Théorie de l'information

'Fundamental limits in Information Theory'

Cours de Télécommunications

Thierry Sartenaer

Mars 2007

ECAM

HAUTE ECOLE • LEONARD DE VINCI •

Outline

Introduction

- La théorie de l'information, dont les bases ont été jetées par Claude Shannon en 1948, permet de calculer les limites fondamentales des performances atteintes par un système de communication.
- Objectif 1: communication efficace = comprimer au maximum la source d'information (codage de source)
- Objectif 2: communication fiable = protéger le système contre les erreurs dues au bruit (codage de canal)
- Résultats remarquables de la théorie de l'information:
  1. Jusqu'à quelle taille minimale peut-on comprimer un signal, sans perdre d'information? Réponse: la limite est donnée par l'entropie de la source du signal, définie en termes de comportement statistique de la source
  2. Quelle est la quantité maximale d'information qui peut être transmise sur un canal bruité? Réponse: la limite est donnée par la capacité du canal, définie par les caractéristiques statistiques du bruit de canal
  3. Si l'entropie de la source est inférieure à la capacité du canal, alors il est théoriquement possible de communiquer l'information produite par cette source avec un taux d'erreur arbitrairement faible à travers ce canal.

Outline

Incertitude, information et entropie

- Une source d'information discrète sans mémoire (discrete memoryless source) peut être modélisée comme une variable aléatoire discrète S produisant des symboles issus d'un alphabet $\mathcal{S} = \{s_0, s_1, \cdots, s_{K-1}\}$ avec des probabilités $0 \leq p_k \leq 1$ satisfaisant $\sum_{k=0}^{K-1} p_k = 1$. Les symboles successifs émis par la source sont supposés indépendants.
- La notion d'information produite par la source est liée à la notion d'incertitude. Avant la production d'un symbole particulier, l'observateur est dans l'incertitude. L'émission du symbole permet d'éliminer cette incertitude, et provoque un certain niveau de surprise, inversement proportionnel à la probabilité du symbole.
- L'information liée à un événement est d'autant plus élevée que la probabilité de cet événement est faible. Dans le cas limite, si un symbole est certain ($p_k = 1$) et tous les autres impossibles ($p_i = 0, \forall i \neq k$), alors il n'y a aucune 'surprise' ni 'information' lorsque la source produit s_k.

Incertitude, information et entropie

- Information (en bits) liée à l'événement $S = s_k$:

$$I(s_k) \triangleq \log_2\left(\frac{1}{p_k}\right)$$

- Cette définition répond aux propriétés intuitives que doit satisfaire la notion d'information:
  1. $I(s_k) \geq 0$ pour $0 \leq p_k \leq 1$,
  2. $I(s_k) = 0$ pour $p_k = 1$,
  3. $I(s_k) > I(s_i)$ pour $p_k < p_i$,
  4. $I(s_k, s_i) = I(s_k) + I(s_i)$ si s_k et s_i sont indépendants
- On a 1 bit quand $p_k = \frac{1}{2}$: le bit est donc défini comme la quantité d'information qui est gagnée lorsque l'on observe un événement particulier parmi deux événements équiprobables

Incertitude, information et entropie

- La quantité d'information produite par la source dépend du symbole particulier qui sera émis par cette source. L'information produite est donc elle-même une variable aléatoire qui peut prendre les valeurs $I(s_0), \cdots, I(s_{K-1})$ avec les probabilités $p_0, \cdots, p_{K-1}$, respectivement. L'entropie $H(\mathcal{S})$ de la source est définie comme la valeur moyenne de l'information produite par la source, en considérant tous les symboles possibles de l'alphabet:

$$H(\mathcal{S}) \triangleq E\left[I(s_k)\right] = \sum_{k=0}^{K-1} p_k \log_2 \frac{1}{p_k}$$

  L'entropie ne dépend que des probabilités p_k, et pas des valeurs s_k des symboles!
- Valeurs limites de l'entropie:

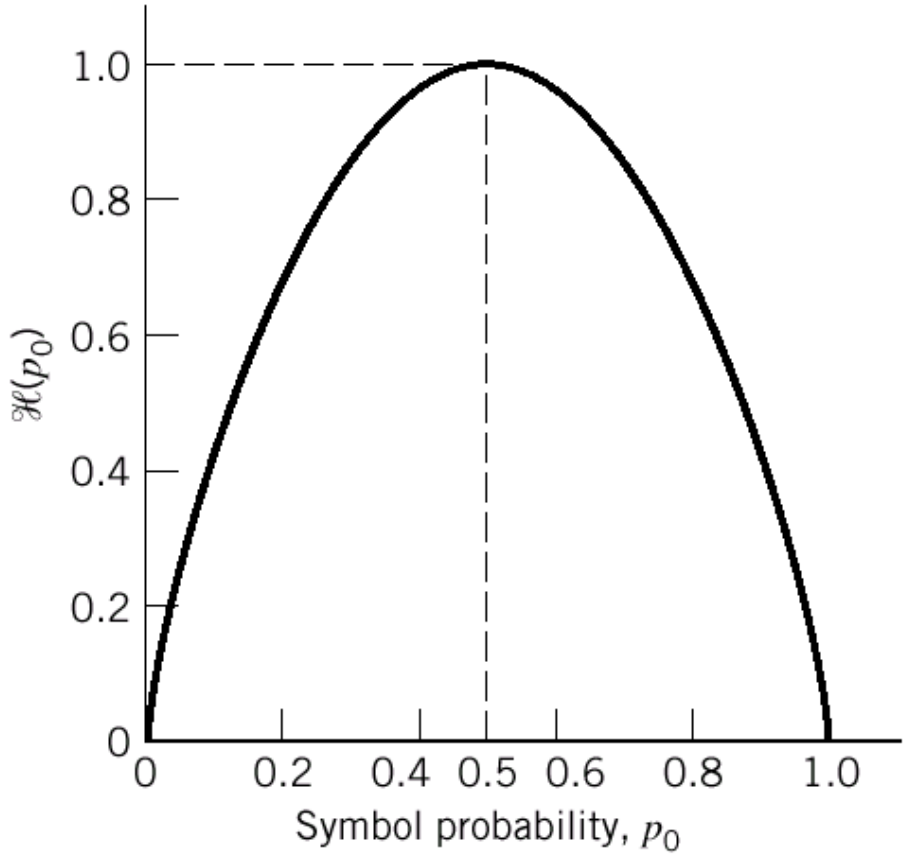
$$0 \leq H(\mathcal{S}) \leq \log_2(K)$$

- Entropie nulle: $H(\mathcal{S}) = 0$ si $p_k = 1$ et $p_i = 0$ pour $i \neq k$, soit une absence totale d'incertitude liée à la source
- Entropie maximale: $H(\mathcal{S}) = \log 2K$ si $p_k = 1/K$ pour tout k, soit une incertitude maximale quand tous les symboles sont équiprobables.

Incertitude, information et entropie

- Cas particulier: source binaire ($K = 2$) produisant des symboles '0' avec une probabilité p_0 et des symboles '1' avec une probabilité $p_1 = 1 - p_0$:

$$H(\mathcal{S}) = -p_0 \log_2(p_0) - (1 - p_0) \log_2(1 - p_0)$$

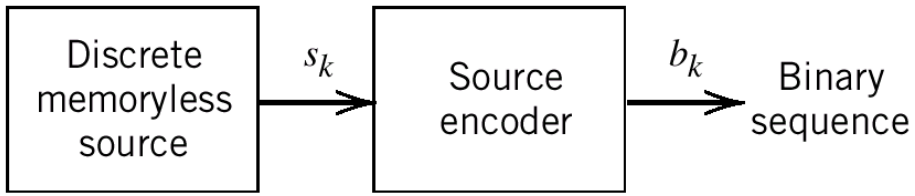


Outline

Codage de source

- Codage de source: comment représenter efficacement les données produites par une source discrète, en connaissant les propriétés statistiques de la source.
- Les mots codes produits par le codeur de source sont supposés être sous forme binaire, et le code doit être sans distorsion, ce qui signifie que la séquence de symboles originale doit pouvoir être reconstruite parfaitement à partir de la séquence binaire codée.
- Codes à longueur variable: code court pour les symboles fréquents, code long pour les symboles rares (exemple: Morse)
- Les symboles s_k de la source sont transformés en mots codes binaires b_k de longueur l_k. Le nombre moyen de bits par symbole associé au codeur de source est donc:

