Examen du cours de Théorie de l'Information et Codage

Temps imparti: 3h.

Dans tout le sujet, on note $H\colon x \mapsto -x\log_2(x) - (1-x)\log_2(1-x)$.

1. Problème 1: Vous avez 8 bouteilles de vin dont une exactement est empoisonnée avec les probabilités suivantes: $p_1 = p_2 = 1/45, p_3 = p_4 = 6/45, p_5 = 7/45, p_6 = p_7 = p_8 = 8/45$. Vous voulez déterminer quelle bouteille est empoisonnée en les testant sur des rats. Un rat qui boit du poison meurt instantanément et un rat qui boit du vin n'est plus fiable pour une nouvelle dégustation! Vous devez donc utiliser chaque rat au plus une fois. Déterminer quelle bouteille est empoisonnée en minimisant le nombre moyen de rats utilisés. Pourquoi Brigitte Bardot n'aime pas votre protocole?

2. Problème 2: Soient $U_1, U_2, \ldots$ les lettres générées par une source sans mémoire d'alphabet $\mathcal{U}$. On suppose que la distribution p_U des lettres est soit p_1 soit p_2, c'est à dire

   (i) $\mathbb{P}(U_i = u) = p_1(u)$ pour tout $u \in \mathcal{U}$ et $i \geq 1$, ou;

   (ii) $\mathbb{P}(U_i = u) = p_2(u)$ pour tout $u \in \mathcal{U}$ et $i \geq 1$.

   Soit $K = |\mathcal{U}|$ le nombre de lettres dans l'alphabet $\mathcal{U}$, $H_1(U)$ l'entropie de U de loi p_1 donnée par (i) et $H_2(U)$ l'entropie de U de loi p_2 donnée par (ii). Soit $p_{j,\min} = \min_{u\in\mathcal{U}} p_j(u)$ la probabilité de la lettre la moins probable sous p_j. Pour un mot $w = u_1u_2\ldots u_n$, sa probabilité est $p_j(w) = \prod_{i=1}^n p_j(u_i)$ et enfin $\hat{p}(w) = \max_{j=1,2} p_j(w)$.

   (a) Soit $n \in \mathbb{N}$ et soit $\mathcal{S}$ l'ensemble des n mots maximisant $\hat{p}$. Montrer que $\mathcal{S}$ peut être représenté comme les nœuds intermédiaires d'un arbre $\mathcal{T}$ ayant $1 + (K-1)n$ feuilles.

   (b) Soit $\mathcal{W}$ l'ensemble des feuilles de $\mathcal{T}$, qui forment un dictionnaire pour la source. Soit $H_1(W)$ et $H_2(W)$ l'entropie des mots du dictionnaire sous p_1 et p_2 respectivement. Montrer que pour $j = 1, 2$, $H_j(W) \geq \log(1/Q)$ où $Q = \min_{v\in\mathcal{S}} \hat{p}(v)$.

   (c) Montrer que $|\mathcal{W}| \leq Q^{-1}(p_{1,\min}^{-1} + p_{2,\min}^{-1})$.

   (d) Soit $\mathbb{E}_j[\text{long(W)}]$ la longueur moyenne d'un mot du dictionnaire sous la probabilité p_j. À partir du dictionnaire $\mathcal{W}$, construire un code instantané. On note ρ_j le nombre de bits émis par lettre de la source si la distribution de la source est p_j. Montrer que

   $$\rho_j < H_j(U) + \frac{1 + \log(p_{1,\min}^{-1} + p_{2,\min}^{-1})}{\mathbb{E}_j[\text{long(W)}]}.$$

   (e) Montrer que cette méthode comprime la source de manière optimale asymptotiquement quand n tend vers l'infini.

3. Problème 3:

   (a) Etant donné un canal discret sans mémoire $(\mathcal{X}, p(y|x), \mathcal{Y})$ de capacité C, on considère le canal où deux sorties indépendantes sont observées pour chaque entrée: $(\mathcal{X}, p(y_1, y_2|x) = p(y_1|x)p(y_2|x), \mathcal{Y}\times\mathcal{Y})$ de capacité C_2. Montrer que $C_2 \leq 2C$.

(b) Une puce mémoire est constituée d'un million de cellules chacune mémorisant un bit. Cependant une cellule est défectueuse avec une probabilité $p = 0.01$ indépendamment des autres. Plutôt que d'attendre (très longtemps) pour une puce parfaite, nous faisons avec les puces imparfaites. La première option est de tester chaque cellule, d'identifier les cellules défectueuses et de n'utiliser que les autres cellules. Quelle est alors la mémoire utile moyenne? Une autre solution qui évite de faire ces tests, est d'utiliser des codes correcteurs d'erreurs. Quelles est alors la mémoire utile dans les deux cas suivants:

i. une cellule défectueuse correspond à un effacement.

ii. une cellule défectueuse donne en lecture l'inverse de ce qui a été écrit.

