathbb{Z}$ et en posant $Z \overset{\text{def}}{=} X+Y$. En effectuant le même raisonnement qu'à la proposition 12, on montre que X, Y, Z réalise bien le diagramme.

Passons maintenant au sens direct, plus difficile. Supposons avoir trois v.a. X, Y, Z qui

réalisent le diagramme pour un certain h. On a entre autres $H(Y \mid X, Z) = 0$, c'est-à-dire que Y est entièrement déterminée par les valeurs de X et de Z. Nous pouvons alors trouver une famille d'applications f_m indicée sur les valeurs de Z qui « choisit » presque sûrement la valeur que Y va prendre étant donné une valeur de X :

$$\mathbf{Pr}(Y = f_m(k) \mid X = k \wedge Z = m) = 1$$

On a également $H(X \mid Y, Z) = 0$, ce qui impose en fait que les f_m sont toutes des bijections : on peut presque sûrement retrouver l'antécédent pour X en connaissant la valeur de Y et de Z. Par un raisonnement identique on peut trouver des bijections entre Y et Z ainsi que X et Z, d'où finalement les trois v.a. prennent leurs valeurs parmi des ensembles de même cardinal. Sans perdre en généralité, nous pouvons donc supposer que les v.a. X, Y, Z prennent leurs valeurs parmi $1, \ldots, N$. Elles sont indépendantes par hypothèse, puisque nous avons par exemple $I(X : Y) = h - h = 0$ avec X et Y, d'où $\mathbf{Pr}(X = k \wedge Y = \ell) = \mathbf{Pr}(X = k) \cdot \mathbf{Pr}(Y = \ell)$ et ainsi de suite. Rappelons par ailleurs les égalités suivantes :

$$\begin{aligned}
\mathbf{Pr}(X = k \wedge Y = \ell \wedge Z = m) &= \overbrace{\mathbf{Pr}(X = k \wedge Y = \ell)}^{= \mathbf{Pr}(X = k) \cdot \mathbf{Pr}(Y = \ell)} \cdot \mathbf{Pr}(Z = m \mid X = k \wedge Y = \ell) \\
&= \mathbf{Pr}(Y = \ell \wedge Z = m) \cdot \mathbf{Pr}(X = k \mid Y = \ell \wedge Z = m) \\
&\vdots
\end{aligned}$$

Or si $\mathbf{Pr}(X = k \wedge Y = \ell \wedge Z = m) \neq 0$ alors la valeur que prend Z est entièrement déterminée par X et Y, d'où $\mathbf{Pr}(Z = m \mid X = k \wedge Y = \ell) = 1, \ldots$ Mais alors :

$$\begin{aligned}
\mathbf{Pr}(X = k \wedge Y = \ell \wedge Z = m) &= \mathbf{Pr}(X = k) \cdot \mathbf{Pr}(Y = \ell) \\
&= \mathbf{Pr}(Y = \ell) \cdot \mathbf{Pr}(Z = m) \\
&= \mathbf{Pr}(Z = m) \cdot \mathbf{Pr}(X = k)
\end{aligned}$$

On en déduit dans le cas où $\mathbf{Pr}(X = k \wedge Y = \ell \wedge Z = m) \neq 0$ que :

$$\mathbf{Pr}(X = k) = \mathbf{Pr}(Y = \ell) = \mathbf{Pr}(Z = m)$$

C'est en particulier le cas lorsque $\ell = f_m(k)$ avec m fixé et k quelconque, c'est-à-dire que pour tout k :

$$\mathbf{Pr}(X = k) = \overbrace{\mathbf{Pr}(Z = m)}^{\text{constant!}}$$

Cela montre bien que X suit une loi uniforme, et par un raisonnement identique c'est également le cas de Y et Z. Leur entropie est donc égale à :

$$H(X) = H(Y) = H(Z) = \log N = h + h - h = h$$

C'est ce qu'il fallait démontrer. ■

Il nous faut donc seulement choisir a, b, c, u, v, w, x positifs quelconques et $h = \log N$ pour fabriquer un très grand nombre de vecteurs entropiques pour trois variables. Nous n'avons cependant pas montré que *tous* les vecteurs entropiques peuvent se décomposer sous cette forme, ce contre quoi nous avions mis en garde dans la remarque précédente (p. 34). Cela nous fournit néanmoins un outil efficace pour tester rapidement si une inégalité d'information pour trois variables semble être — ou ne pas être — vraie.

6 Un exemple d'inégalité à $O(1)$ près

La question de savoir quelles sont les inégalités d'information vraies pour trois variables *à un facteur logarithmique près* semble désormais à peu près entièrement résolue, de ce que nous venons de voir. Une autre question, beaucoup plus difficile, et de savoir quelles sont les inégalités qui sont encore vraies à $O(1)$ près pour la complexité de Kolmogorov. En particulier, le théorème de Kolmogorov-Levin n'est vrai qu'à $O(\log N)$ près, et nous avions remarqué qu'il existait plusieurs définitions incompatibles possibles de l'information mutuelle conditionnelle pour la théorie algorithmique de l'information (p. 29). Si l'on veut « tomber juste » à $O(1)$ près, nous ne pouvons donc pas aussi simplement décomposer la complexité de Kolmogorov en diagrammes comme nous l'avions fait avec l'entropie de Shannon.

6.1 Présentation

Nous allons démontrer que l'inégalité suivante est vraie à constante près :

$$\frac{2}{3}\left[C(x \mid y,z) + C(y \mid z,x) + C(z \mid x,y)\right] + I(x : y \mid z) + I(y : z \mid x) + I(z : x \mid y) \geqslant 0 \quad (9)$$

Plus exactement, nous allons montrer qu'elle est vraie en prenant les deux définitions suivantes possibles de l'information mutuelle conditionnelle :

$$I(x : y \mid z) \stackrel{\text{def}}{=} C(x \mid z) + C(y \mid z) - C(x,y \mid z)$$
$$I(x : y \mid z) \stackrel{\text{def}}{=} C(y \mid z) - C(y \mid x,z)$$

Notons ici que l'on utilise la complexité de Kolmogorov *pleine*. L'une des principales difficultés réside entre autres dans la définition de l'information mutuelle conditionnelle que nous allons choisir. Si l'on ne se pose pas trop de questions pour le moment, nous pouvons déjà nous rassurer sur un point :

Proposition 14. L'inégalité (9) est vraie à $O(\log N)$ près.

Démonstration.. Il suffit de le démontrer pour l'entropie de Shannon. En l'occurrence, l'entropie et l'information mutuelle conditionnelle sont toujours positives, ce qui donne le résultat voulu via le théorème d'équivalence pour la complexité *préfixe* et donc également pour la complexité *pleine* puisque les deux sont égales à un facteur logarithmique près. ∎

6.2 Démonstration

Rappelons que l'inégalité $K(x,y) \leqslant K(x) + K(y)$ est vraie à constante près, tandis que l'inégalité $C(x,y) \leqslant C(x) + C(y)$ ne l'est qu'à un facteur logarithmique. Dans ce cas, nous

n'avons même pas que les informations mutuelles telles que définies ci-après pour la complexité de Kolmogorov *pleine* sont positives à $O(1)$ près :

$$I(x : y) \stackrel{\text{def}}{=} C(x) + C(y) - C(x, y)$$
$$I(x : y \mid z) \stackrel{\text{def}}{=} C(x \mid z) + C(y \mid z) - C(x, y \mid z)$$

Il est néanmoins possible de donner une borne inférieure :

Lemme 16. Soit $\varepsilon > 0$. Alors $I(x : y \mid z) \geqslant C(y \mid z) - C(y \mid x, z) - \varepsilon \cdot C(x \mid z) - O(1)$.