$$\overline{L} = \sum_{k=0}^{K-1} p_k l_k$$



Codage de source

- Premier théorème de Shannon: Source-coding theorem
- Soit une source discrète sans mémoire d'entropie $H(\mathcal{S})$, la longueur moyenne des mots codes $\overline{L}$ obtenus par un encodeur de source sans distorsion est bornée par:

$$\overline{L} \geq H(\mathcal{S})$$

- En pratique, un codeur de source donné n'atteindra pas cette limite et sera caractérisé par son efficacité de codage $\eta = \frac{H(\mathcal{S})}{\overline{L}} \leq 1$
- Compression de données = élimination de la redondance d'information contenue dans un signal numérique avant sa transmission
- Codes préfixes: aucun mot code n'est un préfixe d'un autre mot code
- Algorithmes les plus connus: codage de Huffman, codage de Lempel-Ziv (voir livre Haykin)

TABLE 9.2 *Illustrating the definition of a prefix code*

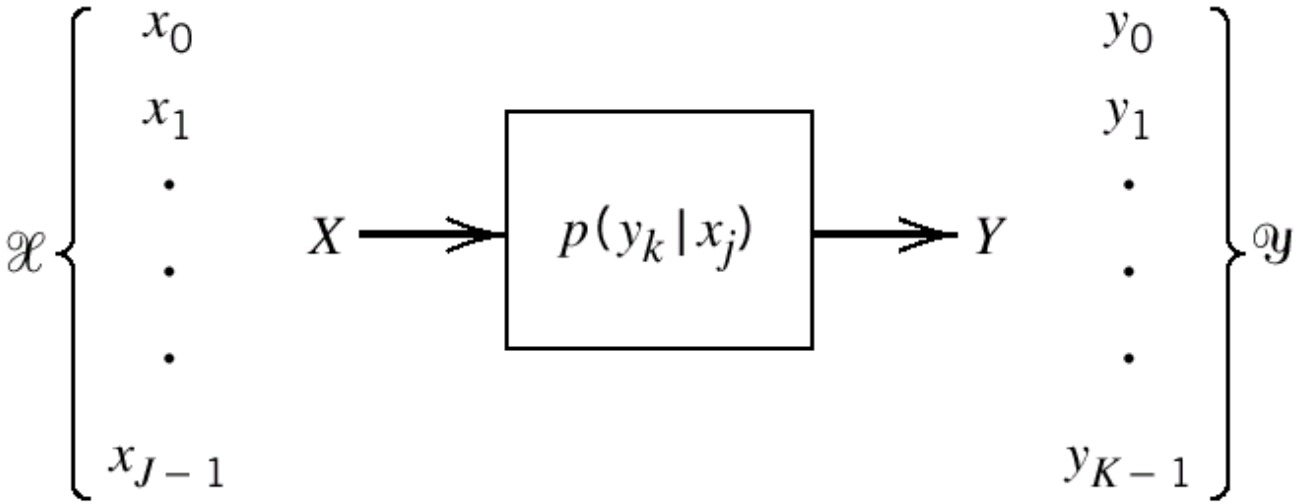
Source Symbol	*Probability of Occurrence*	*Code I*	*Code II*	*Code III*
s_0	0.5	0	0	0
s_1	0.25	1	10	01
s_2	0.125	00	110	011
s_3	0.125	11	111	0111

