Information, inégalités et relations d'incertitude

Steeve Zozor

GIPSA-Lab – CNRS – Grenoble INP, Grenoble, France

17-21 Mai 2010

Plan du cours

1. Entropie – mesure d'incertitude
   - Axiomes, entropie de Shannon et ses propriétés
   - Entropie différentielle – cas vectoriel
   - D'autres mesures d'information
2. Information mutuelle - divergences
   - Entropie conditionnelle, Information mutuelle
   - Quelques exemples
   - Divergences
3. Inégalités – relations entropiques
   - Inégalités "classiques"
   - Chaînes de Markov
   - Inégalité de la puissance entropique
   - Estimation – relations entre informations
4. Relations d'incertitude
   - Relation d'incertitude d'Heisenberg et versions entropiques

Plan du cours

1. Entropie – mesure d'incertitude
   - Axiomes, entropie de Shannon et ses propriétés
   - Entropie différentielle – cas vectoriel
   - D'autres mesures d'information
2. Information mutuelle - divergences
   - Entropie conditionnelle, Information mutuelle
   - Quelques exemples
   - Divergences
3. Inégalités – relations entropiques
   - Inégalités "classiques"
   - Chaînes de Markov
   - Inégalité de la puissance entropique
   - Estimation – relations entre informations
4. Relations d'incertitude
   - Relation d'incertitude d'Heisenberg et versions entropiques

Plan du cours

1. **Entropie – mesure d'incertitude**
   - Axiomes, entropie de Shannon et ses propriétés
   - Entropie différentielle – cas vectoriel
   - D'autres mesures d'information
2. **Information mutuelle - divergences**
   - Entropie conditionnelle, Information mutuelle
   - Quelques exemples
   - Divergences
3. **Inégalités – relations entropiques**
   - Inégalités "classiques"
   - Chaînes de Markov
   - Inégalité de la puissance entropique
   - Estimation – relations entre informations
4. **Relations d'incertitude**
   - Relation d'incertitude d'Heisenberg et versions entropiques

Plan du cours

1. Entropie – mesure d'incertitude
   - Axiomes, entropie de Shannon et ses propriétés
   - Entropie différentielle – cas vectoriel
   - D'autres mesures d'information
2. Information mutuelle - divergences
   - Entropie conditionnelle, Information mutuelle
   - Quelques exemples
   - Divergences
3. Inégalités – relations entropiques
   - Inégalités “classiques”
   - Chaînes de Markov
   - Inégalité de la puissance entropique
   - Estimation – relations entre informations
4. Relations d'incertitude
   - Relation d'incertitude d'Heisenberg et versions entropiques

Plan du cours

1. Entropie – mesure d'incertitude
   - Axiomes, entropie de Shannon et ses propriétés
   - Entropie différentielle – cas vectoriel
   - D'autres mesures d'information
2. Information mutuelle - divergences
   - Entropie conditionnelle, Information mutuelle
   - Quelques exemples
   - Divergences
3. Inégalités – relations entropiques
   - Inégalités "classiques"
   - Chaînes de Markov
   - Inégalité de la puissance entropique
   - Estimation – relations entre informations
4. Relations d'incertitude
   - Relation d'incertitude d'Heisenberg et versions entropiques

Quelques exemples

Quel est le point commun entre les problèmes suivants?

- On dispose d'un dé à 8 faces. On souhaite coder chaque face par une séquence binaire (suite de 0 et 1) suivant les contraintes:
  - à partir d'une séquence binaire correspondant à une séquence de lancers, on souhaite retrouver la séquence des faces (décodage),
  - on souhaite, "en moyenne", utiliser le moins de bits possible.
- Une pièce a une probabilité p de tomber sur pile. X est le nombre de lancers nécessaires pour tomber sur "face" et Vali vient de faire l'expérience. En lui posant le moins possible de questions binaires du type "$X \in \{1, 2, 3\}$" on souhaite deviner X.
- On dispose d'un générateur aléatoire X binaire de ± 1 dont on peut régler $\alpha = \Pr[X = 1]$. On émet une séquence à travers un canal qui distord le symbole (l'inverse, le bruite...). Comment fixer α pour "commettre le moins d'erreurs possible"?
- N particule de gaz parfait occupent un volume Λ avec chacune sa vélocité : quelle la densité de probabilité des n vitesses des particules sachant que l'énergie totale $\sum_i \frac{1}{2} m v_i^2 = E$ est fixe?

Quelques exemples

Quel est le point commun entre les problèmes suivants?

- On dispose d'un dé à 8 faces. On souhaite coder chaque face par une séquence binaire (suite de 0 et 1) suivant les contraintes:
  - à partir d'une séquence binaire correspondant à une séquence de lancers, on souhaite retrouver la séquence des faces (décodage),
  - on souhaite, "en moyenne", utiliser le moins de bits possible.
- Une pièce a une probabilité p de tomber sur pile. X est le nombre de lancers nécessaires pour tomber sur "face" et Vali vient de faire l'expérience. En lui posant le moins possible de questions binaires du type "$X \in \{1,2,3\}$" on souhaite deviner X.
- On dispose d'un générateur aléatoire X binaire de ± 1 dont on peut régler $\alpha = \Pr[X = 1]$. On émet une séquence à travers un canal qui distord le symbole (l'inverse, le bruite...). Comment fixer α pour "commettre le moins d'erreurs possible"?
- N particule de gaz parfait occupent un volume Λ avec chacune sa vélocité : quelle la densité de probabilité des n vitesses des particules sachant que l'énergie totale $\sum_i \frac{1}{2} m v_i^2 = E$ est fixe?

Quelques exemples

Quel est le point commun entre les problèmes suivants?

- On dispose d'un dé à 8 faces. On souhaite coder chaque face par une séquence binaire (suite de 0 et 1) suivant les contraintes:
  - à partir d'une séquence binaire correspondant à une séquence de lancers, on souhaite retrouver la séquence des faces (décodage),
  - on souhaite, "en moyenne", utiliser le moins de bits possible.
- Une pièce a une probabilité p de tomber sur pile. X est le nombre de lancers nécessaires pour tomber sur "face" et Vali vient de faire l'expérience. En lui posant le moins possible de questions binaires du type "$X \in \{1, 2, 3\}$" on souhaite deviner X.
- On dispose d'un générateur aléatoire X binaire de ± 1 dont on peut régler $\alpha = \Pr[X = 1]$. On émet une séquence à travers un canal qui distord le symbole (l'inverse, le bruite...). Comment fixer α pour "commettre le moins d'erreurs possible"?
- N particule de gaz parfait occupent un volume Λ avec chacune sa vélocité : quelle la densité de probabilité des n vitesses des particules sachant que l'énergie totale $\sum_i \frac{1}{2} m v_i^2 = E$ est fixe?

Quelques exemples

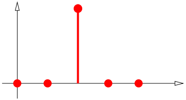
Quel est le point commun entre les problèmes suivants?

- On dispose d'un dé à 8 faces. On souhaite coder chaque face par une séquence binaire (suite de 0 et 1) suivant les contraintes:
  - à partir d'une séquence binaire correspondant à une séquence de lancers, on souhaite retrouver la séquence des faces (décodage),
  - on souhaite, "en moyenne", utiliser le moins de bits possible.
- Une pièce a une probabilité p de tomber sur pile. X est le nombre de lancers nécessaires pour tomber sur "face" et Vali vient de faire l'expérience. En lui posant le moins possible de questions binaires du type "$X \in \{1, 2, 3\}$" on souhaite deviner X.
- On dispose d'un générateur aléatoire X binaire de ± 1 dont on peut régler $\alpha = \Pr[X = 1]$. On émet une séquence à travers un canal qui distord le symbole (l'inverse, le bruite...). Comment fixer α pour "commettre le moins d'erreurs possible"?
- N particule de gaz parfait occupent un volume Λ avec chacune sa vélocité : quelle la densité de probabilité des n vitesses des particules sachant que l'énergie totale $\sum_i \frac{1}{2} m v_i^2 = E$ est fixe?

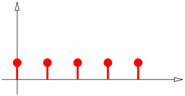
Entropie de Shannon, 1948 – Axiomes

Variable aléatoire discrète X sur $\{x_1, \ldots, x_n\}$, de loi $\{p_k = \Pr[X = x_k]\}_{k=1,\ldots,n}$

But : construire une mesure d'incertitude/d'information $H(X) = H(p_1, \ldots, p_n)$

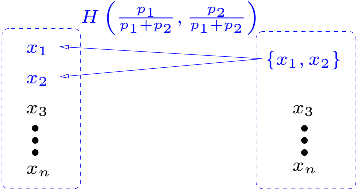


pas d'incertitude
obs. n'apporte pas d'information



incertitude maximum

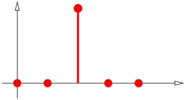
- Invariante par permutation, continue selon les p_k
- $p_k = \frac{1}{n}$ alors $H(n)$ croissante (incertitude croissante)
- Récursivité $H = H(p_1 + p_2, p_3, \ldots) + (p_1 + p_2) H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$



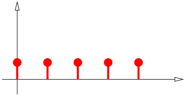
Entropie de Shannon, 1948 – Axiomes équiv.

Variable aléatoire discrète X sur $\{x_1, \ldots, x_n\}$, de loi $\{p_k = \Pr[X = x_k]\}_{k=1,\ldots,n}$

But : construire une mesure d'incertitude/d'information $H(X) = H(p_1, \ldots, p_n)$



pas d'incertitude
obs. n'apporte pas d'information



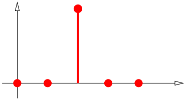
incertitude maximum

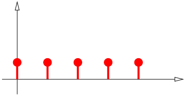
- Incertitude élémentaire – information de Hartley $h(p_k)$, $H(X) = \sum p_k h(p_k)$ continue selon les p_k et symétrique
- X, Y indépendantes, alors $H(X,Y) = H(X) + H(Y)$

Entropie de Shannon, 1948 – Définition

Variable aléatoire discrète X sur $\{x_1, \ldots, x_n\}$, de loi $\{p_k = \Pr[X = x_k]\}_{k=1,\ldots,n}$

But : construire une mesure d'incertitude/d'information $H(X) = H(p_1, \ldots, p_n)$





Intuitivement, $h(p_k q_k) = h(p_k) + h(q_k)$ impose $h(p) = \alpha \log(p)$

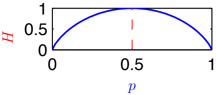
Seule possibilité : $H(X) = -kT \sum_k p_k \log(p_k)$ $\quad (0 \log 0 = 0)$

S'étend au cas $k \in \mathbb{Z}$, $kT \stackrel{\text{def}}{=} \frac{1}{\log b}$.

Quelques propriétés

- $H \geq 0$
- $p_k = \delta_{k,l} \Leftrightarrow H = 0$ minimum (pas d'incertitude)
- $p_k = \frac{1}{n} \Leftrightarrow H = \log n \Leftrightarrow$ (incertitude maximum dans le cas fini)

  e.g. $\{p, 1-p\}$



- Concavité : $P \sim \{p_k\}$, $Q \sim \{q_k\}$ et $R \sim \{\lambda p_k + (1-\lambda) q_k\}, \lambda \in [0\,;\,1]$

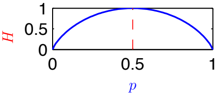
$$H(R) \geq \lambda H(P) + (1-\lambda) H(Q)$$

- $H(aX) = H(X+c) = H(X)$: indépendant des valeurs prises par X

Quelques propriétés

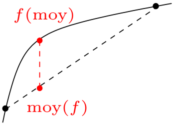
- $H \geq 0$
- $p_k = \delta_{k,l} \Leftrightarrow H = 0$ minimum (pas d'incertitude)
- $p_k = \frac{1}{n} \Leftrightarrow H = \log n \Leftrightarrow$ (incertitude maximum dans le cas fini)

  e.g. $\{p, 1-p\}$



- Concavité : $P \sim \{p_k\}$, $Q \sim \{q_k\}$ et $R \sim \{\lambda p_k + (1-\lambda) q_k\}, \lambda \in [0\,;\,1]$

$$H(R) \geq \lambda H(P) + (1-\lambda) H(Q)$$

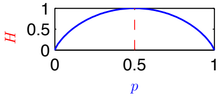


- $H(aX) = H(X+c) = H(X)$: indépendant des valeurs prises par X

Quelques propriétés

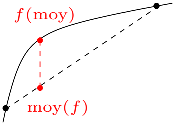
- $H \geq 0$
- $p_k = \delta_{k,l} \Leftrightarrow H = 0$ minimum (pas d'incertitude)
- $p_k = \frac{1}{n} \Leftrightarrow H = \log n \Leftrightarrow$ (incertitude maximum dans le cas fini)

  e.g. $\{p, 1-p\}$



- Concavité : $P \sim \{p_k\}$, $Q \sim \{q_k\}$ et $R \sim \{\lambda p_k + (1-\lambda) q_k\}, \lambda \in [0\,;\,1]$

  $$H(R) \geq \lambda H(P) + (1-\lambda) H(Q)$$



- $H(aX) = H(X+c) = H(X)$: indépendant des valeurs prises par X

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 = 3$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) = 3$ bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
- Code optimum de taille moyenne L^* : $H(X) \le L^* < H(X) + 1$

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 = 3$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) = 3$ bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
- Code optimum de taille moyenne L^* : $H(X) \leq L^* < H(X) + 1$

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 = 3$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) = 3$ bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
- Code optimum de taille moyenne L^* : $H(X) \le L^* < H(X) + 1$

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 =$ 3 bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) =$ 3 bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
  - Comme avant : codage sur 3 bits, & 3 bits en moyenne
  - “0” forte proba. d'apparition $\leadsto$ codage sur peu de bits
    “7” faible proba. d'apparition $\leadsto$ codage sur plus de bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = 2$ bits
- Code optimum de taille moyenne L^* : $H(X) \le L^* < H(X) + 1$

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 =$ 3 bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) =$ 3 bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
  - Comme avant : codage sur 3 bits, & 3 bits en moyenne
  - "0" forte proba. d'apparition $\rightsquigarrow$ codage sur peu de bits
    "7" faible proba. d'apparition $\rightsquigarrow$ codage sur plus de bits
    - séquences décodable…
    - on a besoin de…6 bits !
    - mais en moyenne $\frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + \frac{4}{64}\times 6 = 2$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = 2$ bits
- Code optimum de taille moyenne L^* : $H(X) \le L^* < H(X) + 1$

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 = 3$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) = 3$ bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
  - Comme avant : codage sur 3 bits, & 3 bits en moyenne
  - "0" forte proba. d'apparition $\rightsquigarrow$ codage sur peu de bits
    "7" faible proba. d'apparition $\rightsquigarrow$ codage sur plus de bits
    choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
    - séquences décodable...
    - on a besoin de... 6 bits !
    - mais en moyenne $\frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + \frac{4}{64}\times 6 = 2$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = 2$ bits
- Code optimum de taille moyenne L^* : $H(X) \le L^* < H(X) + 1$

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 = 3$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) = 3$ bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
  - Comme avant : codage sur 3 bits, & 3 bits en moyenne
  - "0" forte proba. d'apparition $\leadsto$ codage sur peu de bits
    "7" faible proba. d'apparition $\leadsto$ codage sur plus de bits

    choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
    - séquences décodable...
    - on a besoin de... 6 bits !
    - mais en moyenne $\frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + \frac{4}{64}\times 6 = 2$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = 2$ bits
- Code optimum de taille moyenne L^* : $H(X) \le L^* < H(X) + 1$

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 = 3$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) = 3$ bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
  - Comme avant : codage sur 3 bits, & 3 bits en moyenne
  - "0" forte proba. d'apparition $\rightsquigarrow$ codage sur peu de bits
    "7" faible proba. d'apparition $\rightsquigarrow$ codage sur plus de bits

    choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
    - séquences décodable…
    - on a besoin de… 6 bits !
    - mais en moyenne $\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + \frac{4}{64} \times 6 = 2$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = 2$ bits
- Code optimum de taille moyenne L^* : $H(X) \le L^* < H(X) + 1$

Un problème de codage

- Soit un dé à 8 faces équiprobables, codage binaire de chaque face :
  - Codage sur 3 bits : $(000, 001, 010, 011, 100, 101, 110, 111)$
  - En moyenne on a besoin de $\sum \frac{1}{8} \times 3 = 3$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = -\sum \frac{1}{8} \log_2\left(\frac{1}{8}\right) = 3$ bits
- Supposons cette fois le dé pipé, $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:
  - Comme avant : codage sur 3 bits, & 3 bits en moyenne
  - "0" forte proba. d'apparition $\rightsquigarrow$ codage sur peu de bits
    "7" faible proba. d'apparition $\rightsquigarrow$ codage sur plus de bits
    choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
    - séquences décodable...
    - on a besoin de... 6 bits !
    - mais en moyenne $\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + \frac{4}{64} \times 6 = 2$ bits
  - Entropie $H(X) = -\sum p_k \log_2(p_k) = 2$ bits
- Code optimum de taille moyenne L^* : $H(X) \le L^* < H(X) + 1$

Application à un problème de pile ou face

Une pièce a une probabilité p de tomber sur le coté "pile".
Soit X le nombre de lancers successifs nécessaires pour tomber sur "face".

- Déterminer $p_k = \Pr[X = k]$ puis l'entropie $H(X) = f(p)$ de X.

- Quelle séquence "naive" de questions permettrait de deviner x ?
- Quel est le nombre moyen N de questions à poser pour trouver x ?
- Comparer N et $H(X)$ dans les cas $p = \frac{1}{2}$ et $p = \frac{\sqrt{2}}{2}$.
- Peut-on trouver une séquence de questions plus courte (en moyenne)?
- Si oui, quel est le nombre moyen N^* de questions? Conclure.

Application à un problème de pile ou face

Une pièce a une probabilité p de tomber sur le coté "pile".
Soit X le nombre de lancers successifs nécessaires pour tomber sur "face".

- Déterminer $p_k = \Pr[X = k]$ puis l'entropie $H(X) = f(p)$ de X.

Vali vient de faire x lancers pour tomber enfin sur "face". On veut deviner x en lui posant une suite de questions binaires "$x \in \{x_1, \ldots x_k\}$?"

- Quelle séquence "naive" de questions permettrait de deviner x ?
- Quel est le nombre moyen N de questions à poser pour trouver x ?
- Comparer N et $H(X)$ dans les cas $p = \frac{1}{2}$ et $p = \frac{\sqrt{2}}{2}$.
- Peut-on trouver une séquence de questions plus courte (en moyenne)?
- Si oui, quel est le nombre moyen N^* de questions? Conclure.