Démonstration.. Montrons d'abord l'inégalité suivante :

$$C(x, y) \leqslant (1 + \varepsilon) \cdot C(x) + C(y \mid x) + O(1)$$

En effet, il n'y a besoin que de $\log C(x)$ bits pour indiquer où couper entre la plus petite description de x et la plus petite description de y sachant x, et pour x suffisamment grand nous avons que $\varepsilon \cdot C(x) > \log C(x)$ d'où l'inégalité. Nous pouvons alors la réécrire de cette façon :

$$C(y) - C(y \mid x) - \varepsilon \cdot C(x) - O(1) \leqslant C(x) + C(y) - C(x, y)$$

On reconnaît à droite la définition de $I(x : y)$, ce qui en relativisant l'inégalité à z donne bien le résultat voulu. ∎

Une autre manière de définir les informations mutuelles est la suivante :

$$I(x : y) \stackrel{\text{def}}{=} C(y) - C(y \mid x)$$
$$I(x : y \mid z) \stackrel{\text{def}}{=} C(y \mid z) - C(y \mid x, z)$$

Elles n'ont *a priori* aucune raison d'être exactement égales aux définitions précédentes. Cela ne change rien cependant :

Lemme 17. Avec ces nouvelles définitions, la borne inférieure donnée par le lemme 16 est encore juste.

Démonstration.. Il suffit de remarquer que :

$$I(x : y) = C(y) - C(y \mid x) \geqslant C(y) - C(y \mid x) - \underbrace{\varepsilon C(x)}_{\geqslant 0}$$

et de conclure de la même façon en relativisant à z. ∎

Remarque. On ne sait pas si le lemme 16 est encore juste avec la troisième définition ci-dessous :

$$I(x : y \mid z) \stackrel{\text{def}}{=} C(x, z) + C(y, z) - C(x, y, z) - C(z)$$

D'ailleurs, on ne sait même pas si $I(x : y) = I(y : x) + O(1)$.

Dans deux des trois définitions possibles de l'information mutuelle conditionnelle, nous avons par conséquent une borne inférieure à constante près. Une sorte de « relation de Chasles » nous est enfin nécessaire pour démontrer le résultat annoncé :

Lemme 18. Soit $\varepsilon > 0$. Alors $C(y \mid x) \leqslant (1 + \varepsilon) \cdot C(z \mid x) + C(y \mid z) + O(1)$.

Démonstration.. La démonstration utilise la même idée que dans le lemme 16 : pour se donner une description de y sachant x, il est suffisant de se donner une description d'une chaîne intermédiaire z sachant x, puis d'une description de y sachant z. Il n'y a pas besoin de plus de $\log C(z \mid x) < \varepsilon C(z \mid x)$ bits pour indiquer où couper lorsque x est assez grand, d'où le résultat. ∎

Proposition 15. L'inégalité (9) est vraie à $O(1)$ près.

Démonstration.. Nous pouvons utiliser trois fois le lemme 16 pour minorer chaque information mutuelle conditionnelle de l'inégalité (9) :

$$\begin{aligned} I(x : y \mid z) + I(y : z \mid x) + I(z : x \mid y) \geqslant\ & C(y \mid z) - C(y \mid x, z) - \varepsilon \cdot C(x \mid z) \\ & + C(z \mid x) - C(z \mid y, x) - \varepsilon \cdot C(y \mid x) \\ & + C(x \mid y) - C(y \mid z, y) - \varepsilon \cdot C(z \mid y) \\ & - O(1) \end{aligned}$$

Remarquons d'autre part que $-\frac{1}{3}C(y \mid z, x) \geqslant -\frac{1}{3}C(y \mid z)$ et dans ce cas :

$$\frac{2}{3}C(y \mid z, x) + C(y \mid z) - C(y \mid x, z) - \varepsilon \cdot C(x \mid z) \geqslant \frac{2}{3}C(y \mid z) - \varepsilon \cdot C(x \mid z) - O(1)$$

En effet, nous avons alors :

$$\begin{aligned} \frac{2}{3}C(y \mid z, x) + C(y \mid z) - C(y \mid x, z) &= C(y \mid z) - \frac{1}{3}C(y \mid z, x) \\ &\geqslant C(y \mid z) - \frac{1}{3}C(y \mid z, x) = \frac{2}{3}C(y \mid z) \end{aligned}$$

et ainsi de suite pour $C(z \mid x)$ et $C(x \mid y)$. En sommant les deux expressions précédentes ensemble, on aboutit à cette inégalité :

$$\frac{2}{3}\left[C(x \mid y,z) + C(y \mid z,x) + C(z \mid x,y)\right] + I(x : y \mid z) + I(y : z \mid x) + I(z : x \mid y)$$
$$\geqslant$$
$$\frac{2}{3}\left[C(x \mid y) + C(y \mid z) + C(z \mid x)\right] - \varepsilon \cdot \left[C(y \mid x) + C(z \mid y) + C(x \mid z)\right] - O(1)$$

Il suffit donc de montrer que la partie droite de l'inégalité est positive à constante près pour démontrer l'inégalité (9). Or, en utilisant le lemme 18 nous obtenons que :

$$\varepsilon \cdot C(y \mid x) \leqslant \varepsilon \cdot \left[O(1) \cdot C(y \mid z) + C(z \mid x)\right] + O(1) \leqslant \frac{1}{3}\left[C(y \mid z) + C(z \mid x)\right] + O(1)$$

pour ε suffisamment petit. Nous pouvons le réécrire sous la forme :

$$\frac{1}{3}\left[C(y \mid z) + C(z \mid x)\right] \geqslant \varepsilon \cdot C(y \mid x) - O(1)$$

et ainsi de suite avec $C(z \mid y)$ et $C(x \mid z)$. En sommant ces inégalités, nous avons bien :

$$\frac{2}{3}\left[C(x \mid y) + C(y \mid z) + C(z \mid x)\right] \geqslant \varepsilon \cdot \left[C(y \mid x) + C(z \mid y) + C(x \mid z)\right] - O(1)$$

ce qu'il fallait démontrer. ■

Remarque. Il est intéressant d'observer que nous nous sommes intéressés à ce qui pourrait s'apparenter à une sorte d'inégalité de *circuits* dans la dernière étape de la démonstration précédente :

$$a \cdot \left[C(x \mid y) + C(y \mid z) + C(z \mid x)\right] \geqslant b \cdot \left[C(y \mid x) + C(z \mid y) + C(x \mid z)\right] - O(1)$$

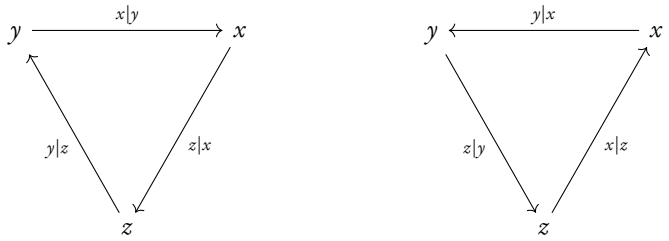

Figure 8 – Deux directions possibles d'un circuit à trois variables.

Il faut considérablement plus d'efforts pour démontrer la proposition 15, là où il n'avait suffit que de quelques lignes pour la proposition 14. En général, établir une inégalité à constante près est bien plus difficile. Voyons maintenant un autre genre d'inégalités qui ne sont pas une trivialité non plus.

