

**Quiz 2          COMP 423          March 5, 2008**

---

Answer all questions in the exam book.  
Calculators or other electronic devices are NOT permitted.  
The Quiz is marked out of 15.

1. (2 points)

- (a) What is the "inverted file" of the following: aabbabaaaaabbaab ?
- (b) Transform this inverted file into a file of gap lengths.

2. (2 points)

Briefly describe how run length coding is used in fax compression. Give at least two important details about how the code is defined.

3. (5 points)

- (a) Encode the following sequence using LZ2 (sliding window):

abaaabababb

Assume the window size is  $n_w = 4$ .

Your answer should include a parsing of the above sequence into phrases.

- (b) In the lecture notes, the number of bits used for Lempel-Ziv version LZ3 is:

$$totbits = \sum_{i=1}^{\phi(n)} (\lceil \log i \rceil + 1)$$

where it was assumed there are two symbols in the alphabet. The term  $\lceil \log i \rceil$  in the summation implies that, as the sequence length  $n$  grows, the number of bits used for phrase  $\phi(n)$  is an increasing function of  $n$ . Does this imply that LZ3 compresses shorter sequences better than longer sequences (i.e. better = smaller compression ratio) ? Briefly explain.

- (c) Give a formula for *totbits* for the case that there are  $N$  symbols in the alphabet, rather than two. Assuming the  $N$  symbols in the alphabet do not occur with equal frequency, how could we change the LZ3 code to reduce *totbits* ?

4. (2 points)

Consider the following conditional probability matrix for a first order Markov model, where the number of symbols in the alphabet is  $N = 3$ . Assume that these conditional probabilities are the same for all  $j$ .

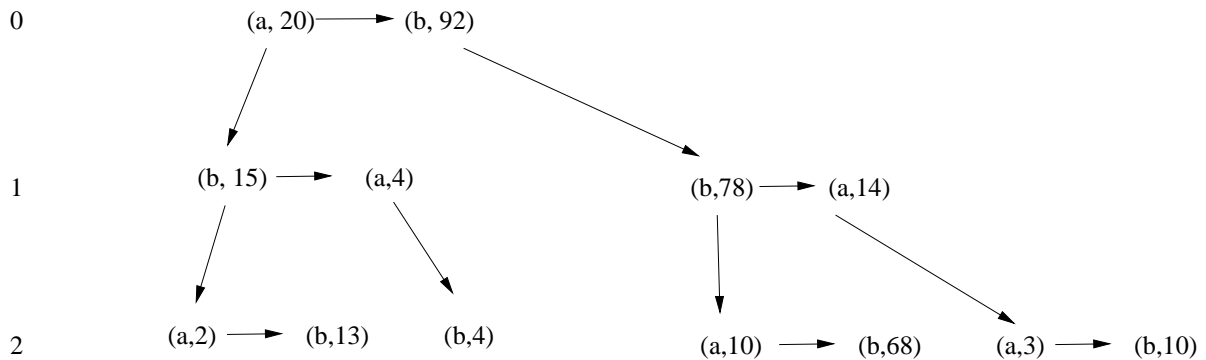
$$P(X_{j+1} | X_j) = \frac{1}{8} \begin{bmatrix} 1 & 6 & 4 \\ 5 & 0 & 1 \\ 2 & 2 & 3 \end{bmatrix}$$

- (a) What are the values of the function  $p(X_2 | X_1 = 1)$  ?
- (b) What is the value of  $p(X_2 = 2, X_3 = 1 | X_1 = 3)$  ?

5. (4 points)

The following data structure represents the frequency counts (in various contexts up to second order) of a sequence, after  $k$  symbols have been encoded.

order (k)



Use these frequency counts to answer the following.

- (a) What is  $k$  ?
- (b) Estimate  $p(X_{k+1} | X_k = a)$ .
- (c) Estimate  $p(X_{k+1} | X_{k-1} = b, X_k = a)$ .
- (d) Suppose PPM were used to encode the sequence, and the next two symbols in the sequence were  $X_{k+1} = a, X_{k+2} = a$ . How would these two symbols be encoded, i.e. what would the encoder send?

Hint: *Do not* compute probabilities.