Outline

Discrete memoryless channel

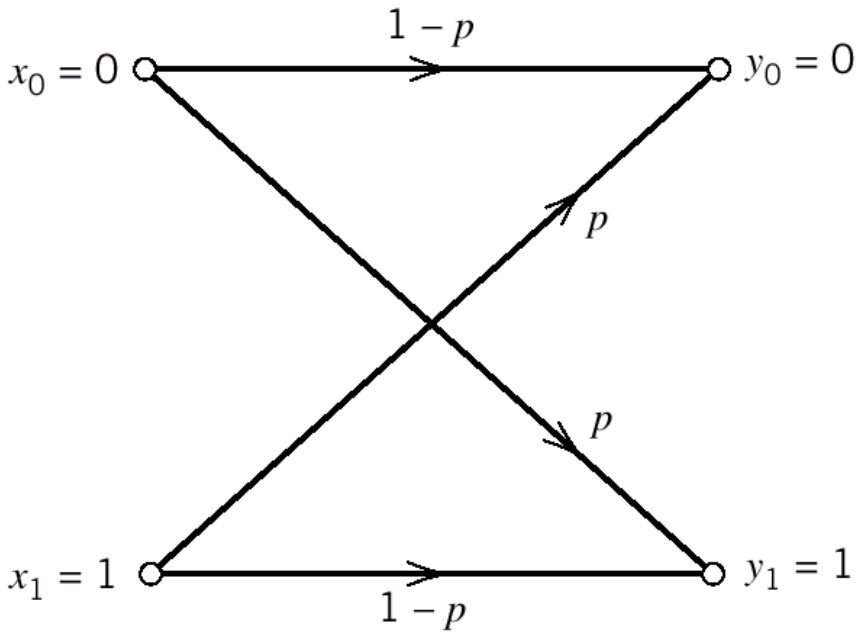
Après la source, le canal de transmission bruité peut aussi être modélisé comme un 'discrete memoryless channel':

- les symboles d'entrée sont issus d'un alphabet $\mathcal{X} = \{x0, \cdots, x_{J-1}\}$ de taille J
- les symboles de sortie sont issus d'un alphabet $\mathcal{Y} = \{y0, \cdots, y_{K-1}\}$ de taille K
- la sortie courante ne dépend que de l'entrée courante, et pas des entrées antérieures
- la matrice de transition caractérise la probabilité d'avoir chaque sortie possible pour chaque entrée possible



Discrete memoryless channel

- Cas particulier: 'Binary Symmetric Channel'
- $J = K = 2$, 2 symboles d'entrée ($x_0 = 0, x_1 = 1$), 2 symboles de sortie ($y_0 = 0, y_1 = 1$)
- Symétrie: $p_{10} = P(y = 1|x = 0) = p_{01} = P(y = 0|x = 1) = p$



Outline

Information mutuelle

- L'incertitude liée à X se mesure par l'entropie $H(\mathcal{X})$. La sortie Y du canal est une version bruitée de l'entrée X. Que reste-t-il donc comme incertitude liée à X après avoir observé Y?
- Pour une observation particulière $Y = y_k$, on peut définir l'entropie conditionnelle de X:

$$H(\mathcal{X}|Y = y_k) = \sum_{j=0}^{J-1} p(x_j|y_k) \log_2[1/p(x_j|y_k)]$$

- L'entropie conditionnelle $H(\mathcal{X}|\mathcal{Y})$ se calcule en moyennant cette quantité sur toutes les observations y_k possibles:

$$H(\mathcal{X}|\mathcal{Y}) = \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} p(x_j, y_k) \log_2 \left[\frac{1}{p(x_j|y_k)} \right]$$

- L'entropie conditionnelle représente la quantité d'incertitude qui reste à propos de l'entrée du canal après en avoir observé la sortie
- Comme l'entropie $H(\mathcal{X})$ représentait l'incertitude liée à X avant d'avoir observé Y, il s'ensuit que la différence $H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y})$ représente la part d'incertitude liée à X qui a été résolue grâce à l'observation de Y. Cette grandeur porte le nom d'information mutuelle du canal:

$$I(\mathcal{X};\mathcal{Y}) \triangleq H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y})$$

Information mutuelle

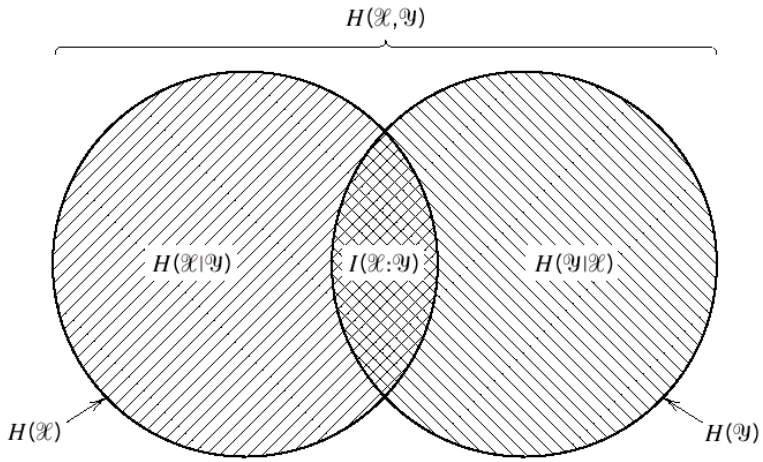
Propriétés de l'information mutuelle:

- Symétrique: $I(\mathcal{X};\mathcal{Y}) = I(\mathcal{Y};\mathcal{X})$
- Non-négative: $I(\mathcal{X};\mathcal{Y}) \geq 0$
- Lien entre l'information mutuelle et l'entropie conjointe des entrée et sortie du canal:

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

où l'entropie conjointe est définie par

$$H(\mathcal{X},\mathcal{Y}) = \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(x_j, y_k)\log_2[1/p(x_j, y_k)]$$



Outline

Capacité de canal

- Supposons que l'on connaît les alphabets d'entrée $\mathcal{X}$ et de sortie $\mathcal{Y}$, ainsi que la matrice de transition du canal $p(y_k|x_j)$. L'information mutuelle peut s'écrire comme:

$$I(\mathcal{X};\mathcal{Y}) = \sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j, y_k)\log_2\left[\frac{p(y_k|x_j)}{p(y_k)}\right]$$

  D'après cette expression, on peut déduire que la connaissance de la matrice de transition du canal n'est pas suffisante pour calculer l'information mutuelle: celle-ci dépend aussi de la distribution de probabilité de l'entrée $\{p(x_j)\}$.
- On définit la capacité du canal comme la valeur maximale de l'information mutuelle $I(\mathcal{X};\mathcal{Y})$ entre l'entrée et la sortie, la maximisation portant sur toutes les distributions de probabilité possibles $\{p(x_j)\}$ sur l'alphabet $\mathcal{X}$:

$$C \triangleq \max_{\{p(x_j)\}} I(\mathcal{X};\mathcal{Y})$$

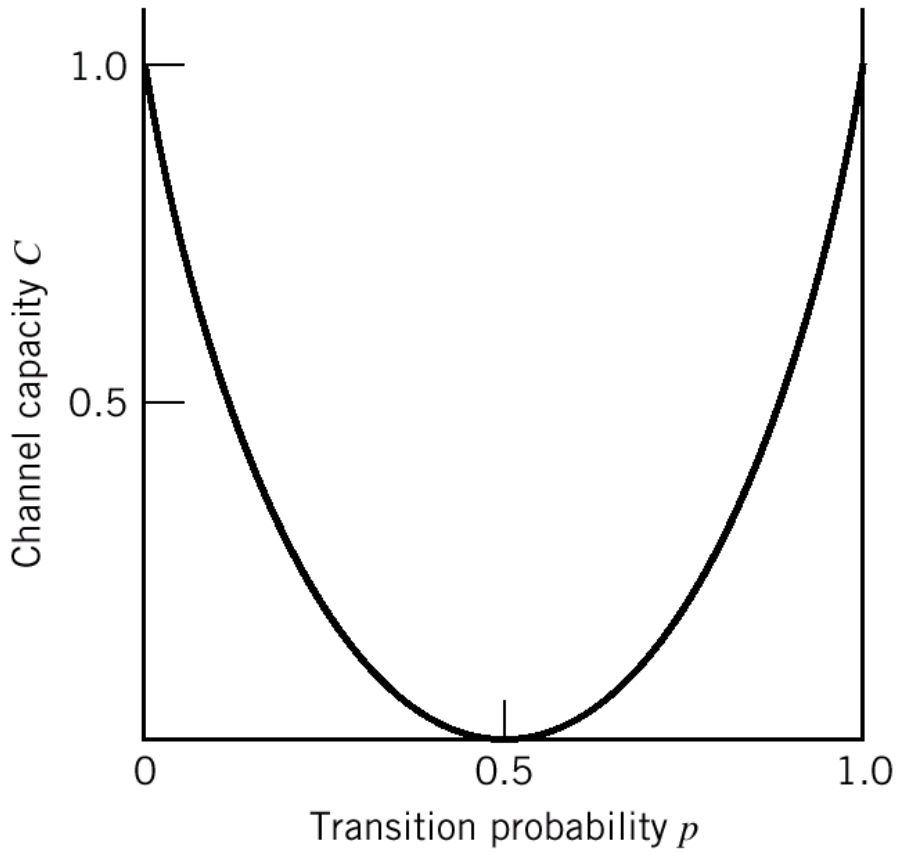
- Unités: bits par utilisation du canal
- Pour un alphabet d'entrée de taille J, le challenge est donc de trouver les J probabilités $p(x_j)$ (strictement positives, et de somme unité!) maximisant l'information mutuelle.

