Compresión de Datos sin Pérdida 2021 - Prueba 2

Justificar todas las respuestas, para lo cual pueden usarse resultados estudiados durante el curso (teórico y práctico) **sin necesidad de demostrarlos**.

Ejercicio 1 *(50 puntos)*
Sea $\{X_n\}_{n\geq 1}$ un proceso de Markov con conjunto de estados $\{0,1,2\}$, donde X_1 tiene distribución uniforme en $\{0,1,2\}$ y la matriz de probabilidades de transición es

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

1. ¿Es este proceso estacionario?

2. Calcule la tasa de entropía del proceso.

3. Definimos $\{Z_n\}_{n\geq 1}$, $Z_n \in \{0,1,2\}$, mediante las ecuaciones (1)-(2). La tabla debajo define Z_n para $n > 1$ equivalentemente a (2).

$$\begin{aligned} Z_1 &= X_1, && (1) \\ Z_n &= X_n - X_{n-1} \mod 3, \quad n > 1. && (2) \end{aligned}$$

		X_n		
		0	1	2
X_{n-1}	0	0	1	2
	1	2	0	1
	2	1	2	0

   Muestre que, para $n > 1$, $P(Z_n = 0) = \frac{1}{2}$ y $P(Z_n = 1) = P(Z_n = 2) = \frac{1}{4}$.

4. Observar que $(X_1, X_2, \ldots, X_n)$ se puede obtener a partir de $(Z_1, Z_2, \ldots, Z_n)$ mediante la siguiente recurrencia

$$\begin{aligned} X_1 &= Z_1, && (3) \\ X_n &= X_{n-1} + Z_n \mod 3, \quad n > 1. && (4) \end{aligned}$$

   A partir de esta observación, definimos el código C_n para $(X_1, X_2, \ldots, X_n)$, como la concatenación de las codificaciones de Z_i, $1 \leq i \leq n$, según un código fijo C,

$$C_n(X_1, X_2, \ldots, X_n) = C(Z_1)C(Z_2)\ldots C(Z_n)\,, \tag{5}$$

   donde $C(0) = 0$, $C(1) = 10$, $C(2) = 11$.

   Calcule el largo de código esperado $E\left[\left|C_n(X_1, X_2, \ldots, X_n)\right|\right]$.

   Sugerencia: Calcule por separado las esperanzas de $\left|C(Z_1)\right|$ y de $\left|C(Z_i)\right|$, $i > 1$.

5. ¿Es posible que exista $\{C'_n\}_{n\geq 1}$, donde C'_n es un código unívocamente decodificable para $(X_1, X_2, \ldots, X_n)$, tal que

$$\lim_{n\to\infty} E\left[\frac{\left|C'_n(X_1, X_2, \ldots, X_n)\right|}{n}\right] < \lim_{n\to\infty} E\left[\frac{\left|C_n(X_1, X_2, \ldots, X_n)\right|}{n}\right]? \tag{6}$$

Solución:

1. El proceso es estacionario porque para el vector de probabilidad para el estado inicial, $\pi = (1/3, 1/3, 1/3)$, se cumple que $\pi P = \pi$.
2. Para un proceso de Markov estacionario tenemos $\mathcal{H} = H(X_2|X_1)$, que en este caso vale

$$\begin{aligned}\mathcal{H} &= \frac{1}{3}H(1/2,1/4,1/4) + \frac{1}{3}H(1/4,1/2,1/4) + \frac{1}{3}H(1/4,1/4,1/2) \\ &= 3 \times \frac{1}{3}H(1/2,1/4,1/4) = \frac{3}{2} \text{ bits/símbolo.}\end{aligned}$$

3.

$$\begin{aligned}P(Z_n = 0) &= \sum_{x=0}^{2} P(X_{n-1} = x, X_n = x) \\ &= \sum_{x=0}^{2} P(X_{n-1} = x)P(X_n = x|X_{n-1} = x) \\ &= \sum_{x=0}^{2} \frac{1}{3}P_{x,x} = \frac{1}{2} \\ P(Z_n = 1) &= \sum_{x=0}^{2} P(X_{n-1} = x, X_n = x+1 \mod 3) \\ &= \sum_{x=0}^{2} P(X_{n-1} = x)P(X_n = x+1 \mod 3|X_{n-1} = x) \\ &= \frac{1}{3}P_{0,1} + \frac{1}{3}P_{1,2} + \frac{1}{3}P_{2,0} = \frac{1}{4} \\ P(Z_n = 2) &= 1 - P(Z_n = 0) - P(Z_n = 1) = \frac{1}{4}\,.\end{aligned}$$