7 Inégalités non-classiques

Les inégalités d'information ne sont pas toutes classiques : c'est le résultat qu'ont montré Zhang & Yeung en 1998 [12]. Comme annoncé il nous faudra travailler avec quatre variables pour en voir un exemple, mais d'abord nous allons regarder ce qu'il se passe avec des objets plus réguliers : les $\mathbb{R}$-espaces vectoriels.

7.1 Inégalité de Ingleton sur les $\mathbb{R}$-espaces vectoriels

Supposons avoir quatre $\mathbb{R}$-espaces vectoriels E, F, G, H de dimension finie. Alors nous pouvons définir une sorte d'analogue de l'information mutuelle avec le rang :

$$I(E : F) \stackrel{\text{def}}{=} \dim E + \dim F - \dim(E + F)$$
$$I(E : F \mid G) \stackrel{\text{def}}{=} \dim(E + G) + \dim(F + G) - \dim(E + F + G) - \dim G$$

Nous avons dans ce cas l'inégalité suivante, appelée *inégalité de Ingleton* :

Lemme 19 (Ingleton [5]). $I(E : F) \leqslant I(E : F \mid G) + I(E : F \mid H) + I(G : H)$

En dessinant un diagramme avec quatre variables et en observant les morceaux qui sont censés correspondre, l'inégalité de Ingleton semble tout-à-fait raisonnable :

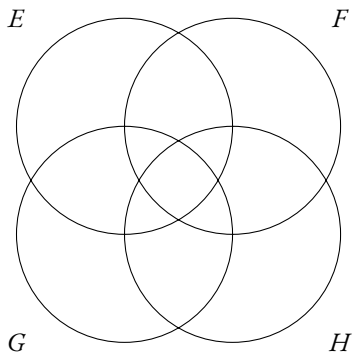

Figure 9 – Intersection de quatre $\mathbb{R}$-espaces vectoriels.

Comme avec l'information mutuelle « triple », l'intuition donnée par les diagrammes ne se prolonge pas aussi simplement avec les variables aléatoires. En effet, l'inégalité de Ingleton est parfois fausse dans le cadre de l'entropie de Shannon :

Proposition 16. Il existe quatre v.a. X, Y, A, B telles que :

$$I(X : Y) \nleqslant I(X : Y \mid A) + I(X : Y \mid B) + I(A : B)$$

Démonstration.. Considérons A, B deux v.a. i.i.d. suivant une loi uniforme sur $\{ 0, 1 \}$.

Posons par ailleurs :

$$X \overset{\text{def}}{=} A \cdot (1 - B) \qquad\qquad Y \overset{\text{def}}{=} B \cdot (1 - A)$$

Nous avons d'une part que $I(A : B) = 0$ par hypothèse. D'autre part :

$$\begin{aligned} I(X : Y \mid A) &= H(A \cdot (1 - B) \mid A) + H(B \cdot (1 - A) \mid A) \\ &\quad - H(A \cdot (1 - B), B \cdot (1 - A) \mid A) \\ &= \frac{1}{2} H(1 - B) + \frac{1}{2} H(B) - \left(\frac{1}{2} H(1 - B) + \frac{1}{2} H(B) \right) \\ &= 0 \end{aligned}$$

De manière symétrique nous avons également $I(X : Y \mid B) = 0$. Enfin, si $X = 1$ alors $A = 1$ et $B = 0$ nécessairement, et donc $Y = 0$. Ceci montre que X et Y ne sont pas indépendantes, c'est-à-dire que $I(X : Y) \neq 0$. D'où finalement :

$$0 < I(X : Y) \nleqslant \underbrace{I(X : Y \mid A) + I(X : Y \mid B) + I(A : B)}_{= 0} \qquad \blacksquare$$

7.2 Matéralisation de l'information mutuelle

Définition 17. Soient X, Y, Z trois variables aléatoires. Alors on dit que Z *matérialise* l'information mutuelle de X et Y lorsque Z satisfait les propriétés suivantes :

$$H(Z) = I(X : Y) \qquad H(Z \mid X) = 0 \qquad H(Z \mid Y) = 0$$

La terminologie est justifiée par le diagramme de la Figure 10 sur la page suivante. La v.a. Z « contient » en quelque sorte toute l'information que X et Y partagent, et aucune autre information. Dans la littérature, on appelle *information commune* (voir par exemple Gács & Körner [3]) la fraction maximale d'information mutuelle que l'on peut matérialiser par une variable aléatoire. En pratique, ces deux notions ne coïncident presque jamais, c'est-à-dire que l'information mutuelle n'est en général pas matérialisable entièrement : en effet c'est une propriété lourde de conséquences, comme l'illustre le prochain lemme.

Lemme 20. Supposons avoir deux v.a. X, Y dont l'information mutuelle est matérialisée par Z. Alors pour toutes v.a. A et B, les v.a. X, Y, A, B satisfont l'inégalité de Ingleton :

$$I(X : Y) \leqslant I(X : Y \mid A) + I(X : Y \mid B) + I(A : B)$$

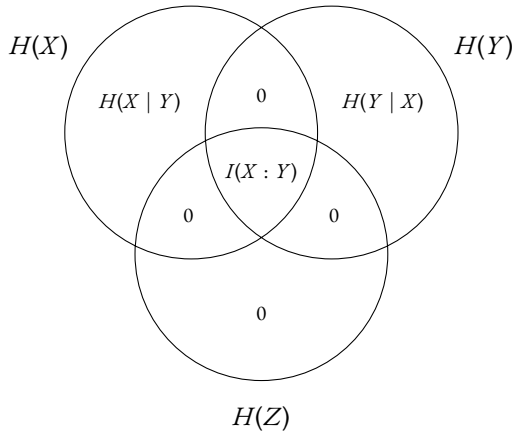

FIGURE 10 – Matérialisation de l'information mutuelle de X et Y par Z.

Démonstration.. Étant donné deux v.a. A et B quelconques, nous avons :

$$H(Z) + H(A,B) \leqslant \overbrace{H(Z) + H(A,B,Z) \leqslant H(Z,A) + H(Z,B)}^{\text{inégalité basique (7)}}$$

Nous pouvons réécrire cette inégalité sous la forme suivante :

$$H(Z) \leqslant \overbrace{H(Z,A) - H(A)}^{= H(Z \mid A)} + \overbrace{H(Z,B) - H(B)}^{= H(Z \mid B)} + \overbrace{H(A) + H(B) - H(A,B)}^{= I(A : B)}$$

Or $H(Z) = I(X : Y)$, d'où $H(Z \mid A) \leqslant I(X : Y \mid A)$ et de manière symétrique on obtient que $H(Z \mid B) \leqslant I(X : Y \mid B)$. Finalement :

$$\begin{aligned} I(X : Y) = H(Z) &\leqslant H(Z \mid A) + H(Z \mid B) + I(A : B) \\ &\leqslant I(X : Y \mid A) + I(X : Y \mid B) + I(A : B) \end{aligned}$$

C'est ce qu'il fallait démontrer. ∎

COROLLAIRE. L'information mutuelle n'est en général pas matérialisable.

Le même résultat a été démontré pour la complexité de Kolmogorov par Gács & Körner dans les années 1970 [3]. Il n'est donc pas possible en général de matérialiser l'information mutuelle, au risque de rendre vraie l'inégalité de Ingleton. Pourtant cette inégalité est en fait *presque* vraie, pour peu que l'on ajoute des termes supplémentaires : c'est le premier exemple d'inégalité d'information non-classique.