Application à un problème de pile ou face

Une pièce a une probabilité p de tomber sur le coté "pile".
Soit X le nombre de lancers successifs nécessaires pour tomber sur "face".

- Déterminer $p_k = \Pr[X = k]$ puis l'entropie $H(X) = f(p)$ de X.

Vali vient de faire x lancers pour tomber enfin sur "face". On veut deviner x en lui posant une suite de questions binaires "$x \in \{x_1, \ldots x_k\}$?"

- Quelle séquence "naive" de questions permettrait de deviner x ?
- Quel est le nombre moyen N de questions à poser pour trouver x ?
- Comparer N et $H(X)$ dans les cas $p = \frac{1}{2}$ et $p = \frac{\sqrt{2}}{2}$.
- Peut-on trouver une séquence de questions plus courte (en moyenne)?
- Si oui, quel est le nombre moyen N^* de questions? Conclure.

Extension au cas continu

Pour X variable aléatoire continue, par analogie :

$$H(X) = -\int p(x) \log p(x) \, dx \qquad \text{(entropie différentielle)}$$

Pas d'approche axiomatique (à ma connaissance)

Extension au cas continu

Pour X variable aléatoire continue, par analogie :

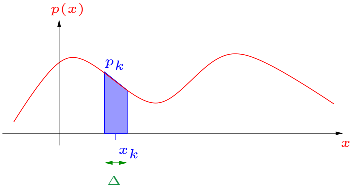
$$H(X) = -\int p(x) \log p(x)\, dx \qquad \text{(entropie différentielle)}$$

Pas d'approche axiomatique (à ma connaissance)

Lien avec l'entropie discrète :

Soit x_k tel que $p_k = p(x_k)\Delta = \int_{k\Delta}^{(k+1)\Delta} p(x)\, dx$

$X^\Delta = x_k$ sur $[k\Delta\,;\,(k+1)\Delta)$, $\quad X^\Delta \sim \{p_k\}$



Extension au cas continu

Pour X variable aléatoire continue, par analogie :

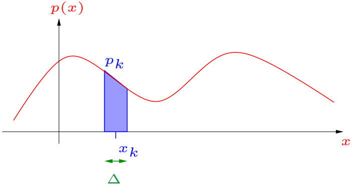
$$H(X) = -\int p(x) \log p(x)\, dx \qquad \text{(entropie différentielle)}$$

Pas d'approche axiomatique (à ma connaissance)

Lien avec l'entropie discrète :

Soit x_k tel que $p_k = p(x_k)\Delta = \int_{k\Delta}^{(k+1)\Delta} p(x)\, dx$

$X^\Delta = x_k$ sur $[k\Delta\,;\,(k+1)\Delta)$, $\quad X^\Delta \sim \{p_k\}$



$$\begin{aligned} H\left(X^\Delta\right) &= -\sum p_k \log p_k = -\sum p(x_k)\Delta \log(p(x_k)\Delta) \\ &= -\log\Delta - \sum p(x_k) \log p(x_k)\, \Delta \end{aligned}$$

Extension au cas continu

Pour X variable aléatoire continue, par analogie :

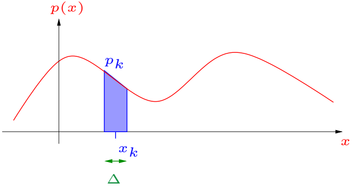
$$H(X) = -\int p(x) \log p(x) \, dx \qquad \text{(entropie différentielle)}$$

Pas d'approche axiomatique (à ma connaissance)

Lien avec l'entropie discrète :

Soit x_k tel que $p_k = p(x_k)\Delta = \int_{k\Delta}^{(k+1)\Delta} p(x) \, dx$

$X^\Delta = x_k$ sur $[k\Delta \,;\, (k+1)\Delta)$, $\quad X^\Delta \sim \{p_k\}$



$$\begin{aligned} H\left(X^\Delta\right) &= -\sum p_k \log p_k = -\sum p(x_k)\Delta \log(p(x_k)\Delta) \\ &= -\log \Delta - \sum p(x_k) \log p(x_k) \, \Delta \end{aligned}$$

$$\lim_{\Delta \to 0} H\left(X^\Delta\right) + \log \Delta = H(X)$$

Quelques propriétés du cas continu

- X définie sur $[a\,;\,b]$ borné, maximum $H = \log(b-a) \Leftrightarrow X$ uniforme
  H peut être négative et $\to -\infty$ pour $b \to a$ (pas d'incertitude)
- Concavité : $H(\lambda p + (1-\lambda) q) \geq \lambda H(p) + (1-\lambda) H(q)$
- Support $\mathbb{R}$, variance/puissance fixée σ^2,
  maximum $H(X) = \frac{1}{2}\log(2\pi e) + \log\sigma \Leftrightarrow p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$
  en Gaussien, σ^2 contient toute l'information/incertitude sur X
- $H(X+c) = H(X)$, mais $H(aX) = H(X) + \log(|a|)$:
  dépend des "valeurs" prises par X
- Puissance entropique $N(X) = \frac{1}{2\pi e}\exp\left(2H(X)\right)$

Quelques propriétés du cas continu

- X définie sur $[a\,;\,b]$ borné, maximum $H = \log(b-a) \Leftrightarrow X$ uniforme
  H peut être négative et $\to -\infty$ pour $b \to a$ (pas d'incertitude)
- Concavité : $H(\lambda p + (1-\lambda) q) \geq \lambda H(p) + (1-\lambda) H(q)$
- Support $\mathbb{R}$, variance/puissance fixée σ^2,
  maximum $H(X) = \frac{1}{2}\log(2\pi e) + \log\sigma \Leftrightarrow p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$
  en Gaussien, σ^2 contient toute l'information/incertitude sur X
- $H(X+c) = H(X)$, mais $H(aX) = H(X) + \log(|a|)$:
  dépend des "valeurs" prises par X
- Puissance entropique $N(X) = \frac{1}{2\pi e}\exp\left(2H(X)\right)$

Quelques propriétés du cas continu

- X définie sur $[a\,;\,b]$ borné, maximum $H = \log(b-a) \Leftrightarrow X$ uniforme
  H peut être négative et $\to -\infty$ pour $b \to a$ (pas d'incertitude)
- Concavité : $H(\lambda p + (1-\lambda) q) \ge \lambda H(p) + (1-\lambda) H(q)$
- Support $\mathbb{R}$, variance/puissance fixée σ^2,
  maximum $H(X) = \frac{1}{2}\log(2\pi \mathrm{e}) + \log\sigma \Leftrightarrow p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$
  en Gaussien, σ^2 contient toute l'information/incertitude sur X
- $H(X+c) = H(X)$, mais $H(aX) = H(X) + \log(|a|)$:
  dépend des "valeurs" prises par X
- Puissance entropique $N(X) = \frac{1}{2\pi\mathrm{e}}\exp\left(2H(X)\right)$

Quelques propriétés du cas continu

- X définie sur $[a\,;\,b]$ borné, maximum $H = \log(b-a) \Leftrightarrow X$ uniforme
  H peut être négative et $\to -\infty$ pour $b \to a$ (pas d'incertitude)
- Concavité : $H(\lambda p + (1-\lambda) q) \ge \lambda H(p) + (1-\lambda) H(q)$
- Support $\mathbb{R}$, variance/puissance fixée σ^2,
  maximum $H(X) = \frac{1}{2}\log(2\pi \mathrm{e}) + \log\sigma \Leftrightarrow p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$
  en Gaussien, σ^2 contient toute l'information/incertitude sur X
- $H(X+c) = H(X)$, mais $H(aX) = H(X) + \log(|a|)$:
  dépend des "valeurs" prises par X
- Puissance entropique $N(X) = \frac{1}{2\pi \mathrm{e}}\exp\left(2H(X)\right)$

Quelques propriétés du cas continu

- X définie sur $[a\,;\,b]$ borné, maximum $H = \log(b-a) \Leftrightarrow X$ uniforme
  H peut être négative et $\to -\infty$ pour $b \to a$ (pas d'incertitude)
- Concavité : $H(\lambda p + (1-\lambda) q) \geq \lambda H(p) + (1-\lambda) H(q)$
- Support $\mathbb{R}$, variance/puissance fixée σ^2,
  maximum $H(X) = \frac{1}{2}\log(2\pi \mathrm{e}) + \log\sigma \Leftrightarrow p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$
  en Gaussien, σ^2 contient toute l'information/incertitude sur X
- $H(X+c) = H(X)$, mais $H(aX) = H(X) + \log(|a|)$:
  dépend des “valeurs” prises par X
- Puissance entropique $N(X) = \frac{1}{2\pi\mathrm{e}}\exp\left(2H(X)\right)$
  - $N(X) \geq 0$
  - Uniforme $N(X) = \frac{(b-a)^2}{2\pi\mathrm{e}} \to 0$ si $b-a \to 0$
  - Gaussien $N(X) = \sigma^2 \to 0$ si $\sigma^2 \to 0$

Entropie d'un vecteur aléatoire

Vecteur aléatoire $\boldsymbol{X} \sim p(\boldsymbol{x})$ sur un domaine $\mathcal{D}$ de $\mathbb{R}^d$ ou $\mathbb{Z}^d$, entropie de $\boldsymbol{X}$

$$H(\boldsymbol{X}) = -\int_{\mathcal{D}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \quad \text{ou} \quad -\sum_{\mathcal{D}} p_{k_1,\ldots,k_d} \log p_{k_1,\ldots,k_d}$$

aussi appelée entropie conjointe des X_i et notée $H(X_1, \ldots, X_d)$

- $\boldsymbol{X}$ défini sur un support $\mathcal{D}$ de $\mathbb{R}^d$ de volume fini vol $\mathcal{D}$

  $H_{\max}(\boldsymbol{X}) = \log \operatorname{vol} \mathcal{D} \;\Leftrightarrow\; \boldsymbol{X}$ uniforme sur $\mathcal{D}$

- support $\mathbb{R}^d$, matrice de covariance $\boldsymbol{R}$ donnée

  $H_{\max}(\boldsymbol{X}) = \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log |\boldsymbol{R}| \Leftrightarrow \boldsymbol{X} \sim \mathcal{N}(\cdot, \boldsymbol{R})$

- Puissance fixée $P = \operatorname{Trace}(\boldsymbol{R})$, l'entropie décroit avec la corrélation
- Puissance entropique $N(\boldsymbol{X}) = \frac{1}{2\pi e} \exp\left(\frac{2}{d} H(\boldsymbol{X})\right)$

Entropie d'un vecteur aléatoire

Vecteur aléatoire $\boldsymbol{X} \sim p(\boldsymbol{x})$ sur un domaine $\mathcal{D}$ de $\mathbb{R}^d$ ou $\mathbb{Z}^d$, entropie de $\boldsymbol{X}$

$$H(\boldsymbol{X}) = -\int_{\mathcal{D}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \quad \text{ou} \quad -\sum_{\mathcal{D}} p_{k_1,\ldots,k_d} \log p_{k_1,\ldots,k_d}$$

aussi appelée entropie conjointe des X_i et notée $H(X_1, \ldots, X_d)$

- $\boldsymbol{X}$ défini sur un support $\mathcal{D}$ de $\mathbb{R}^d$ de volume fini $\operatorname{vol}\mathcal{D}$

  $H_{\max}(\boldsymbol{X}) = \log \operatorname{vol} \mathcal{D} \;\Leftrightarrow\; \boldsymbol{X}$ uniforme sur $\mathcal{D}$

- support $\mathbb{R}^d$, matrice de covariance $\boldsymbol{R}$ donnée

  $H_{\max}(\boldsymbol{X}) = \frac{d}{2} \log(2\pi \mathrm{e}) + \frac{1}{2} \log |\boldsymbol{R}| \Leftrightarrow \boldsymbol{X} \sim \mathcal{N}(\cdot, \boldsymbol{R})$

- Puissance fixée $P = \operatorname{Trace}(\boldsymbol{R})$, l'entropie décroit avec la corrélation

- Puissance entropique $N(\boldsymbol{X}) = \frac{1}{2\pi e} \exp\left(\frac{2}{d} H(\boldsymbol{X})\right)$

Entropie d'un vecteur aléatoire

Vecteur aléatoire $\boldsymbol{X} \sim p(\boldsymbol{x})$ sur un domaine $\mathcal{D}$ de $\mathbb{R}^d$ ou $\mathbb{Z}^d$, entropie de $\boldsymbol{X}$

$$H(\boldsymbol{X}) = -\int_{\mathcal{D}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \quad \text{ou} \quad -\sum_{\mathcal{D}} p_{k_1,\ldots,k_d} \log p_{k_1,\ldots,k_d}$$

aussi appelée entropie conjointe des X_i et notée $H(X_1, \ldots, X_d)$

- $\boldsymbol{X}$ défini sur un support $\mathcal{D}$ de $\mathbb{R}^d$ de volume fini $\operatorname{vol}\mathcal{D}$

  $H_{\max}(\boldsymbol{X}) = \log \operatorname{vol}\mathcal{D} \Leftrightarrow \boldsymbol{X}$ uniforme sur $\mathcal{D}$

- support $\mathbb{R}^d$, matrice de covariance $\boldsymbol{R}$ donnée

  $H_{\max}(\boldsymbol{X}) = \frac{d}{2}\log(2\pi \mathrm{e}) + \frac{1}{2}\log|\boldsymbol{R}| \Leftrightarrow \boldsymbol{X} \sim \mathcal{N}(\cdot, \boldsymbol{R})$

- Puissance fixée $P = \operatorname{Trace}(\boldsymbol{R})$, l'entropie décroit avec la corrélation

- Puissance entropique $N(\boldsymbol{X}) = \frac{1}{2\pi e}\exp\left(\frac{2}{d}H(\boldsymbol{X})\right)$

Entropie d'un vecteur aléatoire

Vecteur aléatoire $\boldsymbol{X} \sim p(\boldsymbol{x})$ sur un domaine $\mathcal{D}$ de $\mathbb{R}^d$ ou $\mathbb{Z}^d$, entropie de $\boldsymbol{X}$

$$H(\boldsymbol{X}) = -\int_{\mathcal{D}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \quad \text{ou} \quad -\sum_{\mathcal{D}} p_{k_1,\ldots,k_d} \log p_{k_1,\ldots,k_d}$$

aussi appelée entropie conjointe des X_i et notée $H(X_1, \ldots, X_d)$

- $\boldsymbol{X}$ défini sur un support $\mathcal{D}$ de $\mathbb{R}^d$ de volume fini $\operatorname{vol} \mathcal{D}$

  $H_{\max}(\boldsymbol{X}) = \log \operatorname{vol} \mathcal{D} \Leftrightarrow \boldsymbol{X}$ uniforme sur $\mathcal{D}$
- support $\mathbb{R}^d$, matrice de covariance $\boldsymbol{R}$ donnée

  $H_{\max}(\boldsymbol{X}) = \frac{d}{2} \log(2\pi \mathrm{e}) + \frac{1}{2} \log |\boldsymbol{R}| \Leftrightarrow \boldsymbol{X} \sim \mathcal{N}(\cdot, \boldsymbol{R})$
- Puissance fixée $P = \operatorname{Trace}(\boldsymbol{R})$, l'entropie décroit avec la corrélation
- Puissance entropique $N(\boldsymbol{X}) = \frac{1}{2\pi \mathrm{e}} \exp\left(\frac{2}{d} H(\boldsymbol{X})\right)$

Taux d'entropie : $\mathcal{H}(\boldsymbol{X}) = \frac{1}{d} H(\boldsymbol{X})$; iid $\mathcal{H}(\mathcal{X}) = H(X_1)$

D'autres mesure d'incertitude

- (Co)variance : l'incertitude est nulle si $\sigma^2 = 0$ et augmente avec σ^2
- Matrice/information de Fisher : $\displaystyle\int \frac{(\nabla p)(\nabla p)^t}{p}$ et sa trace
- Entropie de Rényi : $H_\lambda = \frac{1}{1-\lambda} \log \int p^\lambda$ (phys. stat., détec.)
- Entropie de Havrda et Charvat'67 : $T_q = \frac{1}{q-1} \left(\int p^q - 1 \right)$ (phys.)
- Entropies d'ordre (r,s) (Varma'66) : $H_{s,r} = \frac{1}{s-r} \log \frac{\int p^r}{\int p^s}$
- Entropies de degré (r,s) (Kapur'67) : $T_{s,r} = \frac{1}{s-r} \int (p^r - p^s)$
- f-entropies (Arimoto'71) : $H_f = \inf_{\tilde{p}} \int p f(\tilde{p})$

  f convexe, continuement dérivable ; $A_t = \frac{1}{t-1} \left(\left(\int p^{\frac{1}{t}} \right)^t - 1 \right)$

D'autres mesure d'incertitude

- (Co)variance : l'incertitude est nulle si $\sigma^2 = 0$ et augmente avec σ^2
- Matrice/information de Fisher : $\displaystyle\int \frac{(\boldsymbol{\nabla} p)\,(\boldsymbol{\nabla} p)^t}{p}$ et sa trace
- Entropie de Rényi : $H_\lambda = \frac{1}{1-\lambda} \log \int p^\lambda$ (phys. stat., détec.)
- Entropie de Havrda et Charvat'67 : $T_q = \frac{1}{q-1}\left(\int p^q - 1\right)$ (phys.)
- Entropies d'ordre (r,s) (Varma'66) : $H_{s,r} = \frac{1}{s-r} \log \frac{\int p^r}{\int p^s}$
- Entropies de degré (r,s) (Kapur'67) : $T_{s,r} = \frac{1}{s-r} \int (p^r - p^s)$
- f-entropies (Arimoto'71) : $H_f = \inf_{\tilde{p}} \int p f(\tilde{p})$

  f convexe, continuement dérivable ; $A_t = \frac{1}{t-1}\left(\left(\int p^{\frac{1}{t}}\right)^t - 1\right)$

D'autres mesure d'incertitude

- (Co)variance : l'incertitude est nulle si $\sigma^2 = 0$ et augmente avec σ^2
- Matrice/information de Fisher : $\displaystyle\int \frac{(\boldsymbol{\nabla} p)\,(\boldsymbol{\nabla} p)^t}{p}$ et sa trace
- Entropie de Rényi : $H_\lambda = \frac{1}{1-\lambda} \log \int p^\lambda$ (phys. stat., détec.)

