Université Paris-Sud 11
Faculté des sciences d'Orsay
Master en Information, systèmes et technologie
M2P Réseaux et télécommunications
Année 2007-2008
—

Séance n^o 1
Mardi 11 septembre 2007
Durée : 3 heures
—

Travaux pratiques de OUTILS MATHÉMATIQUES POUR LE TRAITEMENT DU SIGNAL

Menu du jour : Initiation à MATLAB et utilisation des probabilités discrètes

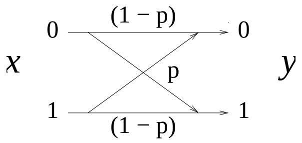
L'objectif de cette séance de travaux pratiques est double :
— d'une part se familiariser avec le langage MATLAB qui est très répandu en traitement du signal, aussi bien dans le monde industriel que le monde académique (voir le manuel résumé « MATLAB en bref ») ; et
— d'autre part s'exercer à utiliser les premières notions probabilistes vues en cours : application des règles de base en vue du calcul de distributions de probabilité. L'exemple choisi est celui du codage de canal lors de la transmission d'un fichier binaire, problème qui reviendra dans les enseignements de communications numériques de la spécialité.

1.1 — Comment peut-on obtenir une communication parfaite avec un canal de communication imparfait, bruyant ?

Une ligne de téléphone analogique sur laquelle communiquent deux modems numériques, une liaison radio avec un satellite d'exploration du système solaire, ou un disque dur d'ordinateur[1] sont des exemples de canaux de communication bruyants. La ligne téléphonique souffre d'intermodulation avec d'autres lignes. L'espace entre les stations radio terrestres et le satellite reçoit divers rayonnements électromagnétiques provenant d'autres sources terrestres ou cosmiques. Un disque dur, qui écrit et lit des chiffres binaires (un zéro ou un un, ou *bit* — pour *binary digit* — à ne pas confondre avec le *bit* — pour *binary unit* — unité d'information binaire) en alignant un volume de matériau magnétique dans l'une de deux orientations possibles, peut très bien échouer à relire plus tard le bit stocké : la magnétisation du volume de matériau a pu basculer spontanément, ou bien une bouffée de bruit ambiant peut faire que le circuit de lecture transmet la mauvaise valeur, ou bien encore la tête d'écriture n'a pas recréé la magnétisation à la place initiale à cause des interférences avec les bits voisins. Dans tous les cas, si nous transmettons les données, c'est-à-dire une suite de bits, à travers le canal, il y a une certaine **probabilité** que le message reçu ne soit pas identique au message transmis.

Considérons un disque dur bruyant qui transmet chaque bit correctement avec la probabilité $(1-p)$ et incorrectement avec la probabilité p. Ce modèle de canal de communication est connu sous le nom de *canal binaire symétrique* (voir figure 1).

[1] Lorsqu'un logiciel d'application a besoin d'un fichier stocké sur le disque dur, la source est le disque magnétique proprement dit, le destinataire est la mémoire vive de l'ordinateur, et le canal est l'ensemble du dispositif qui les relie, depuis la tête de lecture jusqu'aux circuits qui pilotent la mémoire. Les éventuels problèmes de transmission se situant ordinairement au niveau du dispositif de lecture — ou d'écriture dans l'opération inverse — nous dirons, par abus de langage, que le disque dur est son propre canal de transmission.



$$P(y=0\,|\,x=0)=1-p \qquad \text{et} \qquad P(y=0\,|\,x=1)=p$$

$$P(y=1\,|\,x=0)=p \qquad \text{et} \qquad P(y=1\,|\,x=1)=1-p$$

Figure 1 – *Le canal binaire symétrique. Le symbole transmis est x et le symbole reçu y. Le niveau de bruit, ou probabilité qu'un bit soit incorrectement transmis, est p.*

À titre d'exemple, imaginons que $p = 0,1$, c'est-à-dire que 10 % des bits transmis sont modifiés (voir figure 2)[2]. Pour être utilisable, un disque dur ne devrait normalement effectuer aucune erreur de transmission pendant toute la durée de son utilisation. Si nous prévoyons de lire et d'écrire un volume de données d'un milliard de mots (soit un *gigabyte*) par jour pendant 10 ans, il nous faut une probabilité d'erreur de l'ordre de 10^{-15}, ou moins. Il y a deux moyens d'y parvenir.