7.3 Inégalité de Zhang & Yeung

Théorème 6 (Zhang & Yeung [12]). Pour toutes v.a. X, Y, A, B :

$$\begin{aligned} I(X:Y) \leqslant\ & I(X:Y \mid A) + I(X:Y \mid B) + I(A:B) \\ & + I(X:Y \mid A) + I(X:A \mid Y) + I(A:Y \mid X) \end{aligned}$$

La première partie de l'inégalité de Zhang & Yeung correspond bien à l'inégalité de Ingleton. Un simple procédé algorithmique permet de vérifier que cette inégalité n'est *pas* classique (cf. lemme 15). Il nous reste à voir pourquoi celle-ci est vraie : nous allons en donner deux démonstrations.

7.3.1 Démonstration 1

Une première idée consiste à introduire une cinquième v.a. Z. Nous allons pour cela légèrement modifier le « morceau supplémentaire », c'est-à-dire que :

$$\begin{aligned} I(X:Y) \leqslant\ & I(X:Y \mid A) + I(X:Y \mid B) + I(A:B) \\ & + I(X:Y \mid A) + I(X:A \mid Y) + I(A:Y \mid X) \end{aligned}$$

est remplacé par :

$$\begin{aligned} I(X:Y) \leqslant\ & I(X:Y \mid A) + I(X:Y \mid B) + I(A:B) \\ & + I(X:Y \mid Z) + I(X:Z \mid Y) + I(Z:Y \mid X) + 3 \cdot I(Z:A,B \mid X,Y) \end{aligned}$$

Cette nouvelle inégalité est en fait classique : il suffit de le vérifier algorithmiquement. Si nous étions capables de montrer que l'on peut se passer du terme $3 \cdot I(Z:A,B \mid X,Y)$ en général, alors nous en déduirions bien l'inégalité de Zhang & Yeung : c'est ce que nous allons faire. Nous allons pour cela construire une v.a. Z' telle que :

$$\begin{aligned} I(X:Y \mid Z') &= I(X:Y \mid Z) & I(X:Z' \mid Y) &= I(X:Z \mid Y) \\ I(Z:Y \mid X) &= I(Z:Y \mid X) & I(Z':A,B \mid X,Y) &= 0 \end{aligned}$$

Pour cela, nous allons construire une v.a. Z' de telle sorte à ce que celle-ci est « presque » interchangeable avec Z (i.e. elles ont toutes les deux les mêmes distributions dans certains cas), avec cependant la contrainte supplémentaire que Z' est indépendante à A et B sachant X, Y. Pour cela, nous allons partir de la condition d'indépendance, puis sommer les probabilités afin de compléter à l'aide de la formule de Bayes les valeurs qui nous manque, et nous pourrons

alors constater qu'en effet Z' a bien les propriétés voulues :

$$\begin{aligned}
\mathbf{Pr}(A, B, Z' \mid X, Y) &\overset{\text{def}}{=} \mathbf{Pr}(A, B \mid X, Y) \cdot \mathbf{Pr}(Z \mid X, Y) \\
\mathbf{Pr}(X, Y, Z') &= \mathbf{Pr}(X, Y) \cdot \sum_{A,B} \mathbf{Pr}(A, B, Z' \mid X, Y) \\
&= \mathbf{Pr}(X, Y) \cdot \overbrace{\sum_{A,B} \mathbf{Pr}(A, B \mid X, Y)}^{=1} \cdot \mathbf{Pr}(Z \mid X, Y) \\
&= \mathbf{Pr}(X, Y, Z)
\end{aligned}$$

On observe bien que Z et Z' sont interchangeables relativement à X et Y puisque par exemple :

$$\begin{aligned}
\mathbf{Pr}(X \mid Z') &= \frac{\sum_Y \mathbf{Pr}(X, Y, Z')}{\sum_{X,Y} \mathbf{Pr}(X, Y, Z')} = \frac{\sum_Y \mathbf{Pr}(X, Y, Z)}{\sum_{X,Y} \mathbf{Pr}(X, Y, Z)} = \mathbf{Pr}(X \mid Z) \\
\mathbf{Pr}(Y \mid Z') &= \frac{\sum_X \mathbf{Pr}(X, Y, Z')}{\sum_{X,Y} \mathbf{Pr}(X, Y, Z')} = \frac{\sum_X \mathbf{Pr}(X, Y, Z)}{\sum_{X,Y} \mathbf{Pr}(X, Y, Z)} = \mathbf{Pr}(Y \mid Z) \\
&\vdots
\end{aligned}$$

Finalement $I(X : Y \mid Z') = I(X : Y \mid Z), \ldots$ ainsi que $I(Z' : A, B \mid X, Y) = 0$ par construction. L'inégalité suivante est donc vraie :

$$\begin{aligned}
I(X : Y) &\leqslant I(X : Y \mid A) + I(X : Y \mid B) + I(A : B) \\
&\quad + I(X : Y \mid Z') + I(X : Z' \mid Y) + I(Z' : Y \mid X) + 3 \cdot I(Z' : A, B \mid X, Y) \\
&= I(X : Y \mid A) + I(X : Y \mid B) + I(A : B) \\
&\quad + I(X : Y \mid Z) + I(X : Z \mid Y) + I(Z : Y \mid X)
\end{aligned}$$

En posant $Z = A$, nous avons le résultat voulu. ■

7.3.2 Démonstration 2

Une deuxième idée consiste à « presque » matérialiser l'information mutuelle de X, Y avec A. Voici un autre lemme dans le même esprit que le lemme 20 :

Lemme 21. Supposons qu'il existe une v.a. Z qui matérialise l'information mutuelle de X, Y avec A comme sur la Figure 11 ci-après. Alors pour n'importe quelle v.a. B supplémentaire, les quatre v.a. X, Y, A, B satisfont l'inégalité de Zhang & Yeung.

Démonstration.. Par un raisonnement identique à la démonstration du lemme 20, nous avons les inégalités suivantes :

$$\begin{aligned}
H(Z) &\leqslant H(Z \mid X) + H(Z \mid Y) + I(X : Y) \\
H(Z) &\leqslant H(Z \mid A) + H(Z \mid B) + I(A : B)
\end{aligned}$$

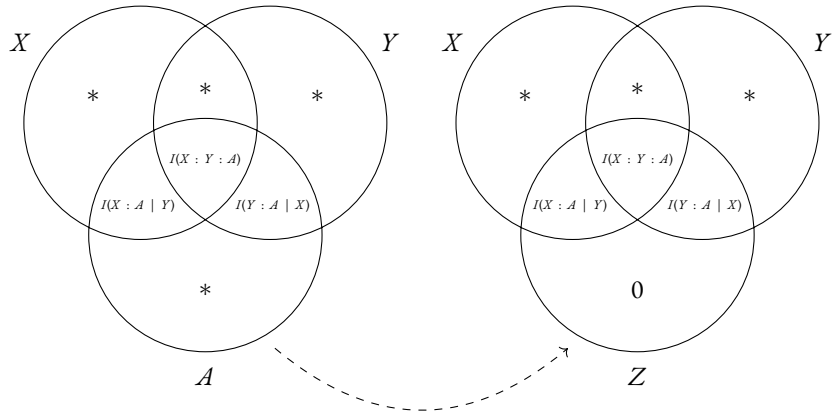

FIGURE II – Matérialisation de l'information mutuelle de X, Y avec A.

