Random Access to a database

Rules for this homework assignment.

• **You can either work in group of 2 students, if working at distance is possible, or alone. Each group of 1 or 2 students should provide one document**.
• The document should contain answers to the questions with proofs and figures.
• You should use LaTeX to write your report.
• Send the pdf file via email to aline.roumy@inria.fr,
• Your mark will be based on correctness but also clarity and quality of your exposition.
• Both French and English are allowed.
• Hard deadline **March 30th 2020, 4 pm** (Paris Time).

1. **Source Model.**
Consider a set of N sources $\{X_1, \ldots, X_N\}$. These sources form an homogeneous Markov chain of order 1 i.e. $X_1 \to X_2 \to ... \to X_N$ and satisfy $\forall\ 2 \le n \le N$,
$$H(X_n) = H(X_1) \quad (1)$$
$$H(X_n|X_{n-1}) = \alpha H(X_1) \quad (2)$$
   (a) Compute the joint entropy of these sources $H(X_1, \ldots, X_N)$, as a function of $N, \alpha, H(X_1)$.
   (b) Justify why $0 \le \alpha \le 1$.

2. **Storage rate for the IPPP scheme.**
Consider a coding scheme where X_1 is encoded losslessly and independently from the other sources. X_1 is said to be intra coded, denoted I in the following. X_2 is encoded knowing X_1 and more generally X_n is encoded knowing X_{n-1}. X_2 to X_n are said to be predicted, denoted P in the following. The overall coding scheme is denoted IPPP.

   Let the *storage* be the average number of bits needed to store a source vector realization $(x_1, \ldots, x_n)$, where the average is computed with respect to the joint distribution of $(X_1, \ldots, X_n)$. Let the *per-source storage rate* be the previously defined storage divided by N.

   (a) Compute the per-source storage rate of this IPPP coding scheme. We will denote it S^{IPPP}.
   (b) Is the IPPP scheme optimal (in the sense is it possible or not to achieve a strictly lower compression rate)?

3. **Storage rate for the IPIP scheme.**
Consider now a coding scheme where blocks of k consecutive source samples are independently encoded. The source samples in each block are encoded according to the IPPP scheme. The resulting coding scheme is called IPIP.

   (a) Compute the per-source storage rate of this IPIP coding scheme. We will denote this $S^{\text{IPIP}}(k)$. Hint: introduce $N = qk + r$.
   (b) Compare the storage rate $S^{\text{IPIP}}(k)$ to the storage rate S^{IPPP} of the scheme with a single Intra coded source. Is the IPIP scheme optimal? If not, find a condition such that the scheme is optimal.

4. **Transmission rate for the IPPP scheme for Requests of $\ell = 2$ consecutive sources.**
Let us assume that the IPPP coding scheme is used to store the data. Now a user may request any set of $\ell = 2$ consecutive sources. No decoding/reencoding is allowed. Only extraction of data within the database is allowed.

   (a) Explain why if the sources X_8 and X_9 are requested, the whole block of sources $(X_1, X_2, ..., X_9)$ need to be sent.

   Let the *transmission rate* be the number of bits that needs to be transmitted (after compression) upon request of $\ell = 2$ consecutive sources, averaged over all possible $N - 1$ requests. Let the *per-source transmission rate* be the previously defined transmission rate divided by $\ell = 2$.

   (b) Compute the per-source transmission rate. We will denote this R^{IPPP}.

5. **Transmission rate for the IPIP scheme for Requests of $\ell = 2$ consecutive sources.**
Let us assume that the IPIP coding scheme is used to store the data. Now a user may request any set of $\ell = 2$ consecutive sources. No decoding/reencoding is allowed. Only extraction of data within the database is allowed.

   Compute the per-source transmission rate, averaged over all possible requests. We will denote this $R^{\text{IPIP}}(k), k \geq 1$.

6. **Storage rate / Transmission rate region for the IPIP scheme.**
   (a) 1. Draw the storage rate vs the size of the encoding block k i.e. $S^{\text{IPIP}}(k)$,
   2. Draw the transmission rate vs the size of the encoding block k i.e. $R^{\text{IPIP}}(k)$,
   3. Draw in the plane $\mathbb{R}^2$, the set of points $\{R^{\text{IPIP}}(k), S^{\text{IPIP}}(k), \forall k\}$, with R on the X-axis and S on the Y-axis. Use logscale for R and linear scale for S.
   Suggested choices of variables: $N = 1024, k \in \{1, 2^1, 2^2, ..., 2^{10}\}, H(X_1) = 1$. Play with different choices of α ranging from low to high correlated sources (for instance $0.2 \leq \alpha \leq 0.8$).

   (b) Explain the tradeoff between the transmission and the storage rate.

   (c) Find numerically

$$k^* = \arg\min_k R^{\text{IPIP}}(k) + S^{\text{IPIP}}(k) \tag{3}$$

   and add this point in the figure above (plane in $\mathbb{R}^2$)

7. **A new source model. Storage rate / Transmission rate for the IPIP scheme.**
Consider a set of N sources $\{X_1, \ldots, X_N\}$, encoded with an IPIP scheme, where a source is intra encoded every k sources. The indices of the intra coded sources are therefore: $1, k+1, 2k+1, ...$ Let us denote $R_{a,b}$ the rate needed to encode the source with index a, when the previous closest source intra coded has index b. These rates satisfy $\forall n \in [1, N]$

$$R_{n,n} = R, \tag{4}$$

$$R_{n,n-1} = \alpha R, \tag{5}$$

$$R_{n,n-2} = R_{n,n-1} + \delta, \tag{6}$$

$$\forall i \in [1, n-1], \; R_{n,n-i} = R_{n,n-i+1} + \delta. \tag{7}$$

where R, α, δ are positive reals , $0 \leq \alpha \leq 1$ and $\delta < R$.

(a) Compute the storage rate of these sources $S^{\text{IPIP}}(k)$ as a function of R, α, δ, k.
Hint: you might consider to introduce

$$S^{\text{IPIP}}(k) = \frac{1}{N} \sum_{n=1}^{N} R_{n,I(n)} \tag{8}$$

where $I(n)$ computes the index of the closest previous source that is intra coded.

(b) Compute the transmission rate of these sources $R^{\text{IPIP}}(k)$, when requests of size $l = 2$ have been made.

(c) 1. Draw the storage rate vs the size of the encoding block k i.e. $S^{\text{IPIP}}(k)$,
2. Draw the transmission rate vs the size of the encoding block k i.e. $R^{\text{IPIP}}(k)$,
3. Draw in the plane $\mathbb{R}^2$, the set of points $\{R^{\text{IPIP}}(k), S^{\text{IPIP}}(k), \forall k\}$, with R on the X-axis and S on the Y-axis..
Suggested choices of variables: $N = 1024, k \in \{1, 2^1, 2^2, ..., 2^{10}\}, R = 1$. Play with different choices of α ranging from low to high correlated sources (for instance $0.2 \leq \alpha \leq 0.8$), and with different choices of δ (for instance $0.1 \leq \delta \leq 0.5$).

(d) Find numerically

$$k^* = \arg\min_k R^{\text{IPIP}}(k) + S^{\text{IPIP}}(k) \tag{9}$$

and add this point in the figure above (plane in $\mathbb{R}^2$)