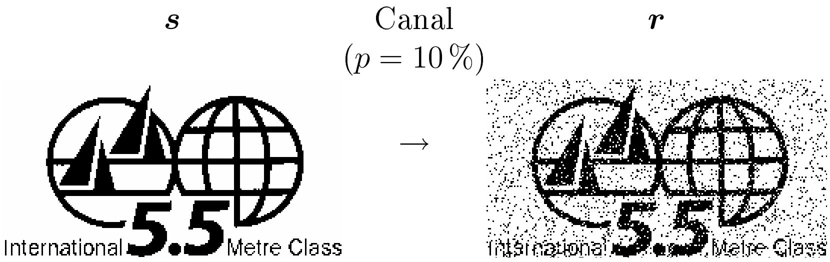


Figure 2 – *Exemple de transmission d'une image binaire dans un canal binaire symétrique ayant une probabilité d'erreur $p = 0,1$. À gauche, le message $\boldsymbol{s}$ émis par la source ; à droite, le message $\boldsymbol{r}$ reçu.*

La solution « physique »

La solution « physique » consisterait à améliorer les caractéristiques physiques du canal pour réduire sa probabilité d'erreur. On pourrait y parvenir en :
— utilisant des composants électroniques plus fiables ;
— évacuant l'air du boitier du disque pour éliminer les turbulences qui perturbent la trajectoire de la tête de lecture ;
— utilisant une plus grande surface de matériau magnétique pour représenter chaque bit ;
— utilisant des signaux de plus grande amplitude ou en refroidissant les circuits électroniques pour réduire le bruit thermique.

Mais ces modifications, pour autant qu'il soit décidé de les faire — la tendance actuelle est plutôt d'augmenter la densité des enregistrements sur le support magnétique, par exemple — augmenteraient certainement le coût du canal.

La solution « système »

La théorie de l'information et celle du codage offrent une alternative : accepter le canal tel qu'il est et lui ajouter des *systèmes* de communication tels que l'on puisse détecter et corriger les erreurs introduites par le canal. Comme le montre la figure 3, nous pouvons ajouter un *codeur* devant le canal et un *décodeur* derrière. Le codeur transforme le message $\boldsymbol{s}$ de la source en un message *transmis* $\boldsymbol{t}$, en ajoutant de la *redondance* au message original.

[2] Le lecteur intrigué par cette image peut se rendre sur le site : http ://www.5.5class.org/ et il comprendra que la torture des étudiants par l'enseignement d'une matière qui leur « *prend la tête* » n'est pas ma seule passion dans la vie.

Le canal ajoute du bruit au message transmis et délivre en sortie le message *reçu* $\boldsymbol{r}$. Le décodeur utilise la redondance introduite par le codeur, et qui est connue, pour en déduire à la fois le signal original $\boldsymbol{s}$ et le bruit qui a été ajouté.

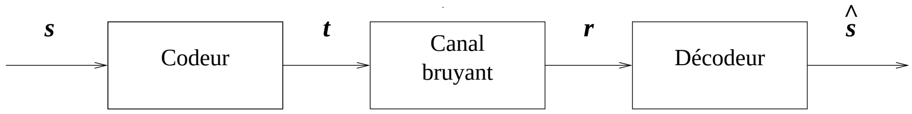


Figure 3 – *La solution « système » pour obtenir une communication fiable avec un canal bruyant.*

Par rapport aux solutions « physiques », la solution « système » n'introduit de surcoût qu'au niveau des charges calculatoires du codeur et du décodeur. La *théorie de l'information* permet d'étudier les limites théoriques de tels systèmes alors que la *théorie du codage* porte sur la réalisation pratique de systèmes de codage et de décodage.