4. Problème 4:

(a) Montrer que le nombre moyen de 1 par mot-code (moyenné sur tous les mots-code) dans un code binaire linéaire de longueur N est au plus $N/2$.

(b) En déduire que la distance minimale d'un code binaire linéaire de longueur N ayant 2^L mots-code satisfait:

$$d_{\min} \leq \frac{2^{L-1}N}{2^L - 1}.$$

(c) Montrer que cette inégalité est valide pour tout code binaire (non nécessairement linéaire).

5. Problème 5:

(a) Montrer que tout code binaire $t-$correcteur $\mathcal{C}$ de longueur n $(2t < n)$ vérifie:

$$|\mathcal{C}| \sum_{i=0}^{t} \binom{n}{i} \leq 2^n,$$

et en déduire une borne supérieure sur le débit du code (en bits utiles par symbole émis).

(b) On fait maintenant tendre la taille n des mots-code vers l'infini et l'on souhaite pouvoir corriger jusqu'à $t = \lfloor \varepsilon n \rfloor$ erreurs par mot-code $(0 \leq \varepsilon < \frac{1}{2})$. Que donne alors la borne précédemment obtenue ? On exprimera la réponse à l'aide de la fonction H.

(c) Comment pouvait-on prévoir directement cette limitation théorique sur le débit ? La borne obtenue est-elle atteignable asymptotiquement ?

6. Problème 6: La classe des codes de Justesen est la seule classe de codes linéaires binaires explicitement connue contenant des codes $(\mathcal{C}_i)_{i\geq 1}$ dont les paramètres $(n_i, k_i, d_i)_{i\geq 1}$ satisfont:

$$n_i \xrightarrow[i\to\infty]{} +\infty, \qquad \liminf_{i\to\infty} \frac{k_i}{n_i} > 0 \qquad \text{et} \qquad \liminf_{i\to\infty} \frac{d_i}{n_i} > 0.$$

Notons $\mathcal{P}_r$ l'ensemble des polynômes de degré au plus r sur le corps fini $\mathbb{F}_{q^m}$ et soit $L = (\alpha_1, \ldots, \alpha_n)$ une famille de $n > r$ éléments 2 à 2 distincts de $\mathbb{F}_{q^m}$.

(a) À chaque $f \in \mathcal{P}_r$, on associe le vecteur de ses évaluations sur L, i.e. le mot-code

$$c(f) = (f(\alpha_1), \ldots, f(\alpha_n)) \in \mathbb{F}_{q^m}^n,$$

et l'on note $\mathcal{C}_{L,r}$ l'ensemble des vecteurs ainsi obtenus. Vérifier que $\mathcal{C}_{L,r}$ est un code linéaire, puis calculer sa dimension et sa distance.

(b) Montrer que cette famille de codes généralise celle des codes de Reed-Solomon.

(c) À chaque $f \in \mathcal{P}_r$, on associe à présent le mot-code

$$\widetilde{c}(f) = (f(\alpha_1), \alpha_1 f(\alpha_1), \ldots, f(\alpha_n), \alpha_n f(\alpha_n)) \in \mathbb{F}_{q^m}^{2n},$$

et l'on note $\widetilde{\mathcal{C}}_{L,r}$ l'ensemble des vecteurs ainsi obtenus. Quelles sont les longueur et dimension de $\widetilde{\mathcal{C}}_{L,r}$? Montrer qu'un mot-code non-nul contient toujours au moins $n - r$ couples $(f(\alpha_i), \alpha_i f(\alpha_i))$ 2 à 2 distincts.

(d) Expliquer comment transformer simplement un code linaire q^m-aire de dimension k et de longueur n en un code linéaire $q-$aire de dimension mk et de longueur mn.

(e) Pour tout $m \geq 1$ et tout $\varrho \in [0, 1)$, on appelle code de Justesen d'ordre m et de paramètre ϱ le code binaire obtenu en appliquant la transformation de la question (d) au code de la question (*c*) avec $q = 2$, $r = \lfloor 2^m \varrho \rfloor$ et $L = \mathbb{F}_{2^m}$. Montrer que la classe des codes de Justesen vérifie bien la propriété annoncée.

Indication: on pourra démontrer le résultat suivant: si $x_1, \ldots, x_M$ sont des mots binaires 2 à 2 distincts de longueur N, alors la proportion totale de 1, $\gamma = \frac{1}{MN} \sum_{i=1}^{M} w(x_i)$, *vérifie:*

$$NH(\gamma) \geq \log_2(M).$$