Ainsi en relativisant la première inégalité à A, on en déduit :

$$\begin{aligned} H(Z \mid A) &\leqslant H(Z \mid X, A) + H(Z \mid Y, A) + I(X : Y \mid A) \\ &\leqslant H(Z \mid X) + H(Z \mid Y) + I(X : Y \mid A) \end{aligned}$$

et de même en relativisant à B. Nous pouvons le réinjecter dans la deuxième inégalité, et ainsi obtenir :

$$\begin{aligned} H(Z) \leqslant\ & H(Z \mid X) + H(Z \mid Y) + I(X : Y \mid A) \\ & + H(Z \mid X) + H(Z \mid Y) + I(X : Y \mid B) + I(A : B) \end{aligned}$$

En simplifiant l'expression, il nous reste :

$$\begin{aligned} I(X : Y : A) \leqslant\ & I(X : Y \mid A) + I(X : Y \mid B) + I(A : B) \\ & + I(X : A \mid Y) + I(Y : A \mid X) \end{aligned}$$

ce qui donne bien l'inégalité de Zhang & Yeung en ajoutant $I(X : Y \mid A)$ de part et d'autre. ■

La prochaine étape consiste à superposer N copies i.i.d. des v.a. X, Y et A puis de matérialiser à $o(N)$ près leur information mutuelle, ce qui permettra de conclure en divisant par N et en passant à la limite sur l'inégalité à la fin de la démonstration du lemme 21. Un dernier lemme, attribué à Ahlswede & Körner, nous donne directement le résultat voulu :

Lemme 22 (Ahlswede & Körner [1], [7, p. 152] pour une démonstration). Étant donné trois v.a. X, Y, A, considérons N copies i.i.d. $\langle X_1, Y_1, A_1\rangle, \ldots, \langle X_N, Y_N, A_N\rangle$ de leur triplet. On définit par ailleurs :

$$X^N \overset{\text{def}}{=} \langle X_1, X_2, \ldots, X_N\rangle$$
$$Y^N \overset{\text{def}}{=} \langle Y_1, Y_2, \ldots, Y_N\rangle$$
$$A^N \overset{\text{def}}{=} \langle A_1, A_2, \ldots, A_N\rangle$$

les « N-puissances » de X, Y et A. Alors il existe une v.a. Z qui matérialise l'information mutuelle de X^N, Y^N avec A^N *à* $o(N)$ *près* (cf. Figure 12 ci-après).

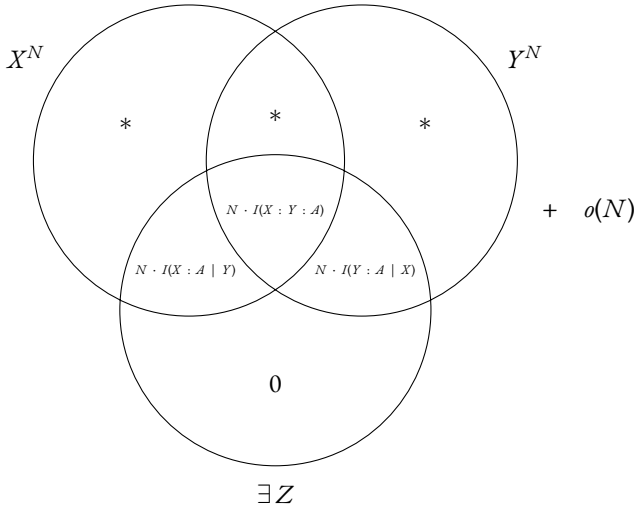

Figure 12 – Matérialisation à $o(N)$ près.

■

8 Suites aléatoires

L'un des points de vue de la complexité de Kolmogorov est que celle-ci représente la difficulté « absolue » ou « intrinsèque » de décrire un objet *fini*. Naturellement, la prochaine question est de savoir si l'on peut développer un point de vue similaire pour des suites *infinies* de 0 et de 1. Une première approche, peut-être naïve, serait de dire qu'une suite infinie est difficile à décrire — on dit plus simplement que la suite est *aléatoire* — lorsque tous ses préfixes le sont à partir d'un certain rang, c'est-à-dire qu'ils ont une grande complexité de Kolmogorov :

$$A \text{ est aléatoire} \iff \forall\, n,\ K(A \upharpoonright n) \geqslant n - O(1) \tag{10}$$

Le mathématicien Per Martin-Löf a proposé dans les années 1960 une approche différente : une suite est aléatoire lorsqu'elle n'est pas *effectivement* capturée par une suite de « propriétés », ce que nous allons voir juste après. Il se trouve que ces deux approches sont équivalentes, ce qui montre finalement que l'aléatoire est une notion robuste. Les notations suivantes nous seront utiles dans cette (dernière !) section :

Notation. On note $A \upharpoonright n$ la chaîne composée des n premiers bits d'une suite A.

Si A et B sont deux suites, alors $A \oplus B$ désigne la suite qui est égale à A sur les indices pairs, et B sur les indices impairs. La suite $A \oplus B$ est appelée le *join* de A et B.

8.1 Aléatoire au sens de Martin-Löf

Rappelons-nous que l'espace de Cantor Ω contient toutes les suites infinies de 0 et de 1, et est muni d'une mesure de probabilités $\square$. L'idée de Martin-Löf est la suivante : quelle que soit la définition que l'on se donnerait de l'aléatoire, si nous étions capables de parfaitement choisir une suite au hasard alors celle-ci serait presque sûrement aléatoire. Cela signifie en particulier que l'ensemble des suites aléatoires est de mesure 1. Si l'on regarde deux ensembles de mesure 1 alors ils ne diffèrent nécessairement que d'un ensemble de mesure nulle, et dans ce cas s'il existait un plus petit ensemble de mesure 1 pour l'inclusion alors par hypothèse il ne contiendrait aucun élément « remarquable », i.e. appartenant à un ensemble de mesure nulle, et cet ensemble serait dans ce cas un excellent candidat à la définition de suite aléatoire. Bien sûr tous les singletons sont de mesure nulle, et donc si nous n'imposons aucune restriction supplémentaire un plus petit ensemble de mesure 1 pour l'inclusion n'existe pas. Martin-Löf propose alors de se limiter aux ensembles de mesure nulle *effectifs*, c'est-à-dire qu'ils sont la limite d'une suite décroissante d'ouverts énumérés par une machine de Turing. C'est cette définition qui va fonctionner :

Définition 18. Une famille $(\mathcal{U}_i)$ d'ouverts est *effective* s'il existe une machine de Turing M qui énumère *uniformément* ses ouverts :

$$\mathcal{U}_i = \bigcup_n [M(i, n)]$$

Définition 19. Une famille effective d'ouverts $(\mathcal{U}_i)$ est un *test de Martin-Löf* si elle satisfait les propriétés suivantes :

i) $\mathcal{U}_{i+1} \subseteq \mathcal{U}_i$

ii) $\square(\mathcal{U}_i) \leqslant 2^{-i}$

On dit qu'une suite A *échoue* le test de Martin-Löf si $A \in \bigcap \mathcal{U}_i$.

On a immédiatement $\square(\bigcap \mathcal{U}_i) = 0$ dans la définition précédente. Nous voulons exactement qu'une suite soit aléatoire lorsqu'elle n'est pas « remarquable », c'est-à-dire qu'elle n'est pas contenue dans un ensemble de mesure nulle :

Définition 20. Une suite A est *aléatoire au sens de Martin-Löf* (abrégé M.L.R. par la suite, pour « *Martin-Löf Random* ») si elle n'échoue aucun test de Martin-Löf.