- Entropie de Havrda et Charvat'67 : $T_q = \frac{1}{q-1} \left(\int p^q - 1 \right)$ (phys.)
- Entropies d'ordre (r,s) (Varma'66) : $H_{s,r} = \frac{1}{s-r} \log \frac{\int p^r}{\int p^s}$
- Entropies de degré (r,s) (Kapur'67) : $T_{s,r} = \frac{1}{s-r} \int (p^r - p^s)$
- f-entropies (Arimoto'71) : $H_f = \inf_{\tilde{p}} \int p f(\tilde{p})$

  f convexe, continuement dérivable ; $A_t = \frac{1}{t-1} \left(\left(\int p^{\frac{1}{t}} \right)^t - 1 \right)$

Entropie de Rényi, 1961

Généralisation de Shannon :

- Continuité ; information élémentaire $h(p_k)$
- Indépendance $H_r(X,Y) = H_r(X) + H_r(Y)$

  $h(p_k q_l) = h(p_k) + h(q_l)$ implique $h(p) = -\log(p)$
- Moyenne quasi-linéaire (exponentielle) :

  $H_r(X) = \varphi^{-1}\left(\sum p_k \varphi(h(p_k))\right)$ avec φ continue croissante (KN)
- $\varphi(x) = ax + b$, conduisant à Shannon
- $\varphi(x) = a\,\mathrm{e}^{(1-\lambda)x} + b$ avec $\lambda \geq 0, \neq 1$ conduisant à Rényi :

$$H_\lambda(X) = \frac{1}{1-\lambda}\log\sum_k p_k^\lambda \qquad \frac{1}{1-\lambda}\log\int p^\lambda(x)\,dx$$

Entropie de Rényi, 1961

Généralisation de Shannon :

- Continuité ; information élémentaire $h(p_k)$
- Indépendance $H_r(X,Y) = H_r(X) + H_r(Y)$

  $h(p_k q_l) = h(p_k) + h(q_l)$ implique $h(p) = -\log(p)$
- Moyenne quasi-linéaire (exponentielle) :

  $H_r(X) = \varphi^{-1}\left(\sum p_k \varphi(h(p_k))\right)$ avec φ continue croissante (KN)

- $\varphi(x) = ax + b$, conduisant à Shannon
- $\varphi(x) = a\,\mathrm{e}^{(1-\lambda)x} + b$ avec $\lambda \geq 0, \neq 1$ conduisant à Rényi :

$$H_\lambda(X) = \frac{1}{1-\lambda}\log\sum_k p_k^\lambda \qquad \frac{1}{1-\lambda}\log\int p^\lambda(x)\,dx$$

Entropie de Rényi, 1961

Généralisation de Shannon :

- Continuité ; information élémentaire $h(p_k)$
- Indépendance $H_r(X,Y) = H_r(X) + H_r(Y)$

  $h(p_k q_l) = h(p_k) + h(q_l)$ implique $h(p) = -\log(p)$
- Moyenne quasi-linéaire (exponentielle) :

  $H_r(X) = \varphi^{-1}\left(\sum p_k \varphi(h(p_k))\right)$ avec φ continue croissante (KN)

Seules solutions admissibles :

- $\varphi(x) = ax + b$, conduisant à Shannon
- $\varphi(x) = a\,\mathrm{e}^{(1-\lambda)x} + b$ avec $\lambda \geq 0, \neq 1$ conduisant à Rényi :

$$H_\lambda(X) = \frac{1}{1-\lambda} \log \sum_k p_k^\lambda \qquad \frac{1}{1-\lambda} \log \int p^\lambda(x)\, dx$$

Entropie de Rényi, 1961

Généralisation de Shannon :

- Continuité ; information élémentaire $h(p_k)$
- Indépendance $H_r(X,Y) = H_r(X) + H_r(Y)$

  $h(p_k q_l) = h(p_k) + h(q_l)$ implique $h(p) = -\log(p)$
- Moyenne quasi-linéaire (exponentielle) :

  $H_r(X) = \varphi^{-1}\left(\sum p_k \varphi(h(p_k))\right)$ avec φ continue croissante (KN)

Seules solutions admissibles :

- $\varphi(x) = ax + b$, conduisant à Shannon
- $\varphi(x) = a\,\mathrm{e}^{(1-\lambda)x} + b$ avec $\lambda \geq 0, \neq 1$ conduisant à Rényi :

$$H_\lambda(\boldsymbol{X}) = \frac{1}{1-\lambda} \log \sum_{\boldsymbol{k}} p_{\boldsymbol{k}}^\lambda \qquad \frac{1}{1-\lambda} \log \int p^\lambda(\boldsymbol{x})\, d\boldsymbol{x}$$

Entropie de Rényi, 1961

Généralisation de Shannon :

- Continuité ; information élémentaire $h(p_k)$
- Indépendance $H_r(X,Y) = H_r(X) + H_r(Y)$

  $h(p_k q_l) = h(p_k) + h(q_l)$ implique $h(p) = -\log(p)$
- Moyenne quasi-linéaire (exponentielle) :

  $H_r(X) = \varphi^{-1}\left(\sum p_k \varphi(h(p_k))\right)$ avec φ continue croissante (KN)

Seules solutions admissibles :

- $\varphi(x) = ax + b$, conduisant à Shannon
- $\varphi(x) = a\,\mathrm{e}^{(1-\lambda)x} + b$ avec $\lambda \geq 0, \neq 1$ conduisant à Rényi :

$$H_\lambda(\boldsymbol{X}) = \frac{1}{1-\lambda} \log \sum_{\boldsymbol{k}} p_{\boldsymbol{k}}^\lambda \qquad \frac{1}{1-\lambda} \log \int p^\lambda(\boldsymbol{x})\, d\boldsymbol{x}$$

$\lambda \to 1$ donne $H_\lambda \to H$ $\qquad$ λ joue un rôle de “loupe”

Quelques propriétés des entropies de Rényi

- Concavité pour $\lambda \le 1$
- H_λ décroît avec λ de $H_0 = \log \operatorname{vol} \mathcal{D}$ à $H_\infty = -\log \sup p$
- $p_{\boldsymbol{k}} = \delta_{\boldsymbol{k},\boldsymbol{l}} \Leftrightarrow H_\lambda = 0$ minimum
- $\boldsymbol{X} \in \prod_i \{1, \ldots, n_i\}$ et $p_{\boldsymbol{k}} = \frac{1}{\prod_i n_i} \Leftrightarrow H_\lambda = \log(\prod n_i)$ maximum
- Discret $H_\lambda \ge 0$ – continu $N_\lambda \propto \exp\left(\frac{2\,H_\lambda}{d}\right)$
- Continu, support fini, maximum en uniforme $H_\lambda = \log(\operatorname{vol} \mathcal{D})$
- Contrainte de (co)variance, maximisantes :

$$\lambda > 1 : \ f(\boldsymbol{x}) \propto \left(1 - \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)_+^{\frac{1}{\lambda-1}} \qquad \text{Student-r} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

$$1 - \frac{2}{d} < \lambda < 1 : \ f(\boldsymbol{x}) \propto \left(1 + \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)^{\frac{1}{\lambda-1}} \qquad \text{Student-t} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

Quelques propriétés des entropies de Rényi

- Concavité pour $\lambda \le 1$
- H_λ décroît avec λ de $H_0 = \log \operatorname{vol} \mathcal{D}$ à $H_\infty = -\log \sup p$
- $p_{\boldsymbol{k}} = \delta_{\boldsymbol{k},\boldsymbol{l}} \Leftrightarrow H_\lambda = 0$ minimum
- $\boldsymbol{X} \in \prod_i \{1, \ldots, n_i\}$ et $p_{\boldsymbol{k}} = \frac{1}{\prod_i n_i} \Leftrightarrow H_\lambda = \log(\prod n_i)$ maximum
- Discret $H_\lambda \ge 0$ – continu $N_\lambda \propto \exp\left(\frac{2\,H_\lambda}{d}\right)$
- Continu, support fini, maximum en uniforme $H_\lambda = \log(\operatorname{vol} \mathcal{D})$
- Contrainte de (co)variance, maximisantes :

$$\lambda > 1 : \; f(\boldsymbol{x}) \propto \left(1 - \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)_+^{\frac{1}{\lambda-1}} \qquad \text{Student-r} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

$$1 - \frac{2}{d} < \lambda < 1 : \; f(\boldsymbol{x}) \propto \left(1 + \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)^{\frac{1}{\lambda-1}} \qquad \text{Student-t} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

Quelques propriétés des entropies de Rényi

- Concavité pour $\lambda \le 1$
- H_λ décroît avec λ de $H_0 = \log \operatorname{vol} \mathcal{D}$ à $H_\infty = -\log \sup p$
- $p_{\boldsymbol{k}} = \delta_{\boldsymbol{k},\boldsymbol{l}} \Leftrightarrow H_\lambda = 0$ minimum
- $\boldsymbol{X} \in \prod_i \{1, \ldots, n_i\}$ et $p_{\boldsymbol{k}} = \frac{1}{\prod_i n_i} \Leftrightarrow H_\lambda = \log(\prod n_i)$ maximum
- Discret $H_\lambda \ge 0$ – continu $N_\lambda \propto \exp\left(\frac{2\,H_\lambda}{d}\right)$
- Continu, support fini, maximum en uniforme $H_\lambda = \log(\operatorname{vol} \mathcal{D})$
- Contrainte de (co)variance, maximisantes :

$$\lambda > 1 : \; f(\boldsymbol{x}) \propto \left(1 - \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)_+^{\frac{1}{\lambda-1}} \qquad \text{Student-r} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

$$1 - \frac{2}{d} < \lambda < 1 : \; f(x) \propto \left(1 + \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)^{\frac{1}{\lambda-1}} \qquad \text{Student-t} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

Quelques propriétés des entropies de Rényi

- Concavité pour $\lambda \leq 1$
- H_λ décroît avec λ de $H_0 = \log \operatorname{vol} \mathcal{D}$ à $H_\infty = -\log \sup p$
- $p_{\boldsymbol{k}} = \delta_{\boldsymbol{k},\boldsymbol{l}} \Leftrightarrow H_\lambda = 0$ minimum
- $\boldsymbol{X} \in \prod_i \{1, \ldots, n_i\}$ et $p_{\boldsymbol{k}} = \frac{1}{\prod_i n_i} \Leftrightarrow H_\lambda = \log(\prod n_i)$ maximum
- Discret $H_\lambda \geq 0$ – continu $N_\lambda \propto \exp\left(\frac{2\, H_\lambda}{d}\right)$
- Continu, support fini, maximum en uniforme $H_\lambda = \log(\operatorname{vol} \mathcal{D})$
- Contrainte de (co)variance, maximisantes :

$$\lambda > 1 : \; f(\boldsymbol{x}) \propto \left(1 - \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)_+^{\frac{1}{\lambda - 1}} \qquad \text{Student-r} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

$$1 - \tfrac{2}{d} < \lambda < 1 : \; f(\boldsymbol{x}) \propto \left(1 + \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)^{\frac{1}{\lambda - 1}} \qquad \text{Student-t} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

Quelques propriétés des entropies de Rényi

- Concavité pour $\lambda \le 1$
- H_λ décroît avec λ de $H_0 = \log \operatorname{vol} \mathcal{D}$ à $H_\infty = -\log \sup p$
- $p_{\boldsymbol{k}} = \delta_{\boldsymbol{k},\boldsymbol{l}} \Leftrightarrow H_\lambda = 0$ minimum
- $\boldsymbol{X} \in \prod_i \{1, \ldots, n_i\}$ et $p_{\boldsymbol{k}} = \frac{1}{\prod_i n_i} \Leftrightarrow H_\lambda = \log(\prod n_i)$ maximum
- Discret $H_\lambda \ge 0$ – continu $N_\lambda \propto \exp\left(\frac{2\,H_\lambda}{d}\right)$
- Continu, support fini, maximum en uniforme $H_\lambda = \log(\operatorname{vol} \mathcal{D})$
- Contrainte de (co)variance, maximisantes :

$$\lambda > 1 \, : \; f(\boldsymbol{x}) \propto \left(1 - \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)_+^{\frac{1}{\lambda-1}} \qquad \text{Student-r} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

$$1 - \tfrac{2}{d} < \lambda < 1 \, : \; f(x) \propto \left(1 + \boldsymbol{x}^t \boldsymbol{R}^{-1} \boldsymbol{x}\right)^{\frac{1}{\lambda-1}} \qquad \text{Student-t} \quad \xrightarrow{\lambda \to 1} \quad \mathcal{N}$$

Rappel : λ joue un rôle de loupe

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale

- Pénalisation des mots longs : minimisation de $C_\alpha = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\to$ Kraft $\sum_i 2^{-l_i} \le 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $l_i = -\log_2 p_i$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \ge H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $l_i = -\log_2 p_i$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\alpha = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $l_i = -\log_2 p_i$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\to$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $l_i = -\log_2 p_i$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$
  - $C_0 = L$ et $C_\infty = \max_i l_i$
  - $l_i = -\log_2 \widetilde{p}_i$ et $C_\alpha = H_\lambda(X)$, avec $\lambda = \frac{1}{1+\alpha}$ et $\widetilde{p}_i = \frac{p_i^\lambda}{\sum_k p_k^\lambda}$
  - Pénalisation $M_\lambda = \sum_i \widetilde{p}_i l_i$ et code / distrib. compagne $\{\widetilde{p}_i\}_i$
  - $\alpha = 1$, $\left(\frac{10\sqrt{2}-12}{7}, \frac{10-6\sqrt{2}}{7}, \frac{5\sqrt{2}-6}{7}, \frac{5-3\sqrt{2}}{7}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}\right)$
  - $L = 2.31 > L^*$, mais besoin de 4 bits
  - $H_{.5}(X) = 2.42$ et $C_1 = 2.46$; $H(\widetilde{X}) = 1.85$ et $M_{.5} = 2.69$.

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $l_i = -\log_2 p_i$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$
  - $C_0 = L$ et $C_\infty = \max_i l_i$
  - $l_i = -\log_2 \widetilde{p}_i$ et $C_\alpha = H_\lambda(X)$, avec $\lambda = \frac{1}{1+\alpha}$ et $\widetilde{p}_i = \frac{p_i^\lambda}{\sum_k p_k^\lambda}$
  - Pénalisation $M_\lambda = \sum_i \widetilde{p}_i l_i$ et code / distrib. compagne $\{\widetilde{p}_i\}_i$
  - $\alpha = 1$, $\left(\frac{10\sqrt{2}-12}{7}, \frac{10-6\sqrt{2}}{7}, \frac{5\sqrt{2}-6}{7}, \frac{5-3\sqrt{2}}{7}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}\right)$
  - $L = 2.31 > L^*$, mais besoin de 4 bits
  - $H_{.5}(X) = 2.42$ et $C_1 = 2.46$; $H(\widetilde{X}) = 1.85$ et $M_{.5} = 2.69$.

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $\underline{l_i = -\log_2 p_i}$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$
  - $C_0 = L$ et $C_\infty = \max_i l_i$
  - $\underline{l_i = -\log_2 \widetilde{p}_i}$ et $C_\alpha = H_\lambda(X)$, avec $\lambda = \frac{1}{1+\alpha}$ et $\widetilde{p}_i = \frac{p_i^\lambda}{\sum_k p_k^\lambda}$
  - Pénalisation $M_\lambda = \sum_i \widetilde{p}_i l_i$ et code / distrib. compagne $\{\widetilde{p}_i\}_i$
  - $\alpha = 1$, $\left(\frac{10\sqrt{2}-12}{7}, \frac{10-6\sqrt{2}}{7}, \frac{5\sqrt{2}-6}{7}, \frac{5-3\sqrt{2}}{7}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}\right)$
  - $L = 2.31 > L^*$, mais besoin de 4 bits
  - $H_{.5}(X) = 2.42$ et $C_1 = 2.46$; $H(\widetilde{X}) = 1.85$ et $M_{.5} = 2.69$.

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $\underline{l_i = -\log_2 p_i}$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$
  - $C_0 = L$ et $C_\infty = \max_i l_i$
  - $\underline{l_i = -\log_2 \widetilde{p}_i}$ et $C_\alpha = H_\lambda(X)$, avec $\lambda = \frac{1}{1+\alpha}$ et $\widetilde{p}_i = \frac{p_i^\lambda}{\sum_k p_k^\lambda}$
  - Pénalisation $M_\lambda = \sum_i \widetilde{p}_i l_i$ et code / distrib. compagne $\{\widetilde{p}_i\}_i$
  - $\alpha = 1$, $\left(\frac{10\sqrt{2}-12}{7}, \frac{10-6\sqrt{2}}{7}, \frac{5\sqrt{2}-6}{7}, \frac{5-3\sqrt{2}}{7}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}\right)$
  - $L = 2.31 > L^*$, mais besoin de 4 bits
  - $H_{.5}(X) = 2.42$ et $C_1 = 2.46$; $H(\widetilde{X}) = 1.85$ et $M_{.5} = 2.69$.

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $l_i = -\log_2 p_i$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$
  - $C_0 = L$ et $C_\infty = \max_i l_i$
  - $l_i = -\log_2 \widetilde{p}_i$ et $C_\alpha = H_\lambda(X)$, avec $\lambda = \frac{1}{1+\alpha}$ et $\widetilde{p}_i = \frac{p_i^\lambda}{\sum_k p_k^\lambda}$
  - Pénalisation $M_\lambda = \sum_i \widetilde{p}_i l_i$ et code / distrib. compagne $\{\widetilde{p}_i\}_i$
  - $\alpha = 1$, $\left(\frac{10\sqrt{2}-12}{7}, \frac{10-6\sqrt{2}}{7}, \frac{5\sqrt{2}-6}{7}, \frac{5-3\sqrt{2}}{7}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}\right)$
  - $L = 2.31 > L^*$, mais besoin de 4 bits
  - $H_{.5}(X) = 2.42$ et $C_1 = 2.46$; $H(\widetilde{X}) = 1.85$ et $M_{.5} = 2.69$.