4. Z_1 tiene distribución uniforme, por lo cual tenemos

$$E\left[\left|C(Z_1)\right|\right] = \frac{1}{3} \times 1 + \frac{1}{3} \times 2 + \frac{1}{3} \times 2 = \frac{5}{3}\,. \tag{7}$$

Para $n > 1$, por la parte anterior tenemos

$$E\left[\left|C(Z_n)\right|\right] = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = \frac{3}{2}\,. \tag{8}$$

Por lo tanto, por la definición (5) y la linealidad de la esperanza, obtenemos

$$E\left[\left|C_n(X_1, X_2, \ldots, X_n)\right|\right] = \sum_{i=1}^{n} \left|C(Z_n)\right| = \frac{5}{3} + \frac{3}{2}(n-1)\,. \tag{9}$$

5. Tenemos

$$\lim_{n\to\infty} E\left[\frac{|C_n(X_1, X_2, \ldots, X_n)|}{n}\right] = \frac{3}{2} = \mathcal{H}\,. \tag{10}$$

Si existiera tal C', para n suficientemente grande se cumpliría

$$E\left[\frac{|C'_n(X_1, X_2, \ldots, X_n)|}{n}\right] < \mathcal{H}\,,$$

lo cual sería una contradicción con el hecho de que la tasa de entropía es una cota inferior para la esperanza del largo de código por símbolo para códigos unívocamente decodificables para $(X_1, X_2, \ldots, X_n)$ (práctico 3, ejercicio 5.1).

Ejercicio 2 *(50 puntos)*
Para cada natural par n, sea x^n la secuencia binaria compuesta por $n/2$ unos consecutivos seguidos de $n/2$ ceros consecutivos

$$x^n = \underbrace{11\ldots1}_{n/2}\underbrace{00\ldots0}_{n/2}\,.$$

1. Sea $\mathcal{C} = \{P_\theta\}_{\theta\in\Theta}$, la familia de modelos de *Bernoulli* (i.i.d. sobre $\{0,1\}$). Calcule la tasa de entropía empírica de x^n con respecto a $\mathcal{C}$.
2. Sea $\mathcal{C}' = \{P'_{\theta'}\}_{\theta'\in\Theta'}$, la familia de modelos de *Markov* de orden 1 sobre $\{0,1\}$ con estado inicial fijo $s_1 = 0$ (es decir, el estado inicial se determina asumiendo la existencia de un símbolo adicional, $x_0 = 0$, que precede a la secuencia x^n). Calcule la tasa de entropía empírica de x^n con respecto a $\mathcal{C}'$.
3. Obtenga el parsing LZ78 de x^n.
4. Muestre que la cantidad de bits de codificación por símbolo de entrada en LZ78 tiende a cero con n.

Solución:

1. Tomando el parámetro θ como la probabilidad del símbolo 0, con $\Theta = [0,1]$, el parámetro de máxima verosimilitud de x^n con respecto a $\mathcal{C}$ es la frecuencia de ocurrencias de 0, que es 1/2. Tenemos entonces

$$\hat{\mathcal{H}}(x^n) = H(\hat{\theta}) = H\left(\frac{1}{2}\right) = 1 \text{ bit/símbolo}\,.$$

2. Con respecto a $\mathcal{C}'$, la mitad de los símbolos de x^n ocurren en estado 0, marcados en negro en (11), y la mitad en estado 1, marcados en rojo en (11):

$$x^n = 11 \ldots 100 \ldots 0\,. \tag{11}$$

En cada estado, uno de los símbolos del alfabeto ocurre solo una vez, y el otro $n/2-1$ veces. En consecuencia, tenemos (práctico 3, ejercicio 3)

$$\hat{\mathcal{H}}(x^n) = \frac{1}{2}H\left(\frac{1}{n/2}\right) + \frac{1}{2}H\left(\frac{1}{n/2}\right) = H\left(\frac{2}{n}\right) \text{ bits/símbolo}\,. \tag{12}$$