Comme il n'existe qu'un nombre dénombrable de machines de Turing, les tests de Martin-Löf n'induisent qu'un nombre dénombrable d'ensembles de mesure nulle, et par conséquent leur réunion est également de mesure nulle. Nous pouvons en quelque sorte reformuler l'ensemble des suites aléatoires comme le plus petit ensemble « co-effectif » de mesure 1, ce qui cristallise bien dans ce cas l'intuition de départ. Il se trouve que la réunion des limites de tests de Martin-Löf est elle-même la limite d'un test de Martin-Löf, que l'on qualifie d'*universel* :

Proposition 17. Il existe un test de Martin-Löf *universel* tel que si une suite n'échoue pas le test universel, alors elle est aléatoire au sens de Martin-Löf.

Démonstration.. Le principe est analogue à une construction que nous avons vue précédemment (p. 18) : pour transformer une machine de Turing en une machine de Turing *préfixe* on teste tous les préfixes possibles et on s'arrête au premier préfixe dont l'exécution se termine. En particulier, cela laisse invariant les machines de Turing préfixe, ce qui rend possible de *toutes* les énumérer, et uniquement elles.

Nous allons procéder de la façon suivante pour transformer une machine de Turing M en un test de Martin-Löf :

i) Exécuter tous les $M(i, n)$ en parallèle

ii) Si l'exécution d'un $M(i, n)$ se termine, alors on affiche $M(i, n)$ s'il vérifie les deux conditions suivantes : on regarde d'une part s'il a déjà été énuméré pour tous les $M(j, -)$ avec $j < i$, et d'autre part on s'assure que ajouter $\square([M(i, n)])$ à ce qui a déjà été énuméré au rang i ne fait pas dépasser le poids total de 2^{-i}

iii) Si on décide d'afficher $M(i, n)$, alors on vérifie de nouveau tous les $M(j, -)$ avec cette fois $j > i$ pour lesquels l'exécution s'est terminée, et qui pourraient éventuellement maintenant satisfaire les deux conditions de l'étape ii)

Il n'y a aucune difficulté à vérifier que cette procédure produit bien des tests de Martin-Löf, et qu'en particulier elle laisse invariant les machines de Turing qui sont déjà des tests de Martin-Löf. Nous pouvons dans ce cas *effectivement* énumérer les tests de Martin-Löf comme une suite (T_e) de machines de Turing. Le test universel $(\mathcal{U}_i)$ est alors défini de la manière suivante :

$$\mathcal{U}_i \overset{\text{def}}{=} \bigcup_e \bigcup_n [T_e(e + i + 1, n)]$$

Là encore il n'y a aucune difficulté à vérifier qu'il s'agit bien d'un test de Martin-Löf :

$$\begin{aligned}\square(\mathcal{U}_i) = \square\left(\bigcup_e \bigcup_n [T_e(e + i + 1, n)]\right) &\leqslant \sum_e \square\left(\bigcup_n [T_e(e + i + 1, n)]\right) \\ &\leqslant \sum_e 2^{-e-i-1} = 2^{-i} \cdot \sum_e 2^{-e-1} \\ &\leqslant 2^{-i}\end{aligned}$$

Par construction, nous avons pour tout e et pour tout i :

$$\bigcup_n [T_e(e + i + 1, n)] \subseteq \mathcal{U}_i$$

Cela impose bien que $\bigcap \mathcal{U}_i$ contient les limites de tous les tests de Martin-Löf, ce qui montre le résultat voulu. ■

8.2 Semi-mesures

Nous allons introduire un nouvel outil qui se prête bien à manipuler des suites infinies, et qui nous permettra de faire le lien ensuite avec la complexité de Kolmogorov :

Définition 21. Une fonction ν définie de $\mathbb{N}$ dans $\mathbb{R}^+$ est une *semi-mesure* si elle possède les deux propriétés suivantes :

i) ν est l.s.-c. (cf. définition 8)

ii) $\sum_n \nu(n) \leqslant 1$

On peut imaginer une semi-mesure comme un processus qui s'exécute éternellement, et qui de temps en temps vient accumuler une masse sur un entier naturel par propriété d'être l.s.-c. sans jamais dépasser le total de 1. Les semi-mesures sont étudiées en détail dans le livre de Shen et al. [10, p. 75] : elles y ont le nom de « semi-mesures discrètes » pour les distinguer des « semi-mesures continues » qui ont une définition similaire sur l'ensemble des chaînes $2^{<\mathbb{N}}$. Les semi-mesures ont de nombreuses bonnes propriétés qui remplissent au moins un chapitre entier d'un livre, si bien que nous ne regarderons ici que l'essentiel. Voici un premier résultat intéressant, important et non trivial qui va nous servir ensuite :

Proposition 18. Il existe une semi-mesure $\widetilde{m}$, appelée *semi-mesure universelle*, qui est plus grande que toutes les autres à constante multiplicative près.

Démonstration.. Une fois encore, nous reprenons l'idée de transformer toutes les machines de Turing en « élaguant » les comportements indésirables afin d'obtenir une énumération de toutes les semi-mesures. Étant donné une machine de Turing M, nous effectuons les opérations suivantes :

i) Exécuter tous les $M(n)$ en parallèle

ii) Lorsqu'un $M(n)$ termine son exécution, alors on affiche $M(n)$ seulement si la masse totale de ce que l'on a déjà affiché ne dépasse pas $1 - M(n)$

Cela produit bien des semi-mesures par construction, et il est trivial de voir que les semi-mesures sont laissées invariantes par cette transformation. Nous pouvons donc supposer avoir une énumération effective (ν_i) de toutes les semi-mesures (avec éventuellement des doublons). Alors on définit $\widetilde{m}$ par :

$$\widetilde{m} \overset{\text{def}}{=} \sum_i 2^{-i-1} \cdot \nu_i$$

C'est la *moyenne pondérée* de toutes les semi-mesures. Par construction, nous avons bien que $\widetilde{m}$ est l.s.-c. d'une part, et d'autre part $\widetilde{m}$ satisfait la condition ii) de la définition d'une semi-mesure :

$$\begin{aligned}\sum_n \widetilde{m}(n) &= \sum_n \sum_i 2^{-i-1} \cdot \nu_i(n) \\ &= \sum_i \sum_n 2^{-i-1} \cdot \nu_i(n) \\ &\leqslant \sum_i 2^{-i-1} \cdot 1 = 1\end{aligned}$$

Enfin, il est immédiat que pour tout i :

$$2^{i+1} \cdot \widetilde{m} \geqslant \nu_i$$

Nous avons montré que $\widetilde{m}$ est bien une semi-mesure ayant les propriétés voulues. ■

Nous avons vu dans la section correspondante que la complexité de Kolmogorov a la propriété d'être u.s.-c., ainsi que de produire des descriptions plus petites à constante près que n'importe quelle autre machine de Turing. Cela n'est pas sans rappeler les propriétés de la semi-mesure universelle, et ce n'est pas un hasard :