1.2 — Codes correcteurs d'erreur pour le canal binaire symétrique

Quel est le moyen le plus simple et le plus efficace d'ajouter de la redondance à une transmission, de manière à être capable de détecter **et** de corriger les erreurs ? Un exemple très simple donne une idée correcte des techniques de codage utilisées dans ce domaine (codage de canal).

Codes à répétition

Une idée immédiate est de répéter chaque bit du message un nombre de fois prédéfini, par exemple trois, comme dans la figure suivante. On obtient ainsi le *code à répétition* appelé R_3.

Mot de la source $\boldsymbol{s}$	Mot transmis $\boldsymbol{t}$
0	000
1	111

Figure 4 – *Le code à répétition R_3.*

Imaginons que nous voulions transmettre le message suivant : $\boldsymbol{s} = 0010110$ à travers un canal binaire symétrique ayant un niveau de bruit $p = 0,1$ en utilisant ce code à répétition. On peut décrire le fonctionnement de ce canal comme l'« addition » d'un vecteur-bruit épars[3] $\boldsymbol{b}$ au vecteur transmis $\boldsymbol{t}$, avec bien entendu une addition en arithmétique modulo 2, comme le montre la figure suivante.

$$
\begin{array}{c|ccccccc}
\boldsymbol{s} & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
\boldsymbol{t} & \overbrace{000} & \overbrace{000} & \overbrace{111} & \overbrace{000} & \overbrace{111} & \overbrace{111} & \overbrace{000} \\
\boldsymbol{b} & 000 & 001 & 000 & 000 & 101 & 000 & 000 \\
\hline
\boldsymbol{r} & 000 & 001 & 111 & 000 & 010 & 111 & 000
\end{array}
$$

Figure 5 – *Un exemple de transmission utilisant R_3.*

[3] Épars car le nombre de un dans ce vecteur, qui vaut p en moyenne, est faible.

Comment devons-nous décoder le vecteur reçu $\boldsymbol{r}$? L'algorithme optimal considère les bits reçus trois par trois et prend une décision à la majorité.

Au risque de paraître vouloir expliquer l'évidence, il est important de démontrer ce résultat, ne serait-ce que pour dévoiler les hypothèses qui lui sont implicites. La décision de décodage optimale (au sens où elle a la probabilité la plus faible d'être fausse) consiste à trouver la valeur de $\boldsymbol{s}$ qui est la plus probable, sachant $\boldsymbol{r}$. Considérons le décodage d'un unique bit s, qui a été codé par $\boldsymbol{t}(s)$ et qui a engendré les trois bits reçus $\boldsymbol{r} = r_1r_2r_3$. Le théorème de Bayes implique que la probabilité *a posteriori* pour s est :

$$P(s \mid r_1r_2r_3) = P(s)\, P(r_1r_2r_3 \mid s)/P(r_1r_2r_3). \tag{1}$$

Nous pouvons expliciter cette probabilité pour les deux termes de l'alternative :

$$P(s=1 \mid r_1r_2r_3) = P(s=1)\, P(r_1r_2r_3 \mid s=1)/P(r_1r_2r_3). \tag{2}$$

$$P(s=0 \mid r_1r_2r_3) = P(s=0)\, P(r_1r_2r_3 \mid s=0)/P(r_1r_2r_3). \tag{3}$$

Cette probabilité *a posteriori* est déterminée par deux facteurs : la ***probabilité a priori*** $P(s)$ et le terme dépendant des données reçues $P(r_1r_2r_3 \mid s)$, qui est appelé la *vraisemblance* de s. La constante $P(r_1r_2r_3)$ n'a pas besoin d'être calculée quand on cherche la décision de décodage optimale qui consiste à supposer que $\widehat{s}=0$ si $P(s=0 \mid \boldsymbol{r}) > P(s=1 \mid \boldsymbol{r})$, et $\widehat{s}=1$ sinon.

