Compresión de datos sin pérdida
Práctico 3

Ejercicio 1
Sea $\{X_n\}_{n\in\mathbb{N}}$ un proceso de Markov estacionario con matriz de probabilidades de transición $\mathbf{P}$ y distribución estacionaria π. Probar que la tasa de entropía del proceso está dada por

$$\mathcal{H} = \mathcal{H}' = -\sum_{ij} \pi_i \mathbf{P}_{ij} \log \mathbf{P}_{ij} \,.$$

Ejercicio 2
Sea X^n una secuencia aleatoria emitida por una fuente de información y sea $\hat{\mathcal{H}}(X^n) = -\frac{1}{n}\log P_{ML}(X^n)$ la tasa de entropía empírica de X^n con respecto a esa fuente. Probar que $E\left[\hat{\mathcal{H}}(X^n)\right] \leq \frac{1}{n}H(X^n)$.

Ejercicio 3
Sea x^n una secuencia de símbolos sobre un alfabeto finito $\mathcal{A} = \{a_1, a_2, \ldots a_m\}$ y sea x^0_{-k+1} una secuencia fija arbitraria, usada para determinar el estado inicial en una cadena de Markov de orden k, $k \geq 0$. La tasa de entropía empírica de orden k para una secuencia x^n está dada por

$$\hat{\mathcal{H}}_k(x^n) = -\sum_{a^k\in\mathcal{A}^k} \hat{p}(a^k) \sum_{b\in\mathcal{A}} \hat{p}(b|a^k) \log \hat{p}(b|a^k) \,,$$

donde $\hat{p}(b|a^k) = n_{b|a^k}/n_{a^k}$ y $\hat{p}(a^k) = n_{a^k}/n$, con $n_{b|a^k} = |\{i = 1\ldots n : x_i = b, x_{i-k}^{i-1} = a^k\}|$ y $n_{a^k} = \sum_{b\in\mathcal{A}} n_{b|a^k}$. Probar que para toda secuencia x^n y todo $k \in \mathbb{N}$ se cumple que $\hat{\mathcal{H}}_{k+1}(x^n) \leq \hat{\mathcal{H}}_k(x^n)$.

Ejercicio 4 *Compresión de una fuente de Markov*
Considere un proceso de Markov *estacionario* con tres estados, $\{1, 2, 3\}$, es decir, un proceso de Markov de orden 1 sobre un alfabeto de tres símbolos $\mathcal{A} = \{1, 2, 3\}$. La matriz de probabilidades de transición es

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} .$$

Por ejemplo, la probabilidad de que observar el símbolo 1 inmediatamente después del símbolo 3 es cero.

Denotamos con $\mathbf{P}_{i*}$ a la fila i de $\mathbf{P}$, y con $\mathcal{A}_i$ al soporte de $\mathbf{P}_{i*}$, es decir, al subconjunto de $\mathcal{A}$ con probabilidad positiva según $\mathbf{P}_{i*}$. Para cada i, $i \in \{1, 2, 3\}$, diseñar un código instantáneo C_i que es óptimo para una variable que toma valores en $\mathcal{A}_i$ con probabilidades determinadas por $\mathbf{P}_{i*}$.

Considere el código C para secuencias de símbolos generadas para este proceso de Markov definido por el siguiente algoritmo:

1. El estado inicial X_1 se codifica de alguna forma arbitraria fijada de antemano, digamos con el código C_1.

2. Para $i > 1$, se codifica el símbolo X_i usando el código C_x, donde $x = X_{i-1}$.

Calcular la cantidad media asintótica de bits de código por símbolo de la fuente,

$$\lim_{n\to\infty} E\left[\frac{|C(X^n)|}{n}\right],$$

y compare con la tasa de entropía del proceso.

Ejercicio 5 *Cota asintótica sobre el largo de código esperado*
Sea $\{X_n\}_{n\in\mathbb{N}}$ un proceso estocástico estacionario sobre un alfabeto $\mathcal{A}$, con tasa de entropía $\mathcal{H}$.

1. Mostrar que para todo código unívocamente decodificable definido sobre bloques de n símbolos, $C : \mathcal{A}^n \to \mathcal{B}^\star$, la esperanza del largo de código satisface

$$\frac{1}{n}E\left[|C(X^n)|\right] \geq \mathcal{H}\,.$$

2. Sea $\epsilon > 0$ arbitrario. Mostrar que existe N (que puede depender de ϵ) tal que para todo natural $n > N$, existe un código $C : \mathcal{A}^n \to \mathcal{B}^\star$ para el cual se cumple

$$\frac{1}{n}E\left[|C(X^n)|\right] < \mathcal{H} + \epsilon\,.$$

Ejercicio 6 *Run length coding*
Una fuente produce una secuencia i.i.d sobre el alfabeto binario $\mathcal{B} = \{0, 1\}$, con probabilidades $p(0) = 9/10$ y $p(1) = 1/10$. Se codifica la secuencia en dos etapas, primero contando la cantidad de ceros consecutivos (hasta un máximo de 8) entre todo par de ocurrencias sucesivas de unos, y luego codificando estos largos de corridas de ceros con una codificación binaria.

Secuencia de la fuente	Dígito intermedio (largo de corrida)
1	0
01	1
001	2
0001	3
00001	4
000001	5
0000001	6
00000001	7
00000000	8

Cuadro 1: Mapeo de bits de la fuente a dígitos intermedios.

La primera etapa mapea secuencias de la fuente en secuencias de *dígitos intermedios* según la correspondencia dada por el Cuadro 1. Por ejemplo, la secuencia binaria 10010000000001100001 se traduce a dígitos intermedios de la siguiente manera

1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
0			2								8			2	0					4

(Podemos suponer que toda secuencia finita de entrada termina con un 1 a los efectos de determinar la secuencia de dígitos intermedios. Este símbolo adicional puede ser retirado por el decodificador si conoce el largo de la secuencia).

La etapa final traduce dígitos intermedios a secuencias de dígitos binarios, que son los que forman la salida del codificador, asignando una palabra de código de cuatro bits, de la forma $1xxx$, a cada uno

de los dígitos intermedios $0 \ldots 7$, y asignando la palabra de código de un bit, 0, al dígito intermedio 8. Si, por ejemplo, se codifica al dígito intermedio 0 con la palabra 1000, al 2 con 1010, y al 4 con 1100, la secuencia del ejemplo anterior se codificaría como $1000, 1010, 0, 1010, 1000, 1100$ (las comas se incluyen sólo a los efectos de facilitar la lectura).

1. Justificar que la codificación es unívocamente decodificable.
2. Observar que la secuencia de dígitos intermedios es i.i.d. y hallar su distribución.
3. Determinar la cantidad media N_1 de bits de la fuente por símbolo intermedio.
4. Determinar la cantidad media N_2 de bits de código (salida) por símbolo intermedio.
5. Usar la ley de los grandes números para mostrar que, para secuencias muy largas, la tasa de compresión (#bits de salida / #bits de la fuente) es cercana a N_2/N_1 con alta probabilidad.
6. Comparar esta tasa de compresión con la que se obtiene aplicando Huffman a bloques de cuatro bits de la fuente.