Licence Informatique
INFO0603 : Compression et cryptographie
Pascal Mignot

Devoir sur table du 3 mars 2016

Notes :
- Seul le support de cours est autorisé. Tout autre document est interdit.
- L'utilisation d'une machine à calculer sans fonction communiquante est autorisée.
- Les questions sont indépendantes.

Soit $\Sigma = (\mathtt{a}, \mathtt{b}, \mathtt{c})$ l'alphabet d'une source S. Cette source produit la sortie S suivant :

```
acbb bbbb bbcb bbbb abbc bbbb bbbb cbbb abbb bcbb bbbb bbcb
abbb bbbc bbbb bbbb cbbb bbbb bcbb bbbb abcb bbbb bbbc bbbb
abbb cbbb bbbb bcbb abbb bbcb bbbb bbbc abbb bbbb cbbb bbbb
```

Les espaces ne font pas partie du message à coder mais sont là pour améliorer sa lisibilité.

1. Si la sortie S est stocké sur un ordinateur en utilisant un code `ASCII`, quelle est la taille de S (en bit) ?
2. **Codage à taille fixe** en ne codant que les symboles effectivement utilisés par la source :
   (a) Combien de bits par symbole sont nécessaires ? On justifiera.
   (b) Proposer un code à taille fixe.
   (c) Quel serait alors la taille de S en utilisant ce codage à taille fixe.
   (d) Quelle est alors le taux de compression obtenu sur S par le codage à taille fixe ?
3. **Codage entropique**
   (a) Donner les probabilités d'apparition des symboles pour la source S.
   (b) Calculer l'entropie de la source S.
   (c) Calculer alors la taille en bits qu'aurait le message S si l'entropie était atteinte.
   (d) Quelle est alors le taux de compression obtenu sur S par un codage qui atteint l'entropie par rapport à un codage à taille fixe ?
4. **Codage prefixe à taille variable**
   (a) Un codage à taille variable donne-t-il **toujours** une meilleure compression qu'un codage à taille fixe ? On justifiera.
   (b) Un codage préfixe de taille variable avec des longueurs de code de $\{1, 2, 3\}$ est-il possible ? On justifiera. Si oui, existe-t-il un code plus court ?
   (c) Donner un code de Huffman pour la source S (les constructions intermédiaires ne sont pas nécessaires).
   (d) Quel serait alors la taille de S en utilisant ce codage de Huffman ? On donnera également le nombre de bits par symbole.
   (e) Le codage de Huffman obtenu est-il optimal ? On justifiera.
5. **Codage Arithmétique :** on utilise la probabilité d'apparition des symboles de la source calculé à la question 3.
   (a) En utilisant la probabilité d'apparition des symboles de la source, appliquer la méthode du codage arithmétique afin de calculer le codage de la chaine *abbc*.
   (b) Donner le codage binaire du centre de l'intervalle trouvé (on utilisera la méthode avec les mises à l'échelle).

(c) En utilisant la self-information, donner le nombre de bits que devrait générer les symboles de la chaine au cours d'un codage arithmétique.
(d) Même question sur la source considérée dans cet exercice.

6. **Codage prefixe à taille variable par bloc** de taille 2
   (a) Effectuer les comptages pour les blocs de taille 2 apparaissant dans cette source.
   (b) Donner le codage de Huffman pour des blocs de taille 2.
   (c) En déduire la longueur du message lorsqu'on utilise ce code.
   (d) Fait-on mieux que l'entropie ? Si oui, pourquoi ?
7. **Codage LZ77** avec une **fenêtre d'historique de 16** et **une fenêtre de codage de 8**.
   (a) Donner le codage `LZ77` pour les deux premières lignes de la sortie de la source.
   (b) Donner le codage binaire minimum d'un triplet du code obtenu à la question précédente.
   (c) Donner la taille du codage `LZ77` sur la sortie S source complète, sachant que 7 codes supplémentaires sont nécessaires pour coder la dernière ligne de la source.
8. **Codage LZ78 :** on veut maintenant effectuer le codage avec la méthode LZ78 **avec une taille maximale de dictionnaire de 16**.
   (a) Donner le codage `LZ78` de la source sur les deux premières lignes.
   (b) En supposant que la troisième ligne génère 10 codes en plus, donner la taille du codage `LZ78` sur la sortie complète. **Attention :** le codage de la sortie devra être optimisé en faisant en sorte que le codage de l'indice dans le dictionnaire dépendant de la taille du dictionnaire au moment du codage (*i.e.* comme vu en TD).
9. **Codage PPM d'ordre 2 :** après la lecture des 136 premiers caractères, on a généré l'ensemble des contextes d'ordre 2 suivants :
   – **ordre 0 :** (3/8/15/113).
   – **ordre 1 :** $a = (2/0/7/1)$, $b = (3/6/92/14)$, $c = (2/1/14/0)$.
   – **ordre 2 :** $\{ac, ca\} = (1/0/1/0)$, $\{cb, bc\} = (2/1/13/0)$, $\{ab, ba\} = (2/0/6/1)$, $\{cb, bc\} = (2/1/13/0)$, $bb = (3/5/73/13)$.
   où l'écriture $(2/0/6/1) = (n_\Delta, n_a, n_b, n_c)$ indique le comptage n_x du caractère x, et $\{ac, ca\} = (1/0/1/0)$ signifie que le comptage (1/0/1/0) a été rencontré dans les contextes ac et ca.
   (a) Appliquer l'algorithme PPM d'ordre 2 des 8 derniers caractères du texte en utilisant les ordres et les contextes donnés ci-dessus. On écrira la totalité des mises à jour des contextes (seules les mise-à-jour seront données).
   (b) Quel est le nombre de bits engendré par le codage de ces derniers caractères ? On calculera explicitement la probabilité conditionnelles dans ces contexte.
   (c) Que fait-on lorsque l'on effectue une codage PPM d'ordre 0 ?
   (d) En supposant que l'ensemble des contextes ont déjà été créé, et que les probabilités des différents contextes ne changent plus, quel est la taille codée avec PPM d'ordre 1.
   (e) Même question avec un codage PPM d'ordre 2.
10. **Codage RLE :** on veut utiliser le codage RLE sans augmenter la taille de l'alphabet à taille fixe donné à la question 4.
    (a) Proposer un codage vérifiant ces contraintes, en faisant en sorte de maximiser la longueur des runs qu'il est possible de représenter.
    (b) Effectuer le codage RLE du premier tiers de la chaine : on écrira la compression en utilisant @ comme caractère spécial et en écrivant les longueurs des runs avec nombres (@4a = run de 4 a).
    (c) Déduire des deux questions précédentes la taille complet du codage RLE associé à la source (en supposant que les deux tiers suivants génèrent le même nombre de caractères), en utilisant cette fois-ci le codage à taille fixe, et le codage.
    (d) Y-a-t-il un gain si l'on utilise un codage à taille variable pour coder le résultat du codage RLE ?