La ecuación (12) pude derivarse definiendo la clase $\mathcal{C}'$ en función del espacio de parámetros $\Theta' = [0,1]\times[0,1]$, donde el parámetro $\theta' = (\theta_0, \theta_1) \in \Theta'$ define la probabilidad del símbolo 0 en estado 0, θ_0, y la probabilidad del símbolo 0 en estado 1, θ_1. Luego tenemos

$$P'_{\theta'}(x^n) = \theta_0^{n/2-1}(1-\theta_0)\theta_1(1-\theta_1)^{n/2-1}\,,$$

que se maximiza en $\theta_0 = \frac{n/2-1}{n/2} = 1-\frac{2}{n}$ y $\theta_1 = \frac{1}{n/2} = \frac{2}{n}$, que maximizan los factores $\theta_0^{n/2-1}(1-\theta_0)$ y $\theta_1(1-\theta_1)^{n/2-1}$, respectivamente. Por lo tanto tenemos

$$\begin{aligned}\frac{-\log P_{ML}(n^n)}{n} &= -\frac{n/2-1}{n}\log\left(1-\frac{2}{n}\right) - \frac{1}{n}\log\frac{2}{n} \\ &\quad -\frac{1}{n}\log\frac{2}{n} - \frac{n/2-1}{n}\log\left(1-\frac{2}{n}\right) \\ &= \frac{1}{2}H\left(\frac{1}{n/2}\right) + \frac{1}{2}H\left(\frac{1}{n/2}\right)\,.\end{aligned}$$

3. El parsing LZ78 de x^n es de la forma

$$x^n = 1, 11, 111, \ldots, z, 0, 00, 000, \ldots w\,, \tag{13}$$

donde w es la concatenación de 0 o más ceros y, dependiendo del valor de n, z es de la forma $11\ldots1$ o de la forma $11\ldots10$. Más específicamente, denotando con f_i y f'_i a las cadenas formada por i unos y por i ceros, respectivamente, el parsing es

$$x^n = f_1, f_2, f_3, \ldots f_k, z, f'_1, f'_2, f'_3, \ldots f'_{k'}, w\,, \quad k \geq 0, k' \geq 0\,, \tag{14}$$

donde $w = f'_{j'}$, para cierto j', $1 \leq j' \leq k'+1$, y $z = f_{k+1}$, si existe un entero k tal que $\sum_{i=1}^{k+1} i = \frac{n}{2}$, y $z = f_j 0$, para cierto j, $1 \leq j \leq k$, en caso contrario.

4. Para cualquier asignación de probabilidad de orden k, Q_k, se cumple que el largo de código por símbolo de entrada satisface

$$\frac{L_{LZ78}(x^n)}{n} = \frac{c(n)\lceil \log c(n)+1\rceil}{n} = \frac{-\log Q_k(x^n)}{n} + o(1)\,, \tag{15}$$

que en particular para $k = 1$ y $Q_k(x^n)$ igual $P_{ML}(x^n)$ con respecto a $\mathcal{C}'$, resulta, por la parte 2, en

$$\frac{L_{LZ78}(x^n)}{n} = H\left(\frac{2}{n}\right) + o(1)\,, \tag{16}$$

que tiende a 0 con n.
Alternativamente, observamos que hay $n/2$ unos en x^n y por lo tanto

$$\frac{n}{2} \geq \sum_{i=1}^{k} |f_i| = \sum_{i=1}^{k} i = \frac{k(k+1)}{2} \geq \frac{k^2}{2}\,,$$

de donde concluimos que $k \leq \sqrt{n}$. Análogamente también concluimos que $k' \leq \sqrt{n}$.
En consecuencia, la cantidad de frases en x^n satisface $c(n) \leq 2\sqrt{n} + 2$, y por lo tanto

$$\frac{L_{LZ78}(x^n)}{n} = \frac{c(n)\lceil \log c(n) + 1 \rceil}{n} = O\left(\frac{\sqrt{n}\log n}{n}\right) = o(1)\,. \tag{17}$$