Capacité de canal

- Cas particulier du canal binaire symétrique
- On peut démontrer que l'information mutuelle est maximale pour des symboles d'entrée équiprobables: $p(x_0) = p(x_1) = 1/2$
- La capacité de canal se calcule alors simplement comme:

$$C = 1 + p \log_2 p + (1-p) \log_2 (1-p)$$

- Canal sans bruit ($p = 0$): $C = 1$ bit par utilisation du canal, correspondant à l'entropie de l'entrée
- Canal inutilisable ($p = 1/2$): $C = 0$



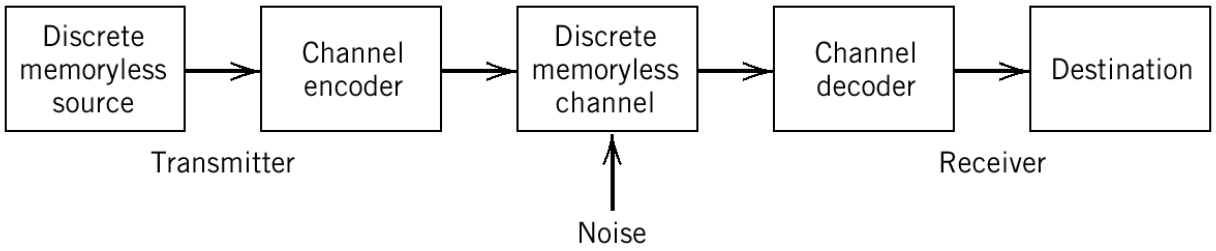
Outline

Codage de canal

- Codage de canal: comment protéger l'information transmise sur un canal bruité dont les propriétés statistiques sont connues, de manière à obtenir un taux d'erreur acceptable
- La protection contre les erreurs implique l'ajout d'une redondance contrôlée à l'information utile (opération duale du codage de source qui consistait à éliminer la redondance intrinsèque de la source)
- Par exemple, dans les codes *en blocs*, on ajoute $n - k$ bits de redondance à des messages de taille k, de manière à obtenir des mots codes de taille n. Le 'taux de codage' est défini par

$$r = \frac{k}{n} \leq 1$$

- Objectif du codage de canal: faire en sorte que le message de départ puisse être récupéré à partir du mot code reçu en sortie du canal bruité, avec une probabilité d'erreur aussi faible que possible



Codage de canal

- Question fondamentale: existe-t-il un système de codage de canal tel que la probabilité qu'un bit de message soit erronné en sortie soit arbitrairement faible (autrement dit: plus faible que ε pour n'importe quel ε positif), tout en restant efficace (taux de code r raisonnable)?
- La réponse est: OUI!
- Second théorème de Shannon: Channel-coding theorem
- Partie 1 ('direct theorem'): Soit une 'discrete memoryless source' d'entropie $H(\mathcal{S})$ produisant des symboles toutes les T_s secondes. Soit un 'discrete memoryless channel' de capacité C utilisé toutes les T_c secondes. Alors, si

$$\frac{H(\mathcal{S})}{T_s} \leq \frac{C}{T_c}$$

il existe un système de codage permettant de récupérer la source avec un taux d'erreur arbitrairement faible. Le rapport C/T_c est appelé taux critique.