Pour trouver $P(s=0 \mid \boldsymbol{r})$ et $P(s=1 \mid \boldsymbol{r})$, nous devons faire une hypothèse sur les probabilités *a priori* des deux hypothèses $s=0$ et $s=1$, et nous devons faire une hypothèse sur la probabilité pour $\boldsymbol{r}$ sachant s. Nous supposons que les probabilités *a priori* sont égales : $P(s=0) = P(s=1) = 0,5$; et donc qu'il est équivalent de maximiser la probabilité *a posteriori* $P(s \mid \boldsymbol{r})$ ou de maximiser la vraisemblance $P(\boldsymbol{r} \mid s)$. Nous supposons aussi que le canal est binaire symétrique avec un niveau de bruit $p < 0,5$ et que les erreurs de transmission sont indépendantes les unes des autres, si bien que la vraisemblance est :

$$P(\boldsymbol{r} \mid s) = P(\boldsymbol{r} \mid \boldsymbol{t}(s)) = \prod_{n=1}^{N} P(r_n \mid t_n(s)), \tag{4}$$

où $N=3$ est le nombre de bits transmis dans le bloc considéré, et :

$$P(r_n \mid t_n) = \begin{cases} (1-p) & \text{si} \quad r_n = t_n \\ p & \text{si} \quad r_n \neq t_n. \end{cases}$$

Le *rapport de vraisemblance* pour les deux hypothèses est :

$$\frac{P(\boldsymbol{r} \mid s=1)}{P(\boldsymbol{r} \mid s=0)} = \prod_{n=1}^{N} \frac{P(r_n \mid t_n(1))}{P(r_n \mid t_n(0))}\,; \tag{5}$$

chaque facteur du second membre valant $(1-p)/p$ si $r_n = 1$ et $p/(1-p)$ si $r_n = 0$. Le rapport $\gamma = (1-p)/p$ est plus grand que 1 puisque $p < 0,5$, si bien que l'hypothèse gagnante est celle qui a recueilli le plus de « votes », chaque vote intervenant avec un facteur de γ dans le rapport de vraisemblance.

Le décodeur par « vote à la majorité » est donc optimal si nous supposons que le canal est binaire symétrique et que les deux mots possibles 0 et 1 émis par la source ont une égale probabilité *a priori*.

Appliquons ce décodage au vecteur reçu de la figure 5. Les trois premiers bits reçus sont tous des 0, nous décodons ce triplet comme un 0. Dans le deuxième triplet, nous avons deux 0 et un 1, et nous décodons donc ce triplet comme un 0 — ce qui corrige l'erreur dans ce cas. Mais toutes les erreurs ne sont pas corrigées, cependant. Si par malchance deux erreurs surviennent dans un seul bloc, comme dans le 5e triplet de la figure 5, alors la règle de décodage donne la mauvaise réponse, comme le montre la figure 6.

$$
\begin{array}{rccccccc}
\boldsymbol{s} & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
\boldsymbol{t} & \overbrace{000} & \overbrace{000} & \overbrace{111} & \overbrace{000} & \overbrace{111} & \overbrace{111} & \overbrace{000} \\
\boldsymbol{b} & 000 & 001 & 000 & 000 & 101 & 000 & 000 \\
\hline
\boldsymbol{r} & \underbrace{000} & \underbrace{001} & \underbrace{111} & \underbrace{000} & \underbrace{010} & \underbrace{111} & \underbrace{000} \\
\widehat{\boldsymbol{s}} & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
\text{Erreurs corrigées} & & * & & & & & \\
\text{Erreurs non-détectées} & & & & & * & &
\end{array}
$$

Figure 6 – *Décodage du vecteur reçu à la figure 5*

Nous voyons donc que **le plus probable n'est pas nécessairement vrai**. La probabilité d'erreur est dominée par la probabilité que deux bits dans un bloc de trois aient été modifiés, et qui vaut p^2 (voir l'exercice). Dans l'exemple du canal binaire symétrique avec $p = 0,1$, le code R_3 a une probabilité d'erreur, après décodage, d'environ $p_b = 0,03$ par bit. La figure 7 montre le résultat de la transmission d'une image binaire par un canal binaire symétrique en utilisant ce code à répétition.