Théorème 7. $K(n) = -\log \widetilde{m}(n) + O(1)$

Démonstration.. Il est assez facile de voir d'une part que la fonction $n \longmapsto 2^{-K(n)}$ est en fait une semi-mesure. Elle est bien l.s.-c. car la complexité de Kolmogorov est u.s.-c., donc en inversant le signe l'approximation change également de sens. Par propriété de la machine universelle pour K d'être préfixe, toutes les plus petites descriptions sont incompatibles, et par conséquent elles induisent des cylindres disjoints de l'espace de Cantor. Dans ce cas la somme de leur mesure et nécessairement inférieure à 1. Ceci nous permet de conclure $2^{-K(n)} \leqslant O(1) \cdot \widetilde{m}(n)$ puisque $\widetilde{m}$ est maximale à constante multiplicative près, ce que l'on peut réécrire comme $-\log \widetilde{m}(n) + O(1) \leqslant K(n)$. D'autre part puisque $-\log \widetilde{m}(n) \leqslant \ell_n \overset{\text{def}}{=} \lceil -\log \widetilde{m}(n) \rceil = -\log \widetilde{m}(n) + O(1)$ par propriété de l'arrondi supérieur, nous avons l'inégalité :

$$\sum_n 2^{-\ell_n} \leqslant \sum_n 2^{-\left(-\log \widetilde{m}(n)\right)} = \sum_n \widetilde{m}(n) \leqslant 1$$

Les entiers ℓ_n sont u.s.-c., nous pouvons supposer avoir une énumération effective $\ell_{n,i}$ telle que $\ell_{n,0} > \ell_{n,1} > \cdots > \ell_{n,N} = \ell_n$ pour N assez grand, et $\ell_{n,N+1}, \ell_{n,N+2}, \ldots$ ne sont pas définis. Dans ce cas il apparaît que :

$$\sum_i 2^{-\ell_{n,i}-1} \leqslant 2^{-\ell_n}$$

car les $\ell_{n,i} + 1$ sont tous distincts strictement plus grands que ℓ_n. Finalement il est possible d'applique le lemme de Kraft-Chaitin (p. 2) aux $\ell_{n,i} + 1$ puisqu'ils satisfont l'inégalité :

$$\sum_n \sum_i 2^{-\ell_{n,i}-1} \leqslant \sum_n 2^{-\ell_n} \leqslant 1$$

On obtient ainsi une énumération effective de chaînes $\tau_{n,i}$ telles que $|\tau_{n,i}| = \ell_{n,i} + 1$. Nous avons alors que la machine M définie par :

$$M(x) \overset{\text{def}}{=} \begin{cases} n \text{ s'il existe } n \text{ et } i \text{ tel que } x = \tau_{n,i} \\ \text{ne s'arrête jamais sinon} \end{cases}$$

est préfixe. Mais alors, étant donné $\tau_{n,N}$ tel que $|\tau_{n,N}| = \ell_n + 1$ pour N assez grand, par définition $M(\tau_{n,N}) = n$, i.e. n possède une description de longueur $\ell_n + O(1)$. Nous en déduisons enfin :

$$K(n) \leqslant |\tau_{n,N}| + O(1) = \lceil -\log \widetilde{m}(n) \rceil + O(1) = -\log \widetilde{m}(n) + O(1) \qquad \blacksquare$$

Remarque. Dans le théorème précédent nous avons implicitement supposé savoir comment mettre en bijection effective les entiers naturels avec les chaînes, ce qui n'est pas difficile à faire.

Ainsi les propriétés de la semi-mesure universelle coïncident avec celles de la complexité de Kolmogorov préfixe. Nous pouvons utiliser ce nouveau langage pour redémontrer plus simplement certains résultats vus précédemment :

Proposition 19 (cf. lemme 7). $K(x, y) \leqslant K(x) + K(y) + O(1)$

Démonstration.. Cela revient à montrer que $O(1) \cdot \widetilde{m}(x, y) \geqslant \widetilde{m}(x) \cdot \widetilde{m}(y)$. Pour cela nous définissons une semi-mesure ν de la manière suivante :

$$\nu(z) \overset{\text{def}}{=} \begin{cases} \widetilde{m}(x) \cdot \widetilde{m}(y) & \text{si } z \text{ est l'encodage de la paire } \langle x, y \rangle \\ 0 & \text{sinon} \end{cases}$$

Elle est clairement l.s.-c., il reste donc à vérifier que la somme ne dépasse pas 1 :

$$\sum_z \nu(z) = \sum_{x,y} \widetilde{m}(x) \cdot \widetilde{m}(y) = \sum_x \widetilde{m}(x) \cdot \sum_y \widetilde{m}(y) \leqslant 1 \cdot 1 = 1$$

C'est ce que l'on voulait. $\blacksquare$

Voici une autre propriété qui n'a rien d'une évidence à première vue, mais dont la démonstration est élémentaire avec les semi-mesures :

Proposition 20. Il existe une constante d telle que pour toute chaîne x, le nombre de plus petites descriptions de x est inférieur à d.

Démonstration.. Notons $\tau_1, \ldots, \tau_N$ toutes les plus petites descriptions d'une chaîne x, c'est-à-dire que $|\tau_1| = \ldots = |\tau_N| = K(x)$. Alors dire que N est inférieur à une constante d est équivalent à montrer l'inégalité suivante :

$$\sum_k^N 2^{-|\tau_k|} \leqslant d \cdot 2^{-K(x)} \tag{11}$$

Si maintenant au lieu de seulement regarder les descriptions les plus courtes nous regardions *toutes* les descriptions, alors montrer l'inégalité suivante implique nécessairement l'inégalité (11) :

$$\sum_{\mathbf{U}(\tau)=x} 2^{-|\tau|} \leqslant O(1) \cdot 2^{-K(x)}$$

Pour rappel, **U** désigne la machine universelle préfixe. Puisque 2^{-K} et $\widetilde{m}$ sont interchangeables à constante multiplicative près, cela signifie qu'il nous suffit de montrer :

$$\sum_{\mathbf{U}(\tau)=x} 2^{-|\tau|} \leqslant O(1) \cdot \widetilde{m}(x)$$

c'est-à-dire finalement que la fonction $x \longmapsto \sum_{\mathbf{U}(\tau)=x} 2^{-|\tau|}$ est une semi-mesure. Elle est clairement l.s.-c. puisqu'au fur et à mesure que les exécutions de tous les τ sur la machine **U** se terminent nous pouvons ajouter la masse $2^{-|\tau|}$. Enfin, puisque **U** est une machine préfixe, tous les τ sur lesquels la machine **U** s'arrête sont incompatibles. On en déduit que :

$$\sum_{x} \sum_{\mathbf{U}(\tau)=x} 2^{-|\tau|} = \sum_{\mathbf{U}(\tau)\text{ s'arrête}} 2^{-|\tau|} \leqslant 1$$

puisque les τ induisent des cylindres disjoints de l'espace de Cantor. ■

À titre indicatif, la démonstration d'une version plus fine du théorème de Kolmogorov-Levin que nous avions déjà énoncé dans une remarque précédente (p. 21) est disponible dans le livre de Shen et al. [10], et fait entre autres abondamment usage des semi-mesures :

Fait (Kolmogorov-Levin, [10, p. 106-108]). $K(x, y) = K(x) + K(y \mid x, K(x)) + O(1)$

8.3 Deux derniers théorèmes

Nous allons enfin voir que l'aléatoire au sens de Martin-Löf coïncide avec l'idée naïve que nous avions présenté avec l'équivalence (10).

Théorème 8 (Levin-Schnorr). Une suite A est M.L.R. si, et seulement s'il existe une constante d telle que $K(A \upharpoonright n) \geqslant n - d$ pour tout n.