Exemple de codage — Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\to$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $\underline{l_i = -\log_2 p_i}$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$
  - $C_0 = L$ et $C_\infty = \max_i l_i$
  - $\underline{l_i = -\log_2 \widetilde{p}_i}$ et $C_\alpha = H_\lambda(X)$, avec $\lambda = \frac{1}{1+\alpha}$ et $\widetilde{p}_i = \frac{p_i^\lambda}{\sum_k p_k^\lambda}$
  - Pénalisation $M_\lambda = \sum_i \widetilde{p}_i l_i$ et code / distrib. compagne $\{\widetilde{p}_i\}_i$
  - $\alpha = 1$, $\left(\frac{10\sqrt{2}-12}{7}, \frac{10-6\sqrt{2}}{7}, \frac{5\sqrt{2}-6}{7}, \frac{5-3\sqrt{2}}{7}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}\right)$
    Choix $(00, 01, 100, 101, 1100, 1101, 1110, 1111)$
  - $L = 2.31 > L^*$, mais besoin de 4 bits
  - $H_{.5}(X) = 2.42$ et $C_1 = 2.46$; $H(\widetilde{X}) = 1.85$ et $M_{.5} = 2.69$.

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \le 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $\underline{l_i = -\log_2 p_i}$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \ge H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$
  - $C_0 = L$ et $C_\infty = \max_i l_i$
  - $\underline{l_i = -\log_2 \widetilde{p}_i}$ et $C_\alpha = H_\lambda(X)$, avec $\lambda = \frac{1}{1+\alpha}$ et $\widetilde{p}_i = \frac{p_i^\lambda}{\sum_k p_k^\lambda}$
  - Pénalisation $M_\lambda = \sum_i \widetilde{p}_i l_i$ et code / distrib. compagne $\{\widetilde{p}_i\}_i$
  - $\alpha = 1$, $\left(\frac{10\sqrt{2}-12}{7}, \frac{10-6\sqrt{2}}{7}, \frac{5\sqrt{2}-6}{7}, \frac{5-3\sqrt{2}}{7}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}\right)$
    Choix $(00, 01, 100, 101, 1100, 1101, 1110, 1111)$
  - $L = 2.31 > L^*$, mais besoin de 4 bits
  - $H_{.5}(X) = 2.42$ et $C_1 = 2.46$; $H(\widetilde{X}) = 1.85$ et $M_{.5} = 2.69$.

Exemple de codage Campbell '65, Bercher '09

Codage binaire des faces du dé à 8 faces pipé de loi $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$.

Code instantanné (mot non préfixe d'un autre) $\rightarrow$ Kraft $\sum_i 2^{-l_i} \leq 1$

- de longueur moyenne: $L = \sum_i p_i l_i$ minimale
  - $l_i = -\log_2 p_i$ et $L^* = H(X)$ $\quad (l_i = \lceil -\log p_i \rceil$ et $L \geq H(X))$
  - choix $(0, 10, 110, 1110, 111100, 111101, 111110, 111111)$
  - $L = 2 = H(X)$ mais malgré tout, besoin de 6 bits
- Pénalisation des mots longs : minimisation de $C_\lambda = \frac{1}{\alpha} \log_2 \sum_i p_i 2^{\alpha l_i}$
  - $C_0 = L$ et $C_\infty = \max_i l_i$
  - $l_i = -\log_2 \widetilde{p}_i$ et $C_\alpha = H_\lambda(X)$, avec $\lambda = \frac{1}{1+\alpha}$ et $\widetilde{p}_i = \frac{p_i^\lambda}{\sum_k p_k^\lambda}$
  - Pénalisation $M_\lambda = \sum_i \widetilde{p}_i l_i$ et code / distrib. compagne $\{\widetilde{p}_i\}_i$
  - $\alpha = 1$, $\left(\frac{10\sqrt{2}-12}{7}, \frac{10-6\sqrt{2}}{7}, \frac{5\sqrt{2}-6}{7}, \frac{5-3\sqrt{2}}{7}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}, \frac{5-3\sqrt{2}}{14}\right)$
    Choix $(00, 01, 100, 101, 1100, 1101, 1110, 1111)$
  - $L = 2.31 > L^*$, mais besoin de 4 bits
  - $H_{.5}(X) = 2.42$ et $C_1 = 2.46$; $H(\widetilde{X}) = 1.85$ et $M_{.5} = 2.69$.

Un peu de physique

Gaz parfait :

N particules, volume Λ,

$$\Omega_{\Lambda,N,E} = \left\{ (v_1, \ldots, v_N) \,\middle|\, \mathcal{H} = \sum_{i=1}^{N} \frac{1}{2} m v_i^2 = E \right\} \qquad \text{(hypersphère)}$$

répartition uniforme sur la surface $\Omega_{\Lambda,N,E}$ (entropie maximum)

Un peu de physique

Gaz parfait :

N particules, volume Λ,

$$\Omega_{\Lambda,N,E} = \left\{ (v_1, \ldots, v_N) \,\middle|\, \mathcal{H} = \sum_{i=1}^{N} \frac{1}{2} m v_i^2 = E \right\} \qquad \text{(hypersphère)}$$

répartition uniforme sur la surface $\Omega_{\Lambda,N,E}$ (entropie maximum)

Sous-ensemble $\Lambda_0 \subset \Lambda$ de $N_0 < N$ particules :

$$p(v_1, \ldots, v_{N_0}) \propto \left(1 - \frac{m}{2E} \, \boldsymbol{v}_0^t \boldsymbol{v}_0 \right)^{\frac{3\,(N - N_0) - 2}{2}}$$

Un peu de physique

Gaz parfait :

N particules, volume Λ,

$$\Omega_{\Lambda,N,E} = \left\{ (v_1, \ldots, v_N) \,\middle|\, \mathcal{H} = \sum_{i=1}^{N} \frac{1}{2} m v_i^2 = E \right\} \qquad \text{(hypersphère)}$$

répartition uniforme sur la surface $\Omega_{\Lambda,N,E}$ (entropie maximum)

Sous-ensemble $\Lambda_0 \subset \Lambda$ de $N_0 < N$ particules :

$$p(v_1, \ldots, v_{N_0}) \propto \left(1 - \frac{m}{2E} \, \boldsymbol{v}_0^t \boldsymbol{v}_0 \right)^{\frac{3\,(N - N_0) - 2}{2}}$$

Limite thermodynamique : $\Lambda \to \mathbb{R}^3$, $\frac{N}{|\Lambda|} \to \rho$ et $\frac{E}{|\Lambda|} \to e$

$$p(v_1, \ldots, v_{N_0}) \propto \exp\left(- \frac{3 m \rho}{4 e} \, \boldsymbol{v}_0^t \boldsymbol{v}_0 \right) \qquad \text{(Boltzmann)}$$

Un peu de physique

Gaz parfait :

N particules, volume Λ,

$$\Omega_{\Lambda,N,E} = \left\{ (v_1, \ldots, v_N) \,\middle|\, \mathcal{H} = \sum_{i=1}^{N} \frac{1}{2} m v_i^2 = E \right\} \qquad \text{(hypersphère)}$$

répartition uniforme sur la surface $\Omega_{\Lambda,N,E}$ (entropie maximum)

Sous-ensemble $\Lambda_0 \subset \Lambda$ de $N_0 < N$ particules :

$$p(v_1, \ldots, v_{N_0}) \propto \left(1 - \frac{m}{2E} \, \boldsymbol{v}_0^t \boldsymbol{v}_0\right)^{\frac{3\,(N-N_0)-2}{2}}$$

Limite thermodynamique : $\Lambda \to \mathbb{R}^3$, $\frac{N}{|\Lambda|} \to \rho$ et $\frac{E}{|\Lambda|} \to e$

$$p(v_1, \ldots, v_{N_0}) \propto \exp\left(- \frac{3m\rho}{4e} \, \boldsymbol{v}_0^t \boldsymbol{v}_0\right) \qquad \text{(Boltzmann)} \quad \Rightarrow \text{ Shannon}$$

Un peu de physique

Gaz parfait :

N particules, volume Λ,

$$\Omega_{\Lambda,N,E} = \left\{ (v_1, \ldots, v_N) \,\middle|\, \mathcal{H} = \sum_{i=1}^{N} \frac{1}{2} m v_i^2 = E \right\} \qquad \text{(hypersphère)}$$

répartition uniforme sur la surface $\Omega_{\Lambda,N,E}$ (entropie maximum)

Sous-ensemble $\Lambda_0 \subset \Lambda$ de $N_0 < N$ particules :

$$p(v_1, \ldots, v_{N_0}) \propto \left(1 - \frac{m}{2E}\, \boldsymbol{v}_0^t \boldsymbol{v}_0\right)^{\frac{3\,(N-N_0)-2}{2}} \Rightarrow \text{Rényi, } \alpha = \frac{3N - 3N_0}{3N - 3N_0 - 2} > 1$$

Limite thermodynamique : $\Lambda \to \mathbb{R}^3$, $\frac{N}{|\Lambda|} \to \rho$ et $\frac{E}{|\Lambda|} \to e$

$$p(v_1, \ldots, v_{N_0}) \propto \exp\left(-\frac{3m\rho}{4e}\, \boldsymbol{v}_0^t \boldsymbol{v}_0\right) \qquad \text{(Boltzmann)} \quad \Rightarrow \text{Shannon}$$

Plan du cours

1 Entropie – mesure d'incertitude
- Axiomes, entropie de Shannon et ses propriétés
- Entropie différentielle – cas vectoriel
- D'autres mesures d'information

2 Information mutuelle - divergences
- Entropie conditionnelle, Information mutuelle
- Quelques exemples
- Divergences

3 Inégalités – relations entropiques
- Inégalités "classiques"
- Chaînes de Markov
- Inégalité de la puissance entropique
- Estimation – relations entre informations

4 Relations d'incertitude
- Relation d'incertitude d'Heisenberg et versions entropiques

Entropie conditionnelle

Soient $\boldsymbol{X}$ et $\boldsymbol{Y}$: degré d'incertitude $H(\boldsymbol{X})$ et $H(\boldsymbol{Y})$

Quelle information/incertitude commune $\boldsymbol{X}$ et $\boldsymbol{Y}$ partagent-ils ?

Entropie conditionnelle

Soient $\boldsymbol{X}$ et $\boldsymbol{Y}$: degré d'incertitude $H(\boldsymbol{X})$ et $H(\boldsymbol{Y})$

Quelle information/incertitude commune $\boldsymbol{X}$ et $\boldsymbol{Y}$ partagent-ils ?

Entropie conditionnelle ou incertitude sur $\boldsymbol{X}$ connaissant $\boldsymbol{Y}$

$$\begin{aligned}
H(\boldsymbol{X}|\boldsymbol{Y}) &= E_{\boldsymbol{Y}}\left[H\left(\boldsymbol{X}\,\middle|_{\boldsymbol{Y}}\right)\right] \\
&= \int p_Y(\boldsymbol{y}) \left(-\int p_{X|Y}(\boldsymbol{x},\boldsymbol{y}) \log p_{X|Y}(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{x}\right) d\boldsymbol{y} \\
&= -\iint p_{X,Y}(\boldsymbol{x},\boldsymbol{y}) \log p_{X|Y}(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{x} d\boldsymbol{y}
\end{aligned}$$

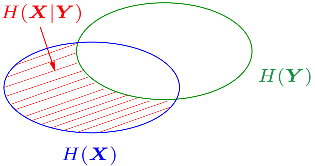


Diagramme de Venn

Entropie conditionnelle

Soient $\boldsymbol{X}$ et $\boldsymbol{Y}$: degré d'incertitude $H(\boldsymbol{X})$ et $H(\boldsymbol{Y})$

Quelle information/incertitude commune $\boldsymbol{X}$ et $\boldsymbol{Y}$ partagent-ils ?

Entropie conditionnelle ou incertitude sur $\boldsymbol{X}$ connaissant $\boldsymbol{Y}$

$$\begin{aligned} H(\boldsymbol{X}|\boldsymbol{Y}) &= E_{\boldsymbol{Y}}\left[H\left(\boldsymbol{X}\left|_{\boldsymbol{Y}}\right.\right)\right] \\ &= \int p_Y(\boldsymbol{y})\left(-\int p_{X|Y}(\boldsymbol{x},\boldsymbol{y})\log p_{X|Y}(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{x}\right)d\boldsymbol{y} \\ &= -\iint p_{X,Y}(\boldsymbol{x},\boldsymbol{y})\log p_{X|Y}(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{x}d\boldsymbol{y} \end{aligned}$$

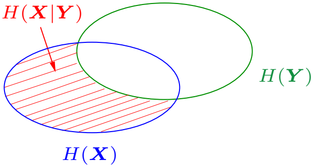
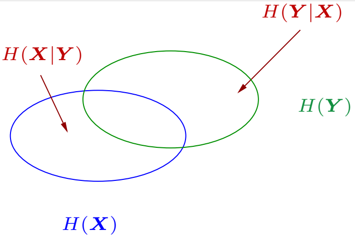


Diagramme de Venn

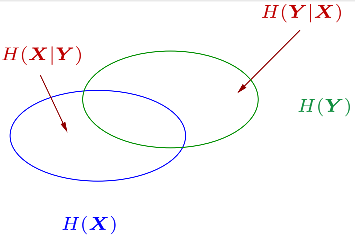
$$H(\boldsymbol{X},\boldsymbol{Y}) = H(\boldsymbol{X}|\boldsymbol{Y}) + H(\boldsymbol{Y})$$

Information mutuelle



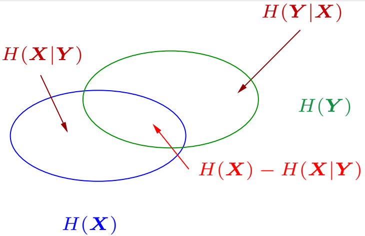
$$H(\boldsymbol{X},\boldsymbol{Y}) \quad = \quad H(\boldsymbol{X}|\boldsymbol{Y}) + H(\boldsymbol{Y})$$

Information mutuelle



$$
\begin{aligned}
H(\boldsymbol{X},\boldsymbol{Y}) &= H(\boldsymbol{X}|\boldsymbol{Y}) + H(\boldsymbol{Y})\\
&= H(\boldsymbol{Y}|\boldsymbol{X}) + H(\boldsymbol{X})
\end{aligned}
$$

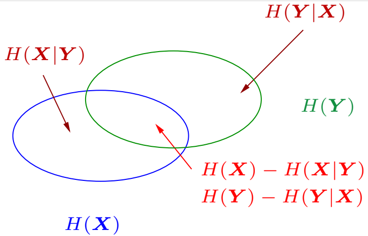
Information mutuelle



$$\begin{aligned} H(\boldsymbol{X},\boldsymbol{Y}) &= H(\boldsymbol{X}|\boldsymbol{Y}) + H(\boldsymbol{Y}) \\ &= H(\boldsymbol{Y}|\boldsymbol{X}) + H(\boldsymbol{X}) \end{aligned}$$

$$H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y})$$

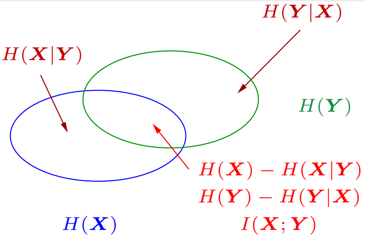
Information mutuelle



$$\begin{aligned} H(\boldsymbol{X},\boldsymbol{Y}) &= H(\boldsymbol{X}|\boldsymbol{Y}) + H(\boldsymbol{Y}) \\ &= H(\boldsymbol{Y}|\boldsymbol{X}) + H(\boldsymbol{X}) \end{aligned}$$

$$H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X})$$

Information mutuelle



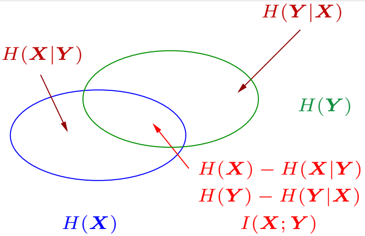
$$\begin{aligned} H(\boldsymbol{X},\boldsymbol{Y}) &= H(\boldsymbol{X}|\boldsymbol{Y}) + H(\boldsymbol{Y}) \\ &= H(\boldsymbol{Y}|\boldsymbol{X}) + H(\boldsymbol{X}) \end{aligned}$$

$$H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X})$$

Information mutuelle :

$$I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) = H(\boldsymbol{X}) + H(\boldsymbol{Y}) - H(\boldsymbol{X},\boldsymbol{Y})$$

Information mutuelle



$$\begin{aligned} H(\boldsymbol{X},\boldsymbol{Y}) &= H(\boldsymbol{X}|\boldsymbol{Y}) + H(\boldsymbol{Y}) \\ &= H(\boldsymbol{Y}|\boldsymbol{X}) + H(\boldsymbol{X}) \end{aligned}$$

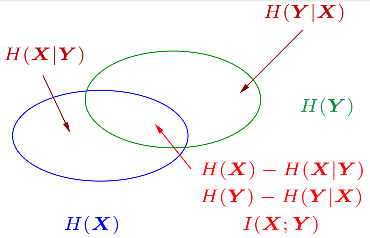
$$H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X})$$

Information mutuelle :

$$I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) = H(\boldsymbol{X}) + H(\boldsymbol{Y}) - H(\boldsymbol{X},\boldsymbol{Y})$$

$$I(\boldsymbol{X};\boldsymbol{Y}) = \int p_{X,Y}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{p_{X,Y}(\boldsymbol{x},\boldsymbol{y})}{p_X(\boldsymbol{x})p_Y(\boldsymbol{y})}\right) d\boldsymbol{x}d\boldsymbol{y}$$

Information mutuelle



$$\begin{aligned} H(\boldsymbol{X},\boldsymbol{Y}) &= H(\boldsymbol{X}|\boldsymbol{Y}) + H(\boldsymbol{Y}) \\ &= H(\boldsymbol{Y}|\boldsymbol{X}) + H(\boldsymbol{X}) \end{aligned}$$

$$H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X})$$

Information mutuelle :

$$I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) = H(\boldsymbol{X}) + H(\boldsymbol{Y}) - H(\boldsymbol{X},\boldsymbol{Y})$$

$$I(\boldsymbol{X};\boldsymbol{Y}) = \int p_{X,Y}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{p_{X,Y}(\boldsymbol{x},\boldsymbol{y})}{p_X(\boldsymbol{x})p_Y(\boldsymbol{y})}\right) d\boldsymbol{x}d\boldsymbol{y}$$

Information conditionnelle : $I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) = H(\boldsymbol{X}|\boldsymbol{Z}) - H(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})$

$$I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) = \int p_{X,Y,Z}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \log\left(\frac{p_{X,Y|Z}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{p_{X|Z}(\boldsymbol{x},\boldsymbol{z})p_{Y|Z}(\boldsymbol{y},\boldsymbol{z})}\right) d\boldsymbol{x}d\boldsymbol{y}d\boldsymbol{z}$$

Information mutuelle : exemples jouets

Calculer entropies, entropies conditionnelles et information mutuelle pour :

Exemple 1

$Y \backslash X$	1	2
1	0	$\frac{1}{4}$
2	$\frac{3}{4}$	0

Exemple 2

$Y \backslash X$	1	2
1	$\frac{3}{16}$	$\frac{1}{16}$
2	$\frac{9}{16}$	$\frac{3}{16}$

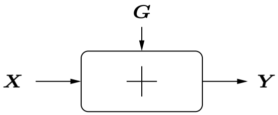
Exemple 3

$Y \backslash X$	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

Information mutuelle : canal Gaussien

Canal Gaussien et capacité canal :

Transmission d'un signal $\boldsymbol{X}$ dans un canal supposé être un "fil", mais perturbé par un bruit Gaussien.