Codage de canal

- Partie 2 ('converse theorem'): inversément, si

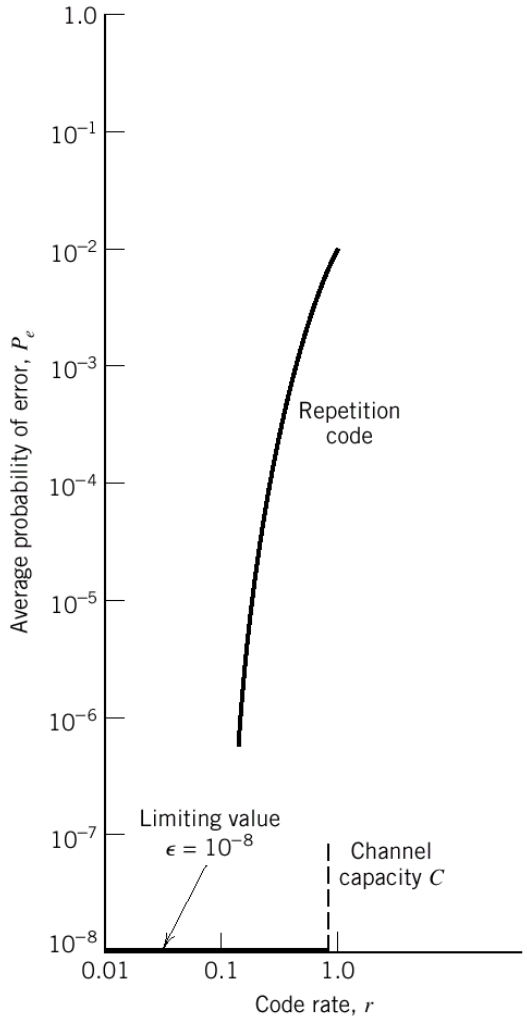
$$\frac{H(\mathcal{S})}{T_s} > \frac{C}{T_c},$$

  alors il n'est pas possible de transmettre l'information de la source sur le canal et de la récupérer avec un taux d'erreur arbitrairement faible.
- Le théorème du codage de canal ne fournit qu'une preuve d'existence de bons codes, mais ne précise absolument pas comment construire ces bons codes!
- Ce théorème ne précise pas non plus quelle sera la valeur pratique du taux d'erreur: il se contente de montrer que le taux d'erreur tend vers 0 quand la longueur n du code tend vers l'infini.

Codage de canal

Application au canal binaire symétrique:

- La source émet des symboles équiprobables ($H(\mathcal{S}) = 1$) chaque T_s seconde.
- Le canal est utilisé chaque T_c seconde, le taux de code vaut donc $r = T_c/T_s$.
- Pour être dans les conditions du théorème de Shannon, il faut donc satisfaire la relation $r \leq C$.
- Pour un canal dont la probabilité de transition est $p = 10^{-2}$, on a $C = 0.9192$. Donc il suffit de choisir un taux de code $r \leq 0.9192$ pour qu'il soit possible de concevoir un système de codage fournissant un taux d'erreur arbitrairement faible!
- En comparaison, l'utilisation d'un simple code à répétition implique le choix d'un taux de codage r très petit pour atteindre un taux d'erreur acceptable.



Outline

Extension aux variables aléatoires continues

- Dans le cas d'une variable aléatoire continue X, de densité de probabilité $f_X(x)$, on peut montrer que l'entropie $H(X)$ (définie à la base pour des variables aléatoires discrètes!) est toujours infinie. On s'intéressera plutôt à la notion d'entropie différentielle $h(X)$:

$$h(X) \triangleq \int_{-\infty}^{\infty} f_X(x) \log_2 \left[\frac{1}{f_X(x)} dx \right]$$

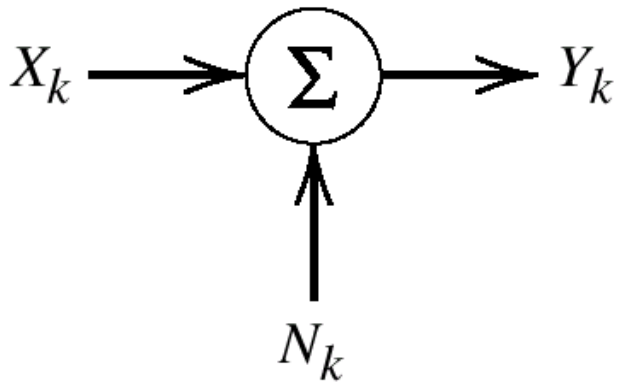
- L'entropie différentielle prend des valeurs finies... par contre, elle peut être négative!
- Distribution gaussienne (moyenne μ, variance σ):
  - Entropie différentielle: $h(X) = \frac{1}{2} \log_2(2\pi e \sigma^2)$
  - Dépend uniquement de σ^2, pas de μ
  - Gaussienne = $h(X)$ plus élevée que n'importe quelle autre distribution de même variance!
- Information mutuelle: $I(X;Y) = h(X) - h(X|Y)$, satisfaisant aux mêmes propriétés que l'information mutuelle définie pour les variables aléatoires discrètes