Démonstration.. Nous allons le démontrer par contraposée. Dans le sens direct, supposons que pour toute constante d, il existe n tel que $K(A \upharpoonright n) < n - d$. Cette propriété peut se réécrire de la manière suivante, en utilisant la semi-mesure universelle :

$$\widetilde{m}(A \upharpoonright n) > 2^{-n+d+c}$$

avec une constante c qui traduit la correspondance entre K et $-\log \widetilde{m}$. On définit une famille effective $(\mathcal{U}_d)$ comme suit :

$$\mathcal{U}_d \overset{\text{def}}{=} \bigcup_\tau [\tau] \quad \text{avec } \tau \text{ tels que} \quad \widetilde{m}(\tau) > 2^{-|\tau|+d+c}$$

Elle est bien effective, puisqu'en calculant $\widetilde{m}(\tau)$ on finit par se rendre compte si la masse accumulée dépasse $2^{-|\tau|+d+c}$. Par ailleurs, nous avons :

$$1 \geqslant \sum_\tau \widetilde{m}(\tau) > \sum_\tau 2^{-|\tau|+d+c}$$

en sommant sur les τ énumérés dans $\mathcal{U}_d$. D'où finalement :

$$2^{-d-c} > \sum_\tau 2^{-|\tau|} = \square(\mathcal{U}_d)$$

Par conséquent $(\mathcal{U}_d)$ définit un test de Martin-Löf que A échoue par construction, ainsi A n'est pas aléatoire au sens de Martin-Löf. Réciproquement, supposons que A ne soit pas M.L.R., c'est-à-dire qu'il existe un test de Martin-Löf $(\mathcal{U}_d)$ que A échoue : nous allons transformer ce test en une semi-mesure et déduire le résultat voulu. On définit la fonction ν de la manière suivante :

$$\nu(\tau) \overset{\text{def}}{=} \begin{cases} 2^{-|\tau|+d+c} & \text{s'il existe } d \text{ tel que } \tau \text{ est énuméré dans } \mathcal{U}_{2\cdot d+c+1} \\ 0 & \text{sinon} \end{cases}$$

où c désigne la même constante que dans le sens direct. Elle est l.s.-c. puisqu'on finit par se rendre compte si τ existe dans l'énumération d'un $\mathcal{U}_{2\cdot d+c+1}$. D'autre part :

$$\begin{aligned} \sum_\tau \nu(\tau) &= \sum_d \sum_\tau 2^{-|\tau|+d+c} \\ &= \sum_d 2^{d+c} \cdot \square(\mathcal{U}_{2\cdot d+c+1}) \\ &\leqslant \sum_d 2^{d+c} \cdot 2^{-2\cdot d-c-1} = \sum_d 2^{-d-1} = 1 \end{aligned}$$

avec τ énuméré dans $\mathcal{U}_{2\cdot d+c+1}$ dans la somme à droite. Cela montre que ν est une semi-mesure, et dans ce cas il existe une constante e telle que :

$$\widetilde{m} \geqslant \nu \cdot 2^{-e}$$

En particulier, pour tout d il existe n tel que $A \upharpoonright n$ est énuméré dans $\mathcal{U}_{d+c+e+1}$, d'où :

$$\begin{aligned} \widetilde{m}(A \upharpoonright n) &\geqslant \nu(A \upharpoonright n) \cdot 2^{-e} \\ &= 2^{-n+d+c+e+1} \cdot 2^{-e} \\ &= 2^{-n+d+c+1} \\ &> 2^{-n+d+c} \end{aligned}$$

ce qui est équivalent à $K(A \upharpoonright n) < n - d$. ∎

L'aléatoire est par conséquent une notion robuste : nous pouvons dire qu'une suite est M.L.R. sans ambiguïté. Comme beaucoup de choses en calculabilité, il est possible de *relativiser* la définition de M.L.R. à la connaissance d'un oracle : relativiser signifie par exemple d'une part que l'on sait mieux compresser des préfixes de suites infinies, et d'autre part que l'on sait énumérer de nouveaux ouverts effectifs, et ainsi créer de nouveaux tests de Martin-Löf. Le théorème de Levin-Schnorr s'avère être encore vrai étant donné un oracle : pour le démontrer il faudrait reprendre toutes les propriétés introduites dans cette section et les reformuler avec un oracle, nous allons donc admettre ce résultat pour la suite.

Définition 22. On dit qu'une suite est A-M.L.R. si elle est aléatoire relativisée à A.

Être A-M.L.R. est strictement plus fort qu'être simplement M.L.R., puisque l'oracle génère potentiellement de nouveaux tests de Martin-Löf qu'une suite ne doit pas échouer. Le dernier théorème que nous allons seulement énoncer est remarquable en ce qu'il établit un parallèle direct avec les théorèmes de symétrie de l'information et de Kolmogorov-Levin (que nous savons déjà équivalents), pointant du doigt la question de recherche suivante : « *Existe-t-il une équivalence entre les énoncés logiques portant sur les M.L.R. et les inégalités d'information ?* »

Fait (van Lambalgen [6], [2, p. 258]). $A \oplus B$ est M.L.R. si, et seulement si A est M.L.R. et B est A-aléatoire au sens de Martin-Löf.

Références

1. AHLSWEDE, Rudolf & KÖRNER, János. On the connection between the entropies of input and output distributions of discrete memoryless channels. In : *Proceedings of the fifth Conference on Probability Theory*. 1977.
2. DOWNEY, Rodney G & HIRSCHFELDT, Denis R. *Algorithmic randomness and complexity*. Springer Science & Business Media, 2010.
3. GÁCS, Peter & KÖRNER, János. Common information is far less than mutual information. *Problems of Control and Information Theory*. 1973, t. 2, n° 2, p. 149-162.
4. HAMMER, Daniel ; ROMASHCHENKO, Andrei ; SHEN, Alexander & VERESHCHAGIN, Nikolai. Inequalities for Shannon entropy and Kolmogorov complexity. *Journal of Computer and System Sciences*. 2000, t. 60, n° 2, p. 442-464.
5. INGLETON, Aubrey W. Representation of matroids. *Combinatorial mathematics and its applications*. 1971, t. 23.
6. LAMBALGEN, Michiel van. The axiomatization of randomness. *The Journal of Symbolic Logic*. 1990, t. 55, n° 3, p. 1143-1167.
7. MAKARYCHEV, Konstantin ; MAKARYCHEV, Yury ; ROMASHCHENKO, Andrei & VERESHCHAGIN, Nikolai. A new class of non-Shannon-type inequalities for entropies. *Communications in Information and Systems*. 2002, t. 2, n° 2, p. 147-166.
8. SHANNON, Claude Elwood. A mathematical theory of communication. *The Bell system technical journal*. 1948, t. 27, n° 3, p. 379-423.
9. SHEN, Alexander. Aksiomaticheskoe opisanie ponyatiya entropii konechnogo objekta (An axiomatic description of the notion of entropy of finite objects). In : *Logic and Foundations of Mathematics* [Abstracts of the 8th All-union conference « Logic and methodology of science »]. 1982, p. 104-105. (En russe).
10. SHEN, Alexander ; USPENSKY, Vladimir A & VERESHCHAGIN, Nikolay. *Kolmogorov complexity and algorithmic randomness*. T. 220. American Mathematical Soc., 2017.
11. YEUNG, Raymond W. *Information theory and network coding*. Springer Science & Business Media, 2008.
12. ZHANG, Zhen & YEUNG, Raymond W. On characterization of entropy function via information inequalities. *IEEE Transactions on Information Theory*. 1998, t. 44, n° 4, p. 1440-1452.

13. ZVONKIN, Alexander K & LEVIN, Leonid A. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*. 1970, t. 25, nº 6, p. 83.