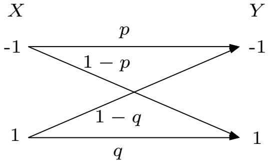


$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{G}$ avec $\boldsymbol{G}$ Gaussien de covariance $\boldsymbol{R}$ et $\boldsymbol{X}$ indépendant de $\boldsymbol{G}$. Par soucis de simplicité et sans perte de généralité on suppose $\boldsymbol{X}$ et $\boldsymbol{G}$ centrés. On suppose de plus que la puissance du signal est limitée à $E[\boldsymbol{X}^t \boldsymbol{X}] \leq P_X$.

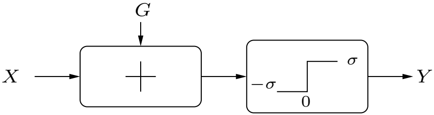
- Quelle est l'information mutuelle maximum possible entre $\boldsymbol{X}$ et $\boldsymbol{Y}$? Cette quantité est appelée capacité du canal.
- Pour quelle distribution de $\boldsymbol{X}$ atteint-t-on ce maximum ? On effectuera la détermination complète de p_X dans le cas 1 dimension.

Information mutuelle : canal symétrique

Canal binaire et capacité canal : transmission d'un signal $X \in \{-1;1\} \sim \{\alpha, 1-\alpha\}$ dans un canal représenté par la figure de droite.



- Les probabilités de transition étant fixées, pour quelle valeur de α atteint-t-on $I(X;Y)$ maximum ?
- Que devient α pour $p = q$ (canal binaire symétrique) ?
- Quelle est la capacité canal dans ce dernier cas ?
- On peut représenter ce canal de la manière suivante :



symbole $X \in \{-\sigma_X, \sigma_X\}$ émis, G Gaussien centré de variance σ^2. Déterminer p, q et conclure.

Divergence de Kullback-Leibler

"Distance" entre vecteurs aléatoires $\boldsymbol{P}$ et $\boldsymbol{Q}$, de distributions p et q :
Divergence de Kullback-Leibler (entropie relative)

$$D_{\mathrm{kl}}(p\|q) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x}$$

- Non-symétrique : q est une "référence"
- $D_{\mathrm{kl}} \geq 0$
- $D_{\mathrm{kl}} = 0 \Leftrightarrow p = q$ (p.p.)
- $I(\boldsymbol{X};\boldsymbol{Y}) = D_{\mathrm{kl}}(p_{X,Y}\|p_X p_Y)$ (symétrique) :
  $I \geq 0$ et est nulle si et seulement si il y a indépendance

Divergence de Kullback-Leibler

"Distance" entre vecteurs aléatoires $\boldsymbol{P}$ et $\boldsymbol{Q}$, de distributions p et q :
Divergence de Kullback-Leibler (entropie relative)

$$D_{\mathrm{kl}}(p\|q) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x}$$

- Non-symétrique : q est une "référence"
- $D_{\mathrm{kl}} \geq 0$
- $D_{\mathrm{kl}} = 0 \Leftrightarrow p = q$ (p.p.)
- $I(\boldsymbol{X};\boldsymbol{Y}) = D_{\mathrm{kl}}(p_{X,Y}\|p_X p_Y)$ (symétrique) :
  $I \geq 0$ et est nulle si et seulement si il y a indépendance

Divergence de Kullback-Leibler

"Distance" entre vecteurs aléatoires $\boldsymbol{P}$ et $\boldsymbol{Q}$, de distributions p et q :
Divergence de Kullback-Leibler (entropie relative)

$$D_{\mathrm{kl}}(p\|q) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x}$$

- Non-symétrique : q est une "référence"
- $D_{\mathrm{kl}} \geq 0$
- $D_{\mathrm{kl}} = 0 \Leftrightarrow p = q$ (p.p.)
- $I(\boldsymbol{X};\boldsymbol{Y}) = D_{\mathrm{kl}}(p_{X,Y}\|p_X p_Y)$ (symétrique) :
  $I \geq 0$ et est nulle si et seulement si il y a indépendance

Divergence de Kullback-Leibler

"Distance" entre vecteurs aléatoires $\boldsymbol{P}$ et $\boldsymbol{Q}$, de distributions p et q :
Divergence de Kullback-Leibler (entropie relative)

$$D_{\mathrm{kl}}(p\|q) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x}$$

- Non-symétrique : q est une "référence"
- $D_{\mathrm{kl}} \geq 0$
- $D_{\mathrm{kl}} = 0 \Leftrightarrow p = q$ (p.p.)
- $I(\boldsymbol{X};\boldsymbol{Y}) = D_{\mathrm{kl}}(p_{X,Y}\|p_X p_Y)$ (symétrique) :
  $I \geq 0$ et est nulle si et seulement si il y a indépendance

Divergence de Kullback-Leibler conditionnelle :

$$D_{\mathrm{kl}}(p_{X|Y}\|q_{X|Y}) = \int p_{X,Y}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{p_{X|Y}(\boldsymbol{x},\boldsymbol{y})}{q_{X|Y}(\boldsymbol{x},\boldsymbol{y})}\right) d\boldsymbol{x}d\boldsymbol{y}$$

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l \log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha - 1}{\alpha-1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0 < \alpha < 1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou $l^2 - 1$: divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha \geq 1$ (Vajda'73)

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l \log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha - 1}{\alpha - 1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0 < \alpha < 1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou $l^2 - 1$: divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha \geq 1$ (Vajda'73)

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l \log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha-1}{\alpha-1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0 < \alpha < 1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou $l^2 - 1$: divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha \geq 1$ (Vajda'73)

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l\log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha-1}{\alpha-1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0<\alpha<1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou l^2-1 : divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha \geq 1$ (Vajda'73)

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l \log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha-1}{\alpha-1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0 < \alpha < 1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou $l^2 - 1$: divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha \geq 1$ (Vajda'73)

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l\log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha-1}{\alpha-1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0<\alpha<1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou l^2-1 : divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha\ge 1$ (Vajda'73)

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l \log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha - 1}{\alpha - 1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0 < \alpha < 1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou $l^2 - 1$: divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha \geq 1$ (Vajda'73)

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l \log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha - 1}{\alpha - 1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0 < \alpha < 1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou $l^2 - 1$: divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha \geq 1$ (Vajda'73)

D'autres divergences

f-divergence (Ali-Silvey'66 ou Csizàr'67)

$$D_f(p\|q) = g\left(E_q\left[f\left(\frac{p}{q}\right)\right]\right)$$

avec f convexe et g fonction croissante

- $f(l) = l \log l$: divergence de Kullback-Leibler
- $f(l) = |\pi_0 - \pi_1 l|$: divergence de Kolmogorov ($P_{e,min}$ détec. bin.)
- $f(l) = l^\lambda$ et $g = \frac{1}{\lambda-1}\log$: $D_\lambda = \frac{1}{\lambda-1}\log\int q\left(\frac{p}{q}\right)^\lambda$ divergence de Rényi
- $f(l) = \frac{l^\alpha-1}{\alpha-1}$: divergence d'Hellinger (classification, estimation)
- $f(l) = -l^\alpha$ pour $0 < \alpha < 1$: la divergence de Chernoff (détection) ; pour $\alpha = 1/2$: divergence de Bhattacharya (estimation, détection)
- $f(l) = |l-1|$: divergence de la variation totale (distance L^1)
- $f(l) = (l-1)^2$ ou $l^2 - 1$: divergence χ^2-Pearson (estimation)
- Plus généralement $f(l) = |l-1|^\alpha$, $\alpha \geq 1$ (Vajda'73)

Zoom sur la matrice information de Fisher

Soit $\boldsymbol{X}_{\boldsymbol{\theta}} \sim p_{\boldsymbol{\theta}}$ paramétré par $\boldsymbol{\theta}$

- Matrice de Fisher :

$$\boldsymbol{J}(\boldsymbol{\theta}) = E\left[(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}})(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}})^t\right]$$

  C'est la covariance de la fonction score $\boldsymbol{S}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}} = \dfrac{\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}}}$

- Information de Fisher :

$$J(\boldsymbol{\theta}) = E\left[(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}})^t(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}})\right]$$

  C'est la trace de la matrice de Fisher

- $\boldsymbol{\theta}$ paramètre de position : $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x} - \boldsymbol{\theta})$ et $\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{x}} p_X(x)$

  le score devient $\boldsymbol{S} = -\nabla_{\boldsymbol{x}} \log p_X$

  Matrice et information de Fisher de $\boldsymbol{X}$: $\boldsymbol{J}(\boldsymbol{X})$ et $J(\boldsymbol{X}) = \mathrm{Tr}\left(\boldsymbol{J}(\boldsymbol{X})\right)$

$\boldsymbol{J}(a\boldsymbol{X}) = \frac{1}{a^2}\boldsymbol{J}(\boldsymbol{X})$: si $a \to \infty$ la matrice de Fisher est nulle. . .

Estimation et information de Fisher

Soit $\boldsymbol{X} \sim p_{\boldsymbol{\theta_0}}$ où $p_{\boldsymbol{\theta_0}} \in \{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$. On veut estimer $\boldsymbol{\theta_0}$ à partir de $\boldsymbol{X}$

Estimation et information de Fisher

Soit $\boldsymbol{X} \sim p_{\boldsymbol{\theta_0}}$ où $p_{\boldsymbol{\theta_0}} \in \{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$. On veut estimer $\boldsymbol{\theta_0}$ à partir de $\boldsymbol{X}$

$D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta_0}}\right) \geq 0$; égalité ssi $\boldsymbol{\theta} = \boldsymbol{\theta_0} \;\Rightarrow\; \boldsymbol{\nabla}_{\boldsymbol{\theta}} D|_{\boldsymbol{\theta_0}} = \mathbf{0}$ et $\boldsymbol{\mathcal{H}}_{\boldsymbol{\theta}} D|_{\boldsymbol{\theta_0}} \geq 0$

Estimation et information de Fisher

Soit $\boldsymbol{X} \sim p_{\boldsymbol{\theta_0}}$ où $p_{\boldsymbol{\theta_0}} \in \{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$. On veut estimer $\boldsymbol{\theta_0}$ à partir de $\boldsymbol{X}$

$D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta_0}}\right) \ge 0$; égalité ssi $\boldsymbol{\theta} = \boldsymbol{\theta_0} \ \Rightarrow \ \boldsymbol{\nabla}_{\boldsymbol{\theta}} D|_{\boldsymbol{\theta_0}} = \mathbf{0}$ et $\boldsymbol{\mathcal{H}}_{\boldsymbol{\theta}} D|_{\boldsymbol{\theta_0}} \ge 0$

On montre que $\boldsymbol{\mathcal{H}}_{\boldsymbol{\theta}} D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta_0}}\right)\Big|_{\boldsymbol{\theta_0}} = \boldsymbol{J}(\boldsymbol{\theta_0})$ matrice de Fisher

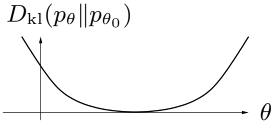
Estimation et information de Fisher

Soit $\boldsymbol{X} \sim p_{\boldsymbol{\theta_0}}$ où $p_{\boldsymbol{\theta_0}} \in \{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$. On veut estimer $\boldsymbol{\theta_0}$ à partir de $\boldsymbol{X}$

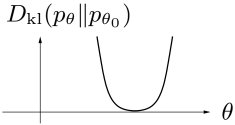
$D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta_0}}\right) \geq 0$; égalité ssi $\boldsymbol{\theta} = \boldsymbol{\theta_0} \;\Rightarrow\; \boldsymbol{\nabla}_{\boldsymbol{\theta}} D|_{\boldsymbol{\theta_0}} = \mathbf{0}$ et $\boldsymbol{\mathcal{H}}_{\boldsymbol{\theta}} D|_{\boldsymbol{\theta_0}} \geq 0$

On montre que $\boldsymbol{\mathcal{H}}_{\boldsymbol{\theta}} D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta_0}}\right)\Big|_{\boldsymbol{\theta_0}} = \boldsymbol{J}(\boldsymbol{\theta_0})$ matrice de Fisher

$$D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta}_0}\right) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta_0})^t \boldsymbol{J}(\boldsymbol{\theta_0})(\boldsymbol{\theta} - \boldsymbol{\theta_0}) + o\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right)$$



$J \ll$, "grosse incertitude"



$J \gg$, "faible incertitude"

Plan du cours

1. Entropie – mesure d'incertitude
   - Axiomes, entropie de Shannon et ses propriétés
   - Entropie différentielle – cas vectoriel
   - D'autres mesures d'information
2. Information mutuelle - divergences
   - Entropie conditionnelle, Information mutuelle
   - Quelques exemples
   - Divergences
3. Inégalités – relations entropiques
   - Inégalités "classiques"
   - Chaînes de Markov
   - Inégalité de la puissance entropique
   - Estimation – relations entre informations
4. Relations d'incertitude
   - Relation d'incertitude d'Heisenberg et versions entropiques

Inégalités triviales

- $\boldsymbol{X}$ discret sur un alphabet $\mathcal{X}$ de taille finie $|\mathcal{X}|$,

$$0 \le H(X) \le \log|\mathcal{X}| \qquad \text{(max. uniforme)}$$

- $\boldsymbol{X}$ continu sur un domaine $\mathcal{D}$ borné,

$$H(X) \le \log \operatorname{vol} \mathcal{D} \qquad \text{(max. uniforme)}$$

- $\boldsymbol{X}$ continu sur $\mathbb{R}^d$, de covariance donnée $\boldsymbol{R}$,

$$H(X) \le \frac{1}{2} \log\left((2\pi e)^d \, |\boldsymbol{R}| \right) \qquad \text{(max. Gaussien)}$$

Inégalités triviales

- $\boldsymbol{X}$ discret sur un alphabet $\mathcal{X}$ de taille finie $|\mathcal{X}|$,

$$0 \le H(X) \le \log|\mathcal{X}| \qquad \text{(max. uniforme)}$$

- $\boldsymbol{X}$ continu sur un domaine $\mathcal{D}$ borné,

$$H(X) \le \log \operatorname{vol} \mathcal{D} \qquad \text{(max. uniforme)}$$

- $\boldsymbol{X}$ continu sur $\mathbb{R}^d$, de covariance donnée R,

$$H(X) \le \frac{1}{2}\log\left((2\pi e)^d\,|R|\right) \qquad \text{(max. Gaussien)}$$

Inégalités triviales

- $\boldsymbol{X}$ discret sur un alphabet $\mathcal{X}$ de taille finie $|\mathcal{X}|$,

$$0 \leq H(X) \leq \log|\mathcal{X}| \qquad \text{(max. uniforme)}$$

- $\boldsymbol{X}$ continu sur un domaine $\mathcal{D}$ borné,

$$H(X) \leq \log \operatorname{vol} \mathcal{D} \qquad \text{(max. uniforme)}$$

- $\boldsymbol{X}$ continu sur $\mathbb{R}^d$, de covariance donnée $\boldsymbol{R}$,

$$H(X) \leq \frac{1}{2} \log\left((2\pi \mathrm{e})^d \, |\boldsymbol{R}|\right) \qquad \text{(max. Gaussien)}$$

Inégalités triviales

- $\boldsymbol{X}$ discret sur un alphabet $\mathcal{X}$ de taille finie $|\mathcal{X}|$,

$$0 \leq H(X) \leq \log|\mathcal{X}| \qquad \text{(max. uniforme)}$$

- $\boldsymbol{X}$ continu sur un domaine $\mathcal{D}$ borné,

$$H(X) \leq \log \operatorname{vol} \mathcal{D} \qquad \text{(max. uniforme)}$$

- $\boldsymbol{X}$ continu sur $\mathbb{R}^d$, de covariance donnée $\boldsymbol{R}$,

$$H(X) \leq \frac{1}{2} \log\left((2\pi \mathrm{e})^d\,|\boldsymbol{R}|\right) \qquad \text{(max. Gaussien)}$$

Toute densité de proba peut être vue comme maximisante de H, sous contrainte adéquate (loi expo./moyenne, loi de Cauchy/contrainte log.)