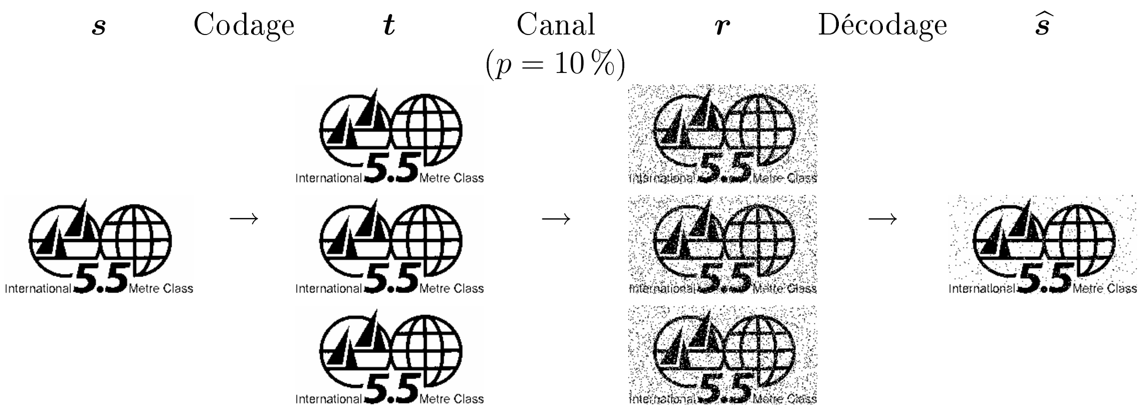


Figure 7 – *Transmission d'une image de 27.000 bits dans un canal binaire symétrique ayant une probabilité d'erreur $p = 0,1$ en utilisant le codage à répétition R_3 et le décodage par vote à la majorité. La probabilité d'erreur passe de 10 % à environ 3 % mais le taux de transmission est réduit à 1/3.*

Le code à répétition R_3 a donc réduit la probabilité d'erreur, comme voulu. Nous avons cependant perdu quelque chose : notre *taux de transfert d'information* a été réduit d'un facteur de trois. Si nous utilisons ce code à répétition pour transmettre des données par une ligne téléphonique, il réduira la fréquence des erreurs, mais il réduira aussi notre vitesse de transmission. Nous devrons payer trois fois plus pour chaque appel téléphonique. De même, comme il faut bien stocker les données répétées quelque part, il nous faudrait disposer de trois disques durs bruyants d'une capacité d'un *gigabyte* pour créer un disque dur équivalent d'un *gigabyte* avec $p_b = 0,03$.

Pouvons-nous encore réduire la probabilité d'erreur jusqu'à la valeur de 10^{-15} exigée pour un disque dur commercialisable ? Nous pourrions utiliser pour cela un code à répétition avec plus de répétitions. Mais on montre (voir l'exercice) qu'il faudrait alors disposer de 60 disques durs bruyants pour créer un disque dur équivalent avec la fiabilité requise. Le compromis entre la probabilité d'erreur et le taux de transmission est indiqué figure 8.

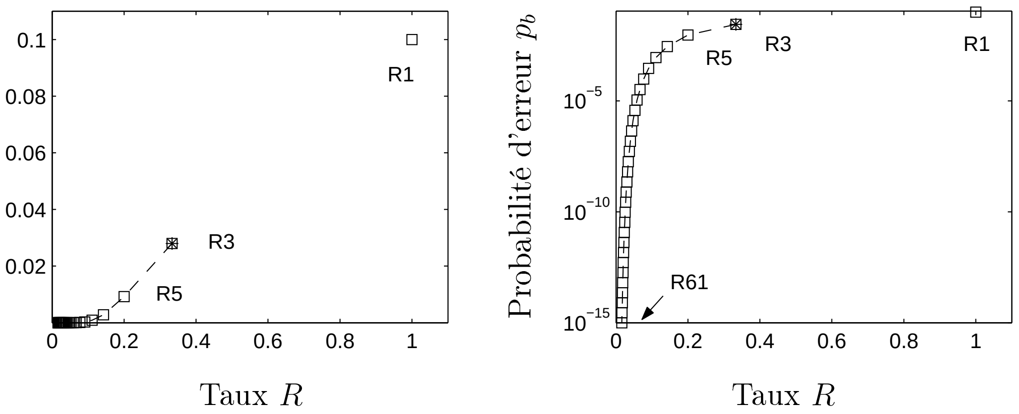


Figure 8 – *Probabilité d'erreur p_b en fonction du taux de transmission pour les codes à répétition et un canal binaire symétrique avec $p = 0,1$. Échelle linéaire à gauche, et logarithmique à droite. Le code R_3 est repéré par une étoile. On aimerait pouvoir se situer en bas à droite, avec un taux élevé et une probabilité d'erreur faible.*

Quelles performances peuvent atteindre les meilleurs codes ?