Capacité d'un canal gaussien

- Soit un canal gaussien de bande passante limitée à B et de puissance limitée à P
- La source $X(t)$, de bande passante B, peut être échantillonnée à cadence $2B$ pour fournir la séquence d' échantillons $(X_1, X_2, \cdots, X_K)$ pendant une durée T avec $K = 2BT$
- Le signal transmis sur le canal est perturbé par un bruit AWGN de densité spectrale de puissance $N_0/2$ dont la bande passante est aussi limitée à B
- Les échantillons du signal reçu s'écrivent:

$$Y_k = X_k + N_k$$

- Variance des échantillons de bruit N_k: $\sigma^2 = N_0 B$
- Les échantillons du signal reçu Y_k sont statistiquement indépendants
- Contrainte sur la puissance du signal transmis: $E[X_k^2] = P$ pour tout k



Capacité d'un canal gaussien

- La capacité du canal gaussien de bande passante B et de puissance P est définie comme:

$$C = \max_{f_{X_k}(x)} \left\{ I(X_k; Y_k) : E[X_k^2] = P \right\}$$

avec l'information mutuelle donnée par

$$I(X_k; Y_k) = h(Y_k) - h(Y_k|X_k) = h(Y_k) - h(N_k)$$

- Le maximum de $I(X_k; Y_k)$ sera donc obtenu en choisissant la distribution $f_{X_k}(x)$ qui maximise $h(Y_k)$ tout en respectant la contrainte de puissance P, c'est à dire la distribution gaussienne!
- Le signal reçu est donc gaussien de variance $P + \sigma^2$
- L'entropie différentielle du signal reçu Y_k et du bruit N_k sont donc:

$$h(Y_k) = \frac{1}{2} \log_2[2\pi e(P + \sigma^2)]$$

$$h(N_k) = \frac{1}{2} \log_2(2\pi e\sigma^2)$$

- Capacité du canal gaussien:

$$C = \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma^2}\right)$$

Unités: bits par utilisation du canal

Capacité d'un canal gaussien

- Si on utilise le canal K fois (transmission des K échantillons X_k) sur une durée de Tsecondes, la capacité par unité de temps devient:

$$C = \frac{K}{T}\frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) = B\log_2\left(1+\frac{P}{N_0 B}\right)$$

  Unité: bits par seconde
- Il s'agit là de l'un des plus remarquables résultats de la théorie de l'information! En une simple formule, on peut comprendre les effets respectifs des 3 caractéristiques principales du système sur la capacité du canal: la bande passante B, la puissance moyenne du signal P, et la densité spectrale de puissance du signal reçu N_0
- La capacité du canal dépend de la bande passante B de manière linéaire, alors qu'elle ne dépend du rapport signal à bruit P/N_0B que de manière logarithmique. Pour une variance de bruit donnée, il est donc plus facile d'augmenter la capacité du canal en augmentant la bande passante qu'en augmentant la puissance de transmission!
- Limite fondamentale d'un système de communication ('channel coding theorem'): si l'on transmet l'information à un débit binaire inférieur ou égal à C bits/s, il est théoriquement possible d'atteindre un taux d'erreur arbitrairement faible. Pour atteindre cette limite, le signal transmis devra avoir des propriétés statistiques semblables à celles d'un bruit gaussien.

Capacité d'un canal gaussien

- Supposons un système idéal où les données sont transmises à un débit binaire R_b égal à la capacité C du canal. La puissance transmise vaut donc $P = E_b C$ où E_b représente l'énergie transmise par bit.
- Le système idéal est défini par la relation suivante:

$$\frac{C}{B} = \log_2\left(1 + \frac{E_b}{N_0}\frac{C}{B}\right)$$

ou encore

$$\frac{E_b}{N_0} = \frac{2^{C/B} - 1}{C/B}$$

- Limite de Shannon (bande passante infinie):

$$\lim_{B\to\infty}\left(\frac{E_b}{N_0}\right) = \log 2 = 0.693$$