Règles de chaînage

- Sur l'entropie :

$$H(\boldsymbol{X}_n, \ldots, \boldsymbol{X}_1) = \sum_i H(\boldsymbol{X}_i | \boldsymbol{X}_{i-1}, \ldots, \boldsymbol{X}_1)$$

- Sur l'information mutuelle :

$$I(\boldsymbol{X}_n, \ldots, \boldsymbol{X}_1 ; \boldsymbol{Y}) = \sum_i I(\boldsymbol{X}_i ; \boldsymbol{Y} | \boldsymbol{X}_{i-1}, \ldots, \boldsymbol{X}_1)$$

- Sur la divergence de Kullback-Leibler :

$$D(p(\boldsymbol{x}_n, \ldots, \boldsymbol{x}_1) \| q(\boldsymbol{x}_n, \ldots, \boldsymbol{x}_1)) = \sum_i D(p(\boldsymbol{x}_i | \boldsymbol{x}_{i-1}, \ldots, \boldsymbol{x}_1) \| q(\boldsymbol{x}_i | \boldsymbol{x}_{i-1}, \ldots, \boldsymbol{x}_1))$$

Règles de chaînage

- Sur l'entropie :

$$H(\boldsymbol{X}_n, \ldots, \boldsymbol{X}_1) = \sum_i H(\boldsymbol{X}_i | \boldsymbol{X}_{i-1}, \ldots, \boldsymbol{X}_1)$$

- Sur l'information mutuelle :

$$I(\boldsymbol{X}_n, \ldots, \boldsymbol{X}_1; \boldsymbol{Y}) = \sum_i I(\boldsymbol{X}_i; \boldsymbol{Y} | \boldsymbol{X}_{i-1}, \ldots, \boldsymbol{X}_1)$$

- Sur la divergence de Kullback-Leibler :

$$D(p(\boldsymbol{x}_n, \ldots, \boldsymbol{x}_1) \| q(\boldsymbol{x}_n, \ldots, \boldsymbol{x}_1)) = \sum_i D(p(\boldsymbol{x}_i | \boldsymbol{x}_{i-1}, \ldots, \boldsymbol{x}_1) \| q(\boldsymbol{x}_i | \boldsymbol{x}_{i-1}, \ldots, \boldsymbol{x}_1))$$

Règles de chaînage

- Sur l'entropie :

$$H(\boldsymbol{X}_n, \ldots, \boldsymbol{X}_1) = \sum_i H(\boldsymbol{X}_i | \boldsymbol{X}_{i-1}, \ldots, \boldsymbol{X}_1)$$

- Sur l'information mutuelle :

$$I(\boldsymbol{X}_n, \ldots, \boldsymbol{X}_1; \boldsymbol{Y}) = \sum_i I(\boldsymbol{X}_i; \boldsymbol{Y} | \boldsymbol{X}_{i-1}, \ldots, \boldsymbol{X}_1)$$

- Sur la divergence de Kullback-Leibler :

$$D(p(\boldsymbol{x}_n, \ldots, \boldsymbol{x}_1) \| q(\boldsymbol{x}_n, \ldots, \boldsymbol{x}_1)) = \sum_i D(p(\boldsymbol{x}_i | \boldsymbol{x}_{i-1}, \ldots, \boldsymbol{x}_1) \| q(\boldsymbol{x}_i | \boldsymbol{x}_{i-1}, \ldots, \boldsymbol{x}_1))$$

Conditionnement, marginales, concavité

- Conditionner réduit l'entropie :

$$H(\boldsymbol{X}|\boldsymbol{Y}) \le H(\boldsymbol{X}) \quad \text{avec égalité ssi } \boldsymbol{X} \text{ et } \boldsymbol{Y} \text{ sont indépendants}$$

- Incertitude globale plus petite que la somme des incertitudes :

$$H(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \le \sum_i H(\boldsymbol{X}_i) \quad \text{égalité ssi les } \boldsymbol{X}_i \text{ sont indépendants}$$

- Concavité de H : pour $\lambda_i \ge 0$ tels que $\sum_i \lambda_i = 1$,

$$H\left(\sum_i \lambda_i \boldsymbol{X}_i\right) \ge \sum_i \lambda_i H(\boldsymbol{X}_i)$$

Conditionnement, marginales, concavité

- Conditionner réduit l'entropie :

$$H(\boldsymbol{X}|\boldsymbol{Y}) \le H(\boldsymbol{X}) \quad \text{avec égalité ssi } \boldsymbol{X} \text{ et } \boldsymbol{Y} \text{ sont indépendants}$$

- Incertitude globale plus petite que la somme des incertitudes :

$$H(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \le \sum_i H(\boldsymbol{X}_i) \quad \text{égalité ssi les } \boldsymbol{X}_i \text{ sont indépendants}$$

    - Inégalité de Hadamard : $|R| \le \prod_i R_{i,i}$

    - Processus stationnaire : $\mathcal{H}(\boldsymbol{X}) \le H(\boldsymbol{X}_1)$, égalité ssi cas iid

- Concavité de H : pour $\lambda_i \ge 0$ tels que $\sum_i \lambda_i = 1$,

$$H\left(\sum_i \lambda_i \boldsymbol{X}_i\right) \ge \sum_i \lambda_i H(\boldsymbol{X}_i)$$

Conditionnement, marginales, concavité

- Conditionner réduit l'entropie :

$$H(\boldsymbol{X}|\boldsymbol{Y}) \le H(\boldsymbol{X}) \quad \text{avec égalité ssi } \boldsymbol{X} \text{ et } \boldsymbol{Y} \text{ sont indépendants}$$

- Incertitude globale plus petite que la somme des incertitudes :

$$H(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \le \sum_i H(\boldsymbol{X}_i) \quad \text{égalité ssi les } \boldsymbol{X}_i \text{ sont indépendants}$$

  - Inégalité de Hadamard : $|\boldsymbol{R}| \le \prod_i R_{i,i}$

  - Processus stationnaire : $\mathcal{H}(\boldsymbol{X}) \le H(\boldsymbol{X}_1)$, égalité ssi cas iid

- Concavité de H : pour $\lambda_i \ge 0$ tels que $\sum_i \lambda_i = 1$,

$$H\left(\sum_i \lambda_i \boldsymbol{X}_i\right) \ge \sum_i \lambda_i H(\boldsymbol{X}_i)$$

Conditionnement, marginales, concavité

- Conditionner réduit l'entropie :

$$H(\boldsymbol{X}|\boldsymbol{Y}) \leq H(\boldsymbol{X}) \quad \text{avec égalité ssi } \boldsymbol{X} \text{ et } \boldsymbol{Y} \text{ sont indépendants}$$

- Incertitude globale plus petite que la somme des incertitudes :

$$H(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \leq \sum_i H(\boldsymbol{X}_i) \quad \text{égalité ssi les } \boldsymbol{X}_i \text{ sont indépendants}$$

  - Inégalité de Hadamard : $|\boldsymbol{R}| \leq \prod_i R_{i,i}$
  - Processus stationnaire : $\mathcal{H}(\boldsymbol{X}) \leq H(\boldsymbol{X_1})$, égalité ssi cas iid

- Concavité de H : pour $\lambda_i \geq 0$ tels que $\sum_i \lambda_i = 1$,

$$H\left(\sum_i \lambda_i \boldsymbol{X}_i\right) \geq \sum_i \lambda_i H(\boldsymbol{X}_i)$$

Conditionnement, marginales, concavité

- Conditionner réduit l'entropie :

$$H(\boldsymbol{X}|\boldsymbol{Y}) \leq H(\boldsymbol{X}) \quad \text{avec égalité ssi } \boldsymbol{X} \text{ et } \boldsymbol{Y} \text{ sont indépendants}$$

- Incertitude globale plus petite que la somme des incertitudes :

$$H(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \leq \sum_i H(\boldsymbol{X}_i) \quad \text{égalité ssi les } \boldsymbol{X}_i \text{ sont indépendants}$$

  - Inégalité de Hadamard : $|\boldsymbol{R}| \leq \prod_i R_{i,i}$
  - Processus stationnaire : $\mathcal{H}(\boldsymbol{X}) \leq H(\boldsymbol{X}_{\boldsymbol{1}})$, égalité ssi cas iid

- Concavité de H : pour $\lambda_i \geq 0$ tels que $\sum_i \lambda_i = 1$,

$$H\left(\sum_i \lambda_i \boldsymbol{X}_i\right) \geq \sum_i \lambda_i H(\boldsymbol{X}_i)$$

Inégalités et chaînes de Markov

Chaîne de Markov $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z}$

si $\quad p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y}) \Leftrightarrow p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{y})$ (relation symétrique)

Processus de Markov : $\boldsymbol{X}_1 \mapsto \ldots \boldsymbol{X}_n \mapsto \ldots$ $p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$ proba. de transition

- Data Processing Theorem :

$$\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z} \Rightarrow I(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{Z})$$

- 2^{nd} loi de la thermodynamique (système isolé, H est croissante)
  Soit $\{\boldsymbol{X}_n\}$ processus de Markov

Inégalités et chaînes de Markov

Chaîne de Markov $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z}$

si $\quad p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y}) \Leftrightarrow p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{y})$ (relation symétrique)

Processus de Markov : $\boldsymbol{X}_1 \mapsto \ldots \boldsymbol{X}_n \mapsto \ldots$ $p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$ proba. de transition

- Data Processing Theorem :
$$\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z} \;\Rightarrow\; I(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{Z})$$

- 2^{nd} loi de la thermodynamique (système isolé, H est croissante)
Soit $\{\boldsymbol{X}_n\}$ processus de Markov

Inégalités et chaînes de Markov

Chaîne de Markov $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z}$

si $\quad p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y}) p(\boldsymbol{z}|\boldsymbol{y}) \Leftrightarrow p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{y})$ (relation symétrique)

Processus de Markov : $\boldsymbol{X}_1 \mapsto \ldots \boldsymbol{X}_n \mapsto \ldots$ $p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$ proba. de transition

- Data Processing Theorem :

  $$\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z} \Rightarrow I(\boldsymbol{X}; \boldsymbol{Y}) \geq I(\boldsymbol{X}; \boldsymbol{Z})$$

  Exemple : $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{f}(\boldsymbol{Y})$ et donc $I(\boldsymbol{X}; \boldsymbol{Y}) \geq I(\boldsymbol{X}; \boldsymbol{f}(\boldsymbol{Y}))$

  Transformer des données ne fait que dégrader l'information

- 2$^{\text{nd}}$ loi de la thermodynamique (système isolé, H est croissante)

  Soit $\{\boldsymbol{X}_n\}$ processus de Markov

Inégalités et chaînes de Markov

Chaîne de Markov $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z}$

si $\quad p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y}) \Leftrightarrow p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{y})$ (relation symétrique)

Processus de Markov : $\boldsymbol{X}_1 \mapsto \ldots \boldsymbol{X}_n \mapsto \ldots$ $p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$ proba. de transition

- Data Processing Theorem :
$$\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z} \;\Rightarrow\; I(\boldsymbol{X};\boldsymbol{Y}) \ge I(\boldsymbol{X};\boldsymbol{Z})$$
Exemple : $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{f}(\boldsymbol{Y})$ et donc $I(\boldsymbol{X};\boldsymbol{Y}) \ge I(\boldsymbol{X};\boldsymbol{f}(\boldsymbol{Y}))$
Transformer des données ne fait que dégrader l'information
- 2^{nd} loi de la thermodynamique (système isolé, H est croissante)
Soit $\{\boldsymbol{X}_n\}$ processus de Markov
  - p_n et q_n deux densités de probabilités de $\boldsymbol{X}_n$ (C.I. $\neq$) :
  $$D_{\mathrm{kl}}(p_{n+1}\|q_{n+1}) \le D_{\mathrm{kl}}(p_n\|q_n)$$
  - soit p^* densité de probabilité stationnaire :
  $$D_{\mathrm{kl}}(p_{n+1}\|p^*) \le D_{\mathrm{kl}}(p_n\|p^*)$$
  - si p^* est uniforme sur D et p_n définie sur le même domaine :
  $$H(\boldsymbol{X}_{n+1}) \ge H(\boldsymbol{X}_n)$$

Inégalités et chaînes de Markov

Chaîne de Markov $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z}$

si $\quad p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y}) \Leftrightarrow p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{y})$ (relation symétrique)

Processus de Markov : $\boldsymbol{X}_1 \mapsto \ldots \boldsymbol{X}_n \mapsto \ldots \; p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$ proba. de transition

- Data Processing Theorem :

  $$\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z} \;\Rightarrow\; I(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{Z})$$

  Exemple : $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{f}(\boldsymbol{Y})$ et donc $I(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{f}(\boldsymbol{Y}))$

  Transformer des données ne fait que dégrader l'information

- 2^{nd} loi de la thermodynamique (système isolé, H est croissante)
  Soit $\{\boldsymbol{X}_n\}$ processus de Markov
  - p_n et q_n deux densités de probabilités de $\boldsymbol{X}_n$ (C.I. $\neq$) :

    $$D_{\mathrm{kl}}(p_{n+1}\|q_{n+1}) \leq D_{\mathrm{kl}}(p_n\|q_n)$$

  - soit p^* densité de probabilité stationnaire :

    $$D_{\mathrm{kl}}(p_{n+1}\|p^*) \leq D_{\mathrm{kl}}(p_n\|p^*)$$

  - si p^* est uniforme sur $\mathcal{D}$ et p_n définie sur le même domaine :

    $$H(\boldsymbol{X}_{n+1}) \geq H(\boldsymbol{X}_n)$$

Inégalités et chaînes de Markov

Chaîne de Markov $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z}$

si $\quad p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y}) \Leftrightarrow p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{y})$ (relation symétrique)

Processus de Markov : $\boldsymbol{X}_1 \mapsto \ldots \boldsymbol{X}_n \mapsto \ldots$ $p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$ proba. de transition

- Data Processing Theorem :

  $$\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z} \;\Rightarrow\; I(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{Z})$$

  Exemple : $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{f}(\boldsymbol{Y})$ et donc $I(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{f}(\boldsymbol{Y}))$

  Transformer des données ne fait que dégrader l'information

- 2^{nd} loi de la thermodynamique (système isolé, H est croissante)
  Soit $\{\boldsymbol{X}_n\}$ processus de Markov
  - p_n et q_n deux densités de probabilités de $\boldsymbol{X}_n$ (C.I. $\neq$) :

    $$D_{\text{kl}}(p_{n+1}\|q_{n+1}) \leq D_{\text{kl}}(p_n\|q_n)$$

  - soit p^* densité de probabilité stationnaire :

    $$D_{\text{kl}}(p_{n+1}\|p^*) \leq D_{\text{kl}}(p_n\|p^*)$$

  - si p^* est uniforme sur $\mathcal{D}$ et p_n définie sur le même domaine :

    $$H(\boldsymbol{X}_{n+1}) \geq H(\boldsymbol{X}_n)$$

Inégalités et chaînes de Markov

Chaîne de Markov $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z}$

si $\quad p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y}) \Leftrightarrow p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{y})$ (relation symétrique)

Processus de Markov : $\boldsymbol{X}_1 \mapsto \ldots \boldsymbol{X}_n \mapsto \ldots$ $p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$ proba. de transition

- Data Processing Theorem :

$$\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{Z} \;\Rightarrow\; I(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{Z})$$

  Exemple : $\boldsymbol{X} \mapsto \boldsymbol{Y} \mapsto \boldsymbol{f(Y)}$ et donc $I(\boldsymbol{X};\boldsymbol{Y}) \geq I(\boldsymbol{X};\boldsymbol{f(Y)})$

  Transformer des données ne fait que dégrader l'information

- 2^{nd} loi de la thermodynamique (système isolé, H est croissante)
  Soit $\{\boldsymbol{X}_n\}$ processus de Markov
  - p_n et q_n deux densités de probabilités de $\boldsymbol{X}_n$ (C.I. $\neq$) :

$$D_{\mathrm{kl}}(p_{n+1}\|q_{n+1}) \leq D_{\mathrm{kl}}(p_n\|q_n)$$

  - soit p^* densité de probabilité stationnaire :

$$D_{\mathrm{kl}}(p_{n+1}\|p^*) \leq D_{\mathrm{kl}}(p_n\|p^*)$$

  - si p^* est uniforme sur $\mathcal{D}$ et p_n définie sur le même domaine :

$$H(\boldsymbol{X}_{n+1}) \geq H(\boldsymbol{X}_n)$$

Exercice sur une chaîne de Markov

Soit la chaîne de Markov $\{X_n\}$ binaire de matrice de transition $\boldsymbol{P} = \begin{pmatrix} p & 1-q \\ 1-p & q \end{pmatrix}$ (processus homogène car $\boldsymbol{P}$ indépendant de n).

$\boldsymbol{P}$: transition du canal binaire.

- Calculer le(s) vecteur(s) de probabilité stationnaire(s) $\boldsymbol{p}^*$. Sous quelle condition y en a-t-il un uniforme ?
- Canal sans erreur, $p = q = 1$: quel est le comportement de $H(X_n)$ et de $D_{\mathrm{kl}}(\boldsymbol{p_n} \| \boldsymbol{q_n})$?
- Canal "inverseur", $p = q = 0$: quel est le comportement de $H(X_n)$, de $D_{\mathrm{kl}}(\boldsymbol{p_n} \| \boldsymbol{q_n})$ et de $D_{\mathrm{kl}}(\boldsymbol{p_n} \| \boldsymbol{p}^*)$? Comment se comporte $\boldsymbol{p_n}$?
- $p \neq 0, 1$ et $q \neq 0, 1$:
  - Montrer que $\boldsymbol{p_n} \to \boldsymbol{p}^*$
  - Soit $p = 0.7$, $q = 0.5$ et $\boldsymbol{p}_0 = \begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix}$. Calculer $H(X_0)$, $H(X_1)$ et $H(X_2)$ ainsi que $H(X_\infty)$. Conclusion.