La figure 9 montre les résultats obtenus avec les codes à répétition et d'autres codes plus efficaces : le code de Hamming et les codes BCH qui sont des codes linéaires par bloc généralisant les codes de Hamming. Cette figure montre que nous pouvons, avec des codes linéaires par bloc, obtenir de meilleurs résultats qu'avec les codes à répétition ; mais la situation asymptotique — lorsque l'on veut faire tendre p_b vers zéro — reste déprimante.

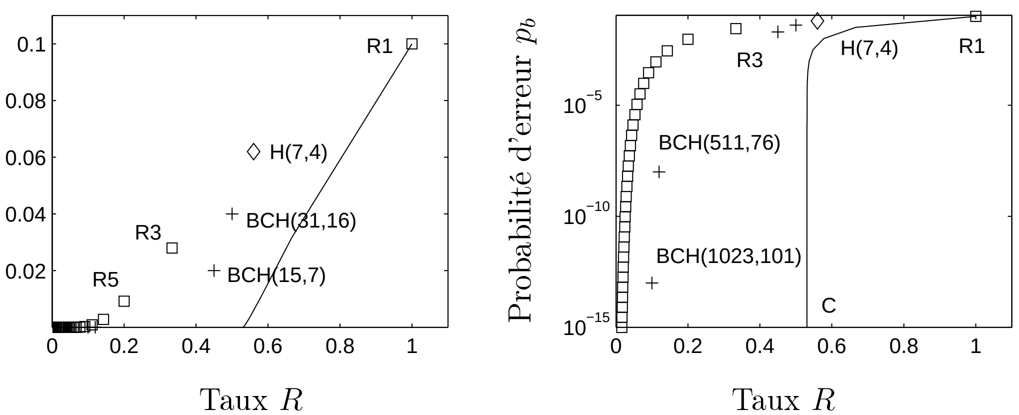


Figure 9 – *Probabilité d'erreur p_b en fonction du taux de transmission R pour les codes à répétition (représentés par des carrés), le (7,4)-code de Hamming (représenté par un losange) et des codes BCH avec des longueurs de bloc allant jusqu'à 1023 (représentés par des croix), pour un canal binaire symétrique avec $p = 0,1$. La courbe en trait plein représente la limite de Shannon. Les points situés à droite ne sont pas atteignables.*

Il semble qu'il doive exister un compromis entre la probabilité d'erreur de décodage par bit p_b (que nous voudrions abaisser) et le taux de transmission R (que nous voudrions maintenir à une valeur élevée). Comment caractériser ce compromis ? Quels points du plan (R, p_b) sont atteignables ? Cette question a été résolue par Shannon dans son article fondateur de 1948, qui a tout à la fois créé la théorie de l'information et résolu la plupart de ses questions fondamentales.

Une opinion répandue à l'époque était que la frontière entre les points atteignables et non atteignables dans le plan (R, p_b) était une courbe passant par l'origine ; s'il en était ainsi, il fallait, pour obtenir une probabilité d'erreur p_b aussi petite que l'on veut, accepter de faire tendre le taux R vers zéro : « *On n'a rien sans peine.* » Shannon démontra cependant le résultat remarquable que cette frontière coupe l'axe des R au point où $R = C$, C étant la *capacité* du canal, comme indiqué figure 9. Pour tout canal, il existe des codes qui permettent de communiquer avec une probabilité d'erreur p_b aussi petite que l'on veut, à des taux non nuls.