Inégalité de la puissance entropique

- Inégalité de la puissance entropique :
  $\boldsymbol{X}$, $\boldsymbol{Y}$ continus sur $\mathbb{R}^d$ et indépendants,

$$N(\boldsymbol{X}+\boldsymbol{Y}) \geq N(\boldsymbol{X}) + N(\boldsymbol{Y}) \qquad \text{(non triviale)}$$

  égalité si et seulement si $\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}(\cdot, \boldsymbol{R}_{X,Y})$ avec $\boldsymbol{R}_Y \propto \boldsymbol{R}_X$

  - Conséquence : inégalité matricielle de Minkovsky ; $\boldsymbol{R}_1$ et $\boldsymbol{R}_2$ symétriques définies positives, $|\boldsymbol{R}_1+\boldsymbol{R}_2|^{\frac{1}{d}} \geq |\boldsymbol{R}_1|^{\frac{1}{d}} + |\boldsymbol{R}_2|^{\frac{1}{d}}$

- Equivalence : $\boldsymbol{X}$ et $\boldsymbol{Y}$ continus sur $\mathbb{R}^d$, indépendants, $\widetilde{\boldsymbol{X}}$ et $\widetilde{\boldsymbol{Y}}$ Gaussiens indépendants de matrices de covariances proportionnelles et tels que $H\left(\widetilde{\boldsymbol{X}}\right) = H(\boldsymbol{X})$ et $H\left(\widetilde{\boldsymbol{Y}}\right) = H(\boldsymbol{Y})$,

$$H(\boldsymbol{X}+\boldsymbol{Y}) \geq H\left(\widetilde{\boldsymbol{X}}+\widetilde{\boldsymbol{Y}}\right)$$

- Inégalité dite de préservation de covariance (équivalente)

$$H\left(\sqrt{\lambda}\,\boldsymbol{X} + \sqrt{1-\lambda}\,\boldsymbol{Y}\right) \geq \lambda H(\boldsymbol{X}) + (1-\lambda) H(\boldsymbol{Y})$$

Inégalité de la puissance entropique

- Inégalité de la puissance entropique :
  $\boldsymbol{X}, \boldsymbol{Y}$ continus sur $\mathbb{R}^d$ et indépendants,

$$N(\boldsymbol{X}+\boldsymbol{Y}) \geq N(\boldsymbol{X}) + N(\boldsymbol{Y}) \qquad \text{(non triviale)}$$

  égalité si et seulement si $\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}(\cdot, \boldsymbol{R}_{X,Y})$ avec $\boldsymbol{R}_Y \propto \boldsymbol{R}_X$

  - Conséquence : inégalité matricielle de Minkovsky ; $\boldsymbol{R_1}$ et $\boldsymbol{R_2}$ symétriques définies positives, $|\boldsymbol{R_1}+\boldsymbol{R_2}|^{\frac{1}{d}} \geq |\boldsymbol{R_1}|^{\frac{1}{d}} + |\boldsymbol{R_2}|^{\frac{1}{d}}$

- Equivalence : $\boldsymbol{X}$ et $\boldsymbol{Y}$ continus sur $\mathbb{R}^d$, indépendants, $\widetilde{\boldsymbol{X}}$ et $\widetilde{\boldsymbol{Y}}$ Gaussiens indépendants de matrices de covariances proportionnelles et tels que $H\left(\widetilde{\boldsymbol{X}}\right) = H(\boldsymbol{X})$ et $H\left(\widetilde{\boldsymbol{Y}}\right) = H(\boldsymbol{Y})$,

$$H(\boldsymbol{X}+\boldsymbol{Y}) \geq H\left(\widetilde{\boldsymbol{X}}+\widetilde{\boldsymbol{Y}}\right)$$

- Inégalité dite de préservation de covariance (équivalente)

$$H\left(\sqrt{\lambda}\,\boldsymbol{X}+\sqrt{1-\lambda}\,\boldsymbol{Y}\right) \geq \lambda H(\boldsymbol{X}) + (1-\lambda) H(\boldsymbol{Y})$$

Inégalité de la puissance entropique

- Inégalité de la puissance entropique :
  $\boldsymbol{X}, \boldsymbol{Y}$ continus sur $\mathbb{R}^d$ et indépendants,
  $$N(\boldsymbol{X}+\boldsymbol{Y}) \geq N(\boldsymbol{X}) + N(\boldsymbol{Y}) \qquad \text{(non triviale)}$$
  égalité si et seulement si $\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}(\cdot, \boldsymbol{R}_{X,Y})$ avec $\boldsymbol{R}_Y \propto \boldsymbol{R}_X$
    - Conséquence : inégalité matricielle de Minkovsky ; $\boldsymbol{R_1}$ et $\boldsymbol{R_2}$ symétriques définies positives, $|\boldsymbol{R_1}+\boldsymbol{R_2}|^{\frac{1}{d}} \geq |\boldsymbol{R_1}|^{\frac{1}{d}} + |\boldsymbol{R_2}|^{\frac{1}{d}}$
- Equivalence : $\boldsymbol{X}$ et $\boldsymbol{Y}$ continus sur $\mathbb{R}^d$, indépendants, $\widetilde{\boldsymbol{X}}$ et $\widetilde{\boldsymbol{Y}}$ Gaussiens indépendants de matrices de covariances proportionnelles et tels que $H\left(\widetilde{\boldsymbol{X}}\right) = H(\boldsymbol{X})$ et $H\left(\widetilde{\boldsymbol{Y}}\right) = H(\boldsymbol{Y})$,
  $$H(\boldsymbol{X}+\boldsymbol{Y}) \geq H\left(\widetilde{\boldsymbol{X}}+\widetilde{\boldsymbol{Y}}\right)$$
- Inégalité dite de préservation de covariance (équivalente)
  $$H\left(\sqrt{\lambda}\,\boldsymbol{X}+\sqrt{1-\lambda}\,\boldsymbol{Y}\right) \geq \lambda H(\boldsymbol{X}) + (1-\lambda) H(\boldsymbol{Y})$$

Inégalité de la puissance entropique

- Inégalité de la puissance entropique :
  $\boldsymbol{X}, \boldsymbol{Y}$ continus sur $\mathbb{R}^d$ et indépendants,

$$N(\boldsymbol{X}+\boldsymbol{Y}) \geq N(\boldsymbol{X}) + N(\boldsymbol{Y}) \qquad \text{(non triviale)}$$

  égalité si et seulement si $\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}\left(\cdot, \boldsymbol{R}_{X,Y}\right)$ avec $\boldsymbol{R}_Y \propto \boldsymbol{R}_X$

  - Conséquence : inégalité matricielle de Minkovsky ; $\boldsymbol{R_1}$ et $\boldsymbol{R_2}$ symétriques définies positives, $|\boldsymbol{R_1}+\boldsymbol{R_2}|^{\frac{1}{d}} \geq |\boldsymbol{R_1}|^{\frac{1}{d}} + |\boldsymbol{R_2}|^{\frac{1}{d}}$

- Equivalence : $\boldsymbol{X}$ et $\boldsymbol{Y}$ continus sur $\mathbb{R}^d$, indépendants, $\widetilde{\boldsymbol{X}}$ et $\widetilde{\boldsymbol{Y}}$ Gaussiens indépendants de matrices de covariances proportionnelles et tels que $H\left(\widetilde{\boldsymbol{X}}\right) = H(\boldsymbol{X})$ et $H\left(\widetilde{\boldsymbol{Y}}\right) = H(\boldsymbol{Y})$,

$$H\left(\boldsymbol{X}+\boldsymbol{Y}\right) \geq H\left(\widetilde{\boldsymbol{X}}+\widetilde{\boldsymbol{Y}}\right)$$

- Inégalité dite de préservation de covariance (équivalente)

$$H\left(\sqrt{\lambda}\,\boldsymbol{X}+\sqrt{1-\lambda}\,\boldsymbol{Y}\right) \geq \lambda H\left(\boldsymbol{X}\right) + (1-\lambda) H\left(\boldsymbol{Y}\right)$$

Inégalité de la puissance entropique

- Inégalité de la puissance entropique :
  $\boldsymbol{X}, \boldsymbol{Y}$ continus sur $\mathbb{R}^d$ et indépendants,

$$N(\boldsymbol{X}+\boldsymbol{Y}) \geq N(\boldsymbol{X}) + N(\boldsymbol{Y}) \qquad \text{(non triviale)}$$

  égalité si et seulement si $\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}(\cdot, \boldsymbol{R}_{X,Y})$ avec $\boldsymbol{R}_Y \propto \boldsymbol{R}_X$

  - Conséquence : inégalité matricielle de Minkovsky ; $\boldsymbol{R_1}$ et $\boldsymbol{R_2}$ symétriques définies positives, $|\boldsymbol{R_1}+\boldsymbol{R_2}|^{\frac{1}{d}} \geq |\boldsymbol{R_1}|^{\frac{1}{d}} + |\boldsymbol{R_2}|^{\frac{1}{d}}$

- Equivalence : $\boldsymbol{X}$ et $\boldsymbol{Y}$ continus sur $\mathbb{R}^d$, indépendants, $\widetilde{\boldsymbol{X}}$ et $\widetilde{\boldsymbol{Y}}$ Gaussiens indépendants de matrices de covariances proportionnelles et tels que $H\left(\widetilde{\boldsymbol{X}}\right) = H(\boldsymbol{X})$ et $H\left(\widetilde{\boldsymbol{Y}}\right) = H(\boldsymbol{Y})$,

$$H(\boldsymbol{X}+\boldsymbol{Y}) \geq H\left(\widetilde{\boldsymbol{X}}+\widetilde{\boldsymbol{Y}}\right)$$

- Inégalité dite de préservation de covariance (équivalente)

$$H\left(\sqrt{\lambda}\,\boldsymbol{X}+\sqrt{1-\lambda}\,\boldsymbol{Y}\right) \geq \lambda H(\boldsymbol{X}) + (1-\lambda) H(\boldsymbol{Y})$$

On peut en tirer entre autre des inégalités matricielles...

Estimation : Fisher, variance et divergence

Soit $\boldsymbol{X} \sim p_{\boldsymbol{\theta_0}}$ où $p_{\boldsymbol{\theta_0}} \in \{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$. On veut estimer $\boldsymbol{\theta_0}$ à partir de $\boldsymbol{X}$

$$D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta}_0}\right) = \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta_0})^t \boldsymbol{J}(\boldsymbol{\theta_0}) (\boldsymbol{\theta} - \boldsymbol{\theta_0}) + o\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right)$$

Estimation : Fisher, variance et divergence

Soit $\boldsymbol{X} \sim p_{\boldsymbol{\theta_0}}$ où $p_{\boldsymbol{\theta_0}} \in \{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$. On veut estimer $\boldsymbol{\theta_0}$ à partir de $\boldsymbol{X}$

$$D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta}_0}\right) = \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta_0}\right)^t \boldsymbol{J}(\boldsymbol{\theta_0})(\boldsymbol{\theta} - \boldsymbol{\theta_0}) + o\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right)$$

Cramér-Rao : estimateur $\widehat{\boldsymbol{\theta}}_{\boldsymbol{0}}$ de $\boldsymbol{\theta_0}$, non-biaisé et de covariance $\boldsymbol{R}_{\boldsymbol{\theta_0}}$,

$$\boldsymbol{R}_{\boldsymbol{\theta_0}} - \boldsymbol{J}(\boldsymbol{\theta_0})^{-1} \geq 0 \qquad \left(\boldsymbol{R}_{\boldsymbol{X}} - \boldsymbol{J}(\boldsymbol{X})^{-1} \geq 0\right)$$

Preuve : $J(\boldsymbol{\theta_0}) = E[\boldsymbol{S}\boldsymbol{S}^t]$ avec $\boldsymbol{S} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta_0}}$ et $E[S(\widehat{\boldsymbol{\theta}}_{\boldsymbol{0}} - \boldsymbol{\theta_0})^t] = \boldsymbol{I}$

Cauchy-Schwarz $E[\boldsymbol{u}^t S(\widehat{\boldsymbol{\theta}}_{\boldsymbol{0}} - \boldsymbol{\theta_0})^t \boldsymbol{v}] \leq \boldsymbol{u}^t \boldsymbol{J}(\boldsymbol{\theta_0}) \boldsymbol{u} \boldsymbol{v}^t \boldsymbol{R}_{\boldsymbol{\theta_0}} \boldsymbol{v}$ en $\boldsymbol{u} = \boldsymbol{J}(\boldsymbol{\theta_0})^{-1} \boldsymbol{v}$

Estimation : Fisher, variance et divergence

Soit $\boldsymbol{X} \sim p_{\boldsymbol{\theta_0}}$ où $p_{\boldsymbol{\theta_0}} \in \{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$. On veut estimer $\boldsymbol{\theta_0}$ à partir de $\boldsymbol{X}$

$$D\left(p_{\boldsymbol{\theta}} \middle\| p_{\boldsymbol{\theta}_0}\right) = \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta_0}\right)^t \boldsymbol{J}(\boldsymbol{\theta_0})(\boldsymbol{\theta} - \boldsymbol{\theta_0}) + o\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right)$$

Cramér-Rao : estimateur $\widehat{\boldsymbol{\theta}}_{\mathbf{0}}$ de $\boldsymbol{\theta_0}$, non-biaisé et de covariance $\boldsymbol{R}_{\boldsymbol{\theta_0}}$,

$$\boldsymbol{R}_{\boldsymbol{\theta_0}} - \boldsymbol{J}(\boldsymbol{\theta_0})^{-1} \geq 0 \qquad \left(\boldsymbol{R}_{\boldsymbol{X}} - \boldsymbol{J}(\boldsymbol{X})^{-1} \geq 0\right)$$

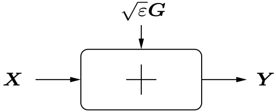
Preuve : $J(\boldsymbol{\theta_0}) = E[\boldsymbol{S}\boldsymbol{S}^t]$ avec $\boldsymbol{S} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta_0}}$ et $E[S(\widehat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta_0})^t] = \boldsymbol{I}$

Cauchy-Schwarz $E[\boldsymbol{u}^t S(\widehat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta_0})^t \boldsymbol{v}] \leq \boldsymbol{u}^t \boldsymbol{J}(\boldsymbol{\theta_0})\boldsymbol{u}\boldsymbol{v}^t \boldsymbol{R}_{\boldsymbol{\theta_0}}\boldsymbol{v}$ en $\boldsymbol{u} = \boldsymbol{J}(\boldsymbol{\theta_0})^{-1}\boldsymbol{v}$

Cas particulier : $R_{ii} \geq \left(\boldsymbol{J}^{-1}\right)_{ii}$ (scalaire $\sigma^2 \geq \frac{1}{J}$) et $\mathrm{Tr}\,(\boldsymbol{R}) \geq \frac{d^2}{\mathrm{Tr}(\boldsymbol{J})}$

Identité de De Bruijn

Soient $\boldsymbol{X}$ et $\boldsymbol{G}$ avec $\boldsymbol{G} \sim \mathcal{N}(\cdot, \boldsymbol{I})$ indépendant de $\boldsymbol{X}$. On a alors



$$\frac{\partial}{\partial \varepsilon} H\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right) = \frac{1}{2} J\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right)$$

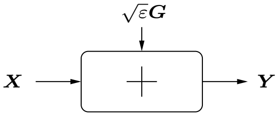
- Inégalité de la puissance entropique
- Inégalité isopérimétrique entropique

$$J(\boldsymbol{X}) N(\boldsymbol{X}) \geq d$$

- "Cramér-Rao" sur la somme des variances $\mathrm{Tr}\left(\boldsymbol{R}_X\right) \geq \frac{d^2}{\mathrm{Tr}(\boldsymbol{J}(\boldsymbol{X}))}$

Identité de De Bruijn

Soient $\boldsymbol{X}$ et $\boldsymbol{G}$ avec $\boldsymbol{G} \sim \mathcal{N}(\cdot, \boldsymbol{I})$ indépendant de $\boldsymbol{X}$. On a alors



$$\frac{\partial}{\partial \varepsilon} H\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right) = \frac{1}{2} J\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right)$$

Conséquences :

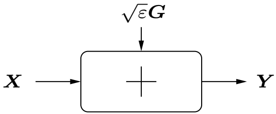
- Inégalité de la puissance entropique
- Inégalité isopérimétrique entropique

$$J(\boldsymbol{X}) N(\boldsymbol{X}) \geq d$$

- "Cramér-Rao" sur la somme des variances $\mathrm{Tr}\left(\boldsymbol{R}_X\right) \geq \frac{d^2}{\mathrm{Tr}(\boldsymbol{J}(\boldsymbol{X}))}$

Identité de De Bruijn

Soient $\boldsymbol{X}$ et $\boldsymbol{G}$ avec $\boldsymbol{G} \sim \mathcal{N}(\cdot, \boldsymbol{I})$ indépendant de $\boldsymbol{X}$. On a alors



$$\frac{\partial}{\partial \varepsilon} H\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right) = \frac{1}{2} J\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right)$$

Conséquences :

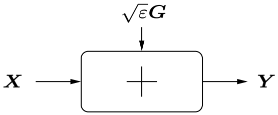
- Inégalité de la puissance entropique
- Inégalité isopérimétrique entropique

$$J(\boldsymbol{X}) N(\boldsymbol{X}) \geq d$$

- "Cramér-Rao" sur la somme des variances $\operatorname{Tr}(\boldsymbol{R}_X) \geq \frac{d^2}{\operatorname{Tr}(\boldsymbol{J}(\boldsymbol{X}))}$

Identité de De Bruijn

Soient $\boldsymbol{X}$ et $\boldsymbol{G}$ avec $\boldsymbol{G} \sim \mathcal{N}(\cdot, \boldsymbol{I})$ indépendant de $\boldsymbol{X}$. On a alors



$$\frac{\partial}{\partial \varepsilon} H\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right) = \frac{1}{2} J\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right)$$

Conséquences :

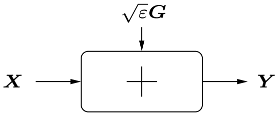
- Inégalité de la puissance entropique
- Inégalité isopérimétrique entropique

$$J(\boldsymbol{X})N(\boldsymbol{X}) \geq d$$

- "Cramér-Rao" sur la somme des variances $\mathrm{Tr}\left(\boldsymbol{R}_{\boldsymbol{X}}\right) \geq \frac{d^2}{\mathrm{Tr}\left(\boldsymbol{J}(\boldsymbol{X})\right)}$

Identité de De Bruijn

Soient $\boldsymbol{X}$ et $\boldsymbol{G}$ avec $\boldsymbol{G} \sim \mathcal{N}(\cdot, \boldsymbol{I})$ indépendant de $\boldsymbol{X}$. On a alors



$$\frac{\partial}{\partial \varepsilon} H\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right) = \frac{1}{2} J\left(\boldsymbol{X} + \sqrt{\varepsilon}\boldsymbol{G}\right)$$

Conséquences :

- Inégalité de la puissance entropique
- Inégalité isopérimétrique entropique

$$J(\boldsymbol{X})N(\boldsymbol{X}) \geq d$$

- "Cramér-Rao" sur la somme des variances $\operatorname{Tr}\left(\boldsymbol{R}_{\boldsymbol{X}}\right) \geq \frac{d^2}{\operatorname{Tr}\left(\boldsymbol{J}(\boldsymbol{X})\right)}$

Verdú'05 : $\displaystyle \frac{\partial}{\partial \varepsilon} I\left(\boldsymbol{X}; \sqrt{\varepsilon}\,\boldsymbol{X} + \boldsymbol{G}\right) = \frac{1}{2} E\left[\left(\boldsymbol{X} - \left[\boldsymbol{X}|_{\sqrt{\varepsilon}\,\boldsymbol{X} + \boldsymbol{G}}\right]\right)^2\right]$

Plan du cours

1. Entropie – mesure d'incertitude
   - Axiomes, entropie de Shannon et ses propriétés
   - Entropie différentielle – cas vectoriel
   - D'autres mesures d'information
2. Information mutuelle - divergences
   - Entropie conditionnelle, Information mutuelle
   - Quelques exemples
   - Divergences
3. Inégalités – relations entropiques
   - Inégalités "classiques"
   - Chaînes de Markov
   - Inégalité de la puissance entropique
   - Estimation – relations entre informations
4. Relations d'incertitude
   - Relation d'incertitude d'Heisenberg et versions entropiques

Principe d'incertitude d'Heisenberg

Fonction d'onde $\Psi_d(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$
i.e. $p = |\Psi_d|^2$ est la densité de probabilité d'un vecteur aléatoire $\boldsymbol{X}$

Transformation de Fourier $\widehat{\Psi}_d(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \Psi_d(\boldsymbol{u}) \, \mathrm{e}^{-\imath \boldsymbol{u}^t \boldsymbol{x}} \, d\boldsymbol{u}$
i.e. $\bar{p} = |\widehat{\Psi}_d|^2$ densité de probabilité du vecteur aléatoire "conjugué" $\bar{\boldsymbol{X}}$

Principe d'incertitude d'Heisenberg

Fonction d'onde $\Psi_d(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$
i.e. $p = |\Psi_d|^2$ est la densité de probabilité d'un vecteur aléatoire $\boldsymbol{X}$

Transformation de Fourier $\widehat{\Psi}_d(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \Psi_d(\boldsymbol{u})\, \mathrm{e}^{-\imath \boldsymbol{u}^t \boldsymbol{x}}\, d\boldsymbol{u}$
i.e. $\bar{p} = |\widehat{\Psi}_d|^2$ densité de probabilité du vecteur aléatoire "conjugué" $\bar{\boldsymbol{X}}$

Principe d'incertitude d'Heisenberg :

$$4\, \boldsymbol{R}_{\boldsymbol{X}} - \boldsymbol{R}_{\bar{\boldsymbol{X}}} \geq 0$$

$$\left(\frac{E\left[\boldsymbol{X}^t \boldsymbol{X}\right]}{d} \, \frac{E\left[\bar{\boldsymbol{X}}^t \bar{\boldsymbol{X}}\right]}{d} \right)^{\frac{1}{2}} \geq \frac{1}{2}$$

égalité ssi $\boldsymbol{X} \sim \mathcal{N}$

Principe d'incertitude d'Heisenberg

Fonction d'onde $\Psi_d(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$
i.e. $p = |\Psi_d|^2$ est la densité de probabilité d'un vecteur aléatoire $\boldsymbol{X}$

Transformation de Fourier $\widehat{\Psi}_d(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \Psi_d(\boldsymbol{u})\, \mathrm{e}^{-\imath \boldsymbol{u}^t \boldsymbol{x}}\, d\boldsymbol{u}$
i.e. $\bar{p} = |\widehat{\Psi}_d|^2$ densité de probabilité du vecteur aléatoire "conjugué" $\bar{\boldsymbol{X}}$

Principe d'incertitude d'Heisenberg :

$$4\, \boldsymbol{R}_{\boldsymbol{X}} - \boldsymbol{R}_{\bar{\boldsymbol{X}}} \geq 0$$

$$\left(\frac{E\left[\boldsymbol{X}^t \boldsymbol{X}\right]}{d} \; \frac{E\left[\bar{\boldsymbol{X}}^t \bar{\boldsymbol{X}}\right]}{d} \right)^{\frac{1}{2}} \geq \frac{1}{2}$$

égalité ssi $\boldsymbol{X} \sim \mathcal{N}$

Ces relations ont un sens si les covariances existent :

Quid *e.g.* du cas Cauchy $p(\boldsymbol{x}) = \dfrac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{\frac{d+1}{2}}\; |\boldsymbol{\Sigma}|^{\frac{1}{2}}\; \left(1 + \boldsymbol{x}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{x}\right)^{\frac{d+1}{2}}}$?

Version entropique (Bialynicki-Birula & Mycielski, Com. in Math. Phys. '75)

$$\frac{H(\boldsymbol{X}) + H(\bar{X})}{d} \geq 1 + \log \pi$$

Version entropique (Bialynicki-Birula & Mycielski, Com. in Math. Phys. '75)

$$\frac{H(\boldsymbol{X}) + H(\bar{X})}{d} \geq 1 + \log \pi$$

Babenko'61 : $\quad \| \widehat{\Psi} \|_{2\alpha^*} \leq (C_{\alpha,\alpha^*}) \; \| \Psi \|_{2\alpha}$

$$C_{\alpha,\alpha^*} = \left(\frac{\pi}{\alpha^*}\right)^{\frac{1}{4\alpha^*}} \left(\frac{\pi}{\alpha}\right)^{-\frac{1}{4\alpha}} \text{ avec } \alpha \in \left[\tfrac{1}{2}\,;\,1\right] \text{ et } \frac{1}{2\alpha} + \frac{1}{2\alpha^*} = 1$$

Version entropique (Bialynicki-Birula & Mycielski, Com. in Math. Phys. '75)

$$\frac{H(\boldsymbol{X}) + H(\bar{X})}{d} \geq 1 + \log \pi$$

Beckner'75 : $\|\widehat{\Psi}_d\|_{2\alpha^*} \leq (C_{\alpha,\alpha^*})^d \, \| \Psi_d\|_{2\alpha}$

$$C_{\alpha,\alpha^*} = \left(\frac{\pi}{\alpha^*}\right)^{\frac{1}{4\alpha^*}} \left(\frac{\pi}{\alpha}\right)^{-\frac{1}{4\alpha}} \text{ avec } \alpha \in \left[\tfrac{1}{2}\,;\,1\right] \text{ et } \frac{1}{2\alpha} + \frac{1}{2\alpha^*} = 1$$

Version entropique (Bialynicki-Birula & Mycielski, Com. in Math. Phys. '75)

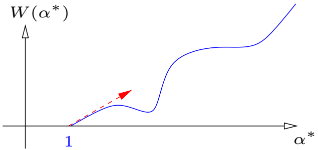
$$\frac{H(\boldsymbol{X}) + H(\bar{X})}{d} \geq 1 + \log \pi$$

Beckner'75 : $\|\widehat{\Psi}_d\|_{2\alpha^*} \leq (C_{\alpha,\alpha^*})^d \, \| \Psi_d\|_{2\alpha}$

$$C_{\alpha,\alpha^*} = \left(\frac{\pi}{\alpha^*}\right)^{\frac{1}{4\alpha^*}} \left(\frac{\pi}{\alpha}\right)^{-\frac{1}{4\alpha}} \text{ avec } \alpha \in \left[\tfrac{1}{2}\,;\,1\right] \text{ et } \frac{1}{2\alpha} + \frac{1}{2\alpha^*} = 1$$

$$W(\alpha^*) = (C_{\alpha,\alpha^*})^d \|\Psi_d\|_{2\alpha} - \|\widehat{\Psi}_d\|_{2\alpha^*} \geq 0 \text{ avec égalité pour } \alpha^* = 1 \text{ (Parseval)}$$

$$\Longrightarrow \left.\frac{dW}{d\beta}\right|_{\alpha^*=1^+} \geq 0$$



VERSION ENTROPIQUE (BIALYNICKI-BIRULA & MYCIELSKI, COM. IN MATH. PHYS. '75)

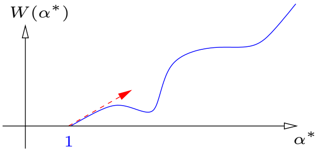
$$\frac{H(\boldsymbol{X}) + H(\bar{X})}{d} \geq 1 + \log \pi$$

Beckner'75 : $\quad \|\widehat{\Psi}_d\|_{2\alpha^*} \;\leq\; (C_{\alpha,\alpha^*})^d \,\|\,\Psi_d\|_{2\alpha}$

$$C_{\alpha,\alpha^*} = \left(\tfrac{\pi}{\alpha^*}\right)^{\frac{1}{4\alpha^*}} \left(\tfrac{\pi}{\alpha}\right)^{-\frac{1}{4\alpha}} \text{ avec } \alpha \in \left[\tfrac{1}{2}\,;\,1\right] \text{ et } \tfrac{1}{2\alpha} + \tfrac{1}{2\alpha^*} = 1$$

$$W(\alpha^*) = (C_{\alpha,\alpha^*})^d \|\Psi_d\|_{2\alpha} - \|\widehat{\Psi}_d\|_{2\alpha^*} \geq 0 \text{ avec égalité pour } \alpha^* = 1 \text{ (Parseval)}$$

$$\Longrightarrow \left.\frac{dW}{d\beta}\right|_{\alpha^*=1^+} \geq 0$$



- Insensible aux facteurs d'échelle $H(\boldsymbol{AX}) + H(\overline{\boldsymbol{AX}}) = H(\boldsymbol{X}) + H(\bar{\boldsymbol{X}})$
- Implique Heisenberg : $\left(N(\boldsymbol{X})N(\bar{\boldsymbol{X}})\right)^{\frac{1}{2}} \geq \frac{1}{2}$
- Egalité ssi $\boldsymbol{X}$ est Gaussien, minimisante de $N(\boldsymbol{X})N(\bar{\boldsymbol{X}})$

Généralisation à Rényi

Directement de la relation de Beckner, via $H_\alpha(\boldsymbol{X}) = \frac{1}{1-\alpha} \log \|\psi_d\|_{2\alpha}^{2\alpha}$:

$$\frac{H_\alpha(\boldsymbol{X}) + H_{\alpha^*}(\bar{\boldsymbol{X}})}{d} \geq \log \pi + \frac{\log \alpha}{2(\alpha - 1)} + \frac{\log \alpha^*}{2(\alpha^* - 1)} = B(\alpha)$$

- Valable pour $\alpha \geq \frac{1}{2}$ par involution $\bar{\bar{\boldsymbol{X}}} \overset{d}{=} \boldsymbol{X}$
- Insensible aux facteurs d'échelle (matriciel inversible)
- $\alpha \to 1$ Bialynicki-Birula & Mycielski est retrouvé
- égalité si et seulement si $\boldsymbol{X}$ est Gaussien
- Décroissance de H_λ versus λ :

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \begin{cases} 2\pi & \text{si} \quad (\alpha, \beta) \in [0; 1/2]^2 \\ B(\max(\alpha, \beta)) & \text{sinon et} \quad \beta \leq \alpha^* \\ \text{pas de rel. incert.} & \text{sinon} \end{cases}$$

Généralisation à Rényi

Directement de la relation de Beckner, via $H_\alpha(\boldsymbol{X}) = \frac{1}{1-\alpha} \log \|\psi_d\|_{2\alpha}^{2\alpha}$:

$$\frac{H_\alpha(\boldsymbol{X}) + H_{\alpha^*}(\bar{\boldsymbol{X}})}{d} \geq \log \pi + \frac{\log \alpha}{2(\alpha-1)} + \frac{\log \alpha^*}{2(\alpha^*-1)} = B(\alpha)$$

- Valable pour $\alpha \geq \frac{1}{2}$ par involution $\bar{\bar{\boldsymbol{X}}} \stackrel{\mathrm{d}}{=} \boldsymbol{X}$
- Insensible aux facteurs d'échelle (matriciel inversible)
- $\alpha \to 1$ Bialynicki-Birula & Mycielski est retrouvé
- égalité si et seulement si $\boldsymbol{X}$ est Gaussien

- Décroissance de H_λ versus λ :

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \begin{cases} 2\pi & \text{si} \quad (\alpha,\beta) \in [0;1/2]^2 \\ B(\max(\alpha,\beta)) & \text{sinon et} \quad \beta \leq \alpha^* \\ \text{pas de rel. incert.} & \text{sinon} \end{cases}$$

Généralisation à Rényi

Directement de la relation de Beckner, via $H_\alpha(\boldsymbol{X}) = \frac{1}{1-\alpha} \log \|\psi_d\|_{2\alpha}^{2\alpha}$:

$$\frac{H_\alpha(\boldsymbol{X}) + H_{\alpha^*}(\bar{\boldsymbol{X}})}{d} \geq \log \pi + \frac{\log \alpha}{2(\alpha - 1)} + \frac{\log \alpha^*}{2(\alpha^* - 1)} = B(\alpha)$$

- Valable pour $\alpha \geq \frac{1}{2}$ par involution $\bar{\bar{\boldsymbol{X}}} \stackrel{\mathrm{d}}{=} \boldsymbol{X}$
- Insensible aux facteurs d'échelle (matriciel inversible)
- $\alpha \to 1$ Bialynicki-Birula & Mycielski est retrouvé
- égalité si et seulement si $\boldsymbol{X}$ est Gaussien

- Décroissance de H_λ versus λ :

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \begin{cases} 2\pi & \text{si} \quad (\alpha, \beta) \in [0; 1/2]^2 \\ B(\max(\alpha, \beta)) & \text{sinon et} \quad \beta \leq \alpha^* \\ \text{pas de rel. incert.} & \text{sinon} \end{cases}$$

Cas discret – cas périodique

Ψ_d défini sur $\mathcal{A} = \{0, \ldots, n-1\}^d$, $\boldsymbol{X}$ vecteur aléatoire associé

- TF "réduite" : $\widehat{\Psi}_d^{(r)}(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \sum_{\boldsymbol{u} \in \mathcal{A}} \Psi_d(\boldsymbol{u})\, \mathrm{e}^{-\imath \boldsymbol{u}^t \boldsymbol{x}} \qquad \boldsymbol{x} \in [0; 2\pi)^d$

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \log(2\pi) \quad \text{pour } \beta \leq \alpha^*, \text{ égalité ssi } \boldsymbol{X} \text{ est déterministe}$$

- TF "discrète" : $\widehat{\Psi}_d^{(d)}(\boldsymbol{k}) = n^{-\frac{d}{2}} \sum_{\boldsymbol{u} \in \mathcal{A}} \Psi_d(\boldsymbol{u}) \mathrm{e}^{-\frac{2\imath\pi}{n} \boldsymbol{u}^t \boldsymbol{k}} \qquad \boldsymbol{k} \in \mathcal{A}^d$

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \log n \quad \text{pour } \beta \leq \alpha^*, \text{ égalité ssi } \boldsymbol{X} \text{ est déterministe}$$

Cas discret – cas périodique

Ψ_d défini sur $\mathcal{A} = \{0, \ldots, n-1\}^d$, $\boldsymbol{X}$ vecteur aléatoire associé

- TF "réduite" : $\widehat{\Psi}_d^{(r)}(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \sum_{\boldsymbol{u}\in\mathcal{A}} \Psi_d(\boldsymbol{u})\, e^{-\imath \boldsymbol{u}^t \boldsymbol{x}} \qquad \boldsymbol{x} \in [0; 2\pi)^d$

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \log(2\pi) \quad \text{pour } \beta \leq \alpha^*, \text{ égalité ssi } \boldsymbol{X} \text{ est déterministe}$$

- TF "discrète" : $\widehat{\Psi}_d^{(d)}(\boldsymbol{k}) = n^{-\frac{d}{2}} \sum_{\boldsymbol{u}\in\mathcal{A}} \Psi_d(\boldsymbol{u})\, e^{-\frac{2\imath\pi}{n} \boldsymbol{u}^t \boldsymbol{k}} \qquad \boldsymbol{k} \in \mathcal{A}^d$

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \log n \quad \text{pour } \beta \leq \alpha^*, \text{ égalité ssi } \boldsymbol{X} \text{ est déterministe}$$

Cas discret – cas périodique

Ψ_d défini sur $\mathcal{A} = \{0, \ldots, n-1\}^d$, $\boldsymbol{X}$ vecteur aléatoire associé

- TF "réduite" : $\widehat{\Psi}_d^{(r)}(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \sum_{\boldsymbol{u} \in \mathcal{A}} \Psi_d(\boldsymbol{u}) \, e^{-\imath \boldsymbol{u}^t \boldsymbol{x}} \qquad \boldsymbol{x} \in [0; 2\pi)^d$

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \log(2\pi) \quad \text{pour } \beta \leq \alpha^*, \text{ égalité ssi } \boldsymbol{X} \text{ est déterministe}$$

- TF "discrète" : $\widehat{\Psi}_d^{(d)}(\boldsymbol{k}) = n^{-\frac{d}{2}} \sum_{\boldsymbol{u} \in \mathcal{A}} \Psi_d(\boldsymbol{u}) e^{-\frac{2\imath\pi}{n} \boldsymbol{u}^t \boldsymbol{k}} \qquad \boldsymbol{k} \in \mathcal{A}^d$

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \log n \quad \text{pour } \beta \leq \alpha^*, \text{ égalité ssi } \boldsymbol{X} \text{ est déterministe}$$

Cas discret – cas périodique

Ψ_d défini sur $\mathcal{A} = \{0, \ldots, n-1\}^d$, $\boldsymbol{X}$ vecteur aléatoire associé

- TF "réduite" : $\widehat{\Psi}_d^{(r)}(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \sum_{\boldsymbol{u} \in \mathcal{A}} \Psi_d(\boldsymbol{u})\, e^{-\imath \boldsymbol{u}^t \boldsymbol{x}} \qquad \boldsymbol{x} \in [0; 2\pi)^d$

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \log(2\pi) \quad \text{pour } \beta \leq \alpha^*, \text{ égalité ssi } \boldsymbol{X} \text{ est déterministe}$$

- TF "discrète" : $\widehat{\Psi}_d^{(d)}(\boldsymbol{k}) = n^{-\frac{d}{2}} \sum_{\boldsymbol{u} \in \mathcal{A}} \Psi_d(\boldsymbol{u}) e^{-\frac{2\imath\pi}{n} \boldsymbol{u}^t \boldsymbol{k}} \qquad \boldsymbol{k} \in \mathcal{A}^d$

$$\frac{H_\alpha(\boldsymbol{X}) + H_\beta(\bar{\boldsymbol{X}})}{d} \geq \log n \quad \text{pour } \beta \leq \alpha^*, \text{ égalité ssi } \boldsymbol{X} \text{ est déterministe}$$

Basés sur l'inégalité de Young-Hausdorff $\|\widehat{\Psi}\|_{2\alpha^*} \leq \left(C^{\frac{1}{4\alpha^*} - \frac{1}{4\alpha}}\right)^d \|\Psi\|_{2\alpha}$
avec $\alpha \in \left[\frac{1}{2}\,;\,1\right]$ et $C = 2\pi$ (TFR) ou n (TFD)

Quelques références

T. M. Cover and J. A. Thomas.
Elements of Information Theory.
John Wiley & Sons, New-York, 1991.

A. Dembo, T. M. Cover, and J. A. Thomas.
Information theoretic inequalities.
IEEE Transactions on Information Theory, 37(6):1501–1518, November 1991.

A. Rényi.
On measures of entropy and information.
in Proceeding of the 4th Berkeley Symposium on Mathematical Statistics and Probability, 1:547–561, 1961.

C. E. Shannon.
A mathematical theory of communication.
The Bell System Technical Journal, 27:623–656, October 1948.