La capacité du canal binaire symétrique de notre exemple du disque dur, avec le niveau de bruit $p = 0,1$, vaut $C \simeq 0,53$. Voyons ce que cela signifie en termes de probabilité d'erreur de transmission. Nous savons qu'avec le code à répétition R_3 nous pouvons atteindre $p_b \simeq 0,03$ avec un taux $R = 1/3$. Nous savons aussi (voir l'exercice) comment obtenir $p_b \simeq 10^{-15}$ avec soixante disques durs. Et voici Shannon qui débarque, s'enquiert de ce que nous faisons, et déclare : « Quelles performances voulez-vous atteindre ? 10^{-15} ? Vous n'avez pas besoin de soixante disques durs, il vous suffit juste de deux disques durs (car 1/2 est plus petit que 0,53). Votre taux vaudra donc 1/2. Et s'il vous faut 10^{-18} ou 10^{-24} ou encore mieux, vous pouvez aussi l'obtenir avec deux disques durs ![4] » Étonnant, non ?

1.3 — Travail à effectuer

Exercice

(a) — Montrer que la probabilité d'erreur est réduite par l'utilisation du code à répétition R_3 en calculant la probabilité d'erreur de ce code pour un canal binaire symétrique avec un niveau de bruit p.

(b) — Montrer que la probabilité d'erreur de R_N, le code à répétition avec N répétitions, est :

$$p_b = \sum_{n=(N+1)/2}^{N} \mathrm{C}_n^N \, p^n \, (1-p)^{N-n} \tag{6}$$

pour N impair.

(c) — En supposant que $p = 0,1$, quel est le plus grand terme de la somme ?

(d) — En utilisant la formule de Stirling pour approcher le facteur C_n^N dans ce terme, trouver une expression approchée de la probabilité d'erreur d'un code à répétition avec N répétitions.

(e) — En supposant que $p = 0,1$, combien faut-il de répétitions pour abaisser la probabilité d'erreur à 10^{-15} ?

Simulation numérique sous MATLAB

Le but de cette partie est de simuler la transmission d'un fichier binaire dans un canal binaire symétrique et bruyant, en utilisant le code à répétition R3, de vérifier que celui-ci permet effectivement de réduire la probabilité d'erreur par bit, conformément au calcul effectué dans l'exercice ci-dessus.

Il est vivement conseillé de créer, sous l'éditeur, un fichier exécutable par MATLAB (que l'on appellera, par exemple, `tp1.m`), plutôt que de taper des instructions à suivre dans la fenêtre de commande. Ceci permet d'éviter d'avoir à les retaper à chaque nouvelle exécution et facilite considérablement la mise au point. Pour l'exécution, il suffit de taper le nom du fichier, sans le suffixe `.m`, dans la fenêtre de commande.

[4] Ces affirmations pourraient ne pas être strictement vraies puisque le théorème de Shannon implique des codes par bloc avec des longueurs de bloc toujours croissantes, et que la longueur de bloc requise pourrait être supérieure à un *gigabyte*. Mais il suffirait alors d'utiliser des disques de plus grande capacité.

On commencera par charger le fichier fourni (`5point5.jpg`) qui est une image en couleur au format J-PEG. Il faudra ensuite la convertir en niveaux de gris réels (MATLAB ne connaît que les réels et les chaînes de caractères), puis la binariser avec les deux niveaux 0 et 1. On la visualisera ensuite à l'écran pour se familiariser avec les instructions graphiques.

On engendrera ensuite un bruit blanc binaire, à valeurs 0 ou 1, de probabilités respectives $1 - p$ et p, à partir du générateur de bruit pseudo-aléatoire uniforme. On choisira $p = 0,1$.

Trois réalisations différentes de ce bruit seront ensuite ajoutées, modulo 2, à l'image à transmettre pour simuler le fonctionnement du canal bruyant avec le code R3.

Un vote à la majorité sur les trois images transmises permettra d'obtenir le résultat du décodage optimal associé au code R3. On s'assurera que le nombre moyen d'erreurs subsistant, qui est un estimateur de la probabilité p_b calculée dans l'exercice ci-dessus, n'en est pas trop éloigné.

—oOo—