Teoría Avanzada de la Termodinámica – 2do cuatrimestre de 2023

Guía 3: *Entropía e infomación*

1. Considere un sistema que, a medida que pasa el tiempo, va cambiando de estado de forma aleatoria. Entonces se dice que el sistema experimenta un *proceso aleatorio*. El proceso es *de Markov* si la probabilidad (o densidad de probabilidad, según el caso) condicional satisface

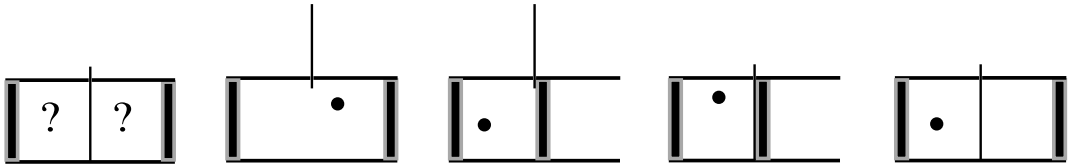
$$p(x_n, t_n|x_1, t_1; \ldots; x_{n-1}, t_{n-1}) = p(x_n, t_n|x_{n-1}, t_{n-1}).$$

   Es decir, si conocemos el estado del sistema a tiempo t_{n-1}, entonces conocer la historia previa no afecta a las probabilidades a tiempo t_n. Demuestre que todas las probabilidades conjuntas de n tiempos, $p_n(x_1, t_1; x_2, t_2; \ldots; x_n, t_n)$, pueden construirse a partir de $p(x, t)$ y $p(x, t|x', t')$, con $t \geq t'$.

2. Una memoria elemental consiste en un conjunto de cajas o "bits", cada una de volumen V y conteniendo una sola partícula. A cada lado de la caja hay un pistón y en el medio una partición removible. Si la partícula está en el lado izquierdo, el estado es **0** y si está en el lado derecho es **1**.

   Las cajas están en contacto con un foco térmico a temperatura T. La partícula puede tratarse como un gas ideal, con ecuación de estado $PV = kT$ y entropía $S = S_\mathrm{r} + k\log(V/V_\mathrm{r}) + ck\log(T/T_\mathrm{r})$, donde S_r, V_r y T_r corresponden a un estado de referencia y c es constante. Un bit es borrado cuando, independientemente de su estado inicial, y sin conocimiento de este estado, se lo fuerza al estado **0** mediante las siguientes operaciones:

   - Se quita la partición, permitiendo que el gas se expanda libremente de $V/2$ a V.
   - Isotérmica y cuasiestáticamente se comprime el gas hasta un volumen $V/2$ mediante el pistón de la derecha, de manera que, sea cuál sea el estado inicial, la partícula termina en el lado izquierdo.
   - Por último, se reinserta la partición y se vuelve el pistón a su estado inicial a la derecha de la caja.



   a) Calcule el calor cedido por la memoria durante este proceso de borrado de un bit, y compare con el principio de Landauer.

   b) El segundo pistón, que hasta aquí no se usó, tiene una función: proponga uno o más procedimientos para pasar del estado **0** al **1**, de manera reversible, sin transferencia de calor ni realización de trabajo.

   c) ¿Cuál es el trabajo total entregado por una máquina de Szilard que opere usando esta memoria como demonio de Maxwell?

3. La entropía de una distribución de probabilidad p se define como

$$S(p) = -\sum_{r=1}^{M} p(r) \log p(r),$$

donde $p(r)$ es la probabilidad del resultado r, entre M resultados posibles. Esta cantidad mide el grado de incertidumbre acerca del resultado del experimento o, en otras palabras, la cantidad de información que adquirimos al conocer ese resultado. Para ver que esto es así, muestre que S es mínima cuando el resultado del experimento se conoce con certeza (es decir, un resultado tiene probabilidad 1 y los demás tienen probabilidad 0), y máxima cuando la distribución es uniforme, $p(r) = 1/M$.

Sugerencia: para ver lo segundo, empiece graficando la función $x \log x$ en el intervalo $[0, 1]$, y después use la desigualdad de Jensen: si Φ es convexa, entonces

$$\Phi\left(\frac{1}{M}\sum_{k=1}^{M} a_k\right) \leq \frac{1}{M}\sum_{k=1}^{M} \Phi(a_k).$$

Comentario: la base en la que se calcula el logaritmo que aparece en la definición de la entropía define la unidad de medida de la información. En el contexto de Teoría de la Información suele usarse el logaritmo en base 2 y la entropía se mide en *bits*.

4. Sean A y B dos experimentos aleatorios. Dado un resultado α de A y un resultado β de B, sea $p(\alpha|\beta)$ la probabilidad de α condicionada a β. La *entropía condicional* de A dado B es el valor medio de las entropías de las distribuciones $p(\ |\beta)$ sobre todos los valores posibles de β,

$$S(A|B) = \sum_{\beta} p(\beta) S(p(\ |\beta)) = -\sum_{\alpha,\beta} p(\beta) p(\alpha|\beta) \log p(\alpha|\beta) = -\sum_{\alpha,\beta} p(\alpha,\beta) \log \frac{p(\alpha,\beta)}{p(\beta)},$$

y se interpreta como el grado de incertidumbre acerca del resultado de A si conocemos el resultado de B. Esta cantidad cumple dos desigualdades importantes, que se derivan de la desigualdad $\log x \leq x - 1$.

a) Muestre que $S(A|B) \leq S(A)$; es decir, conocer el resultado de B no agrega incertidumbre acerca del resultado de A.

b) Muestre que $S(A|BC) \leq S(A|B)$; es decir, conocer el resultado de C además del de B no agrega incertidumbre acerca del resultado de A. Nótese que la desigualdad del ítem anterior es un caso particular de ésta, correspondiente al caso en que el experimento B consiste en no hacer nada.

La entropía condicional se puede escribir en términos de las entropías de los experimentos AB y B. Las desigualdades anteriores, entonces, dan lugar a desigualdades para la entropía.

c) Muestre que $S(A|B) = S(AB) - S(B)$.

d) A partir de este resultado, reescriba las desigualdades de los ítems *a)* y *b)* en términos de $S(A)$, $S(B)$, $S(AB)$, $S(BC)$ y $S(ABC)$. La primera de estas desigualdades se conoce como la *subaditividad* de la entropía, y la segunda como *subaditividad fuerte*.

5. También puede hablarse de la entropía de una fuente de mensajes. Supongamos que la fuente emite mensajes con un número arbitrario de caracteres. La entropía asociada a los mensajes de longitud N es

$$S_N = -\sum_{r=1}^{M_N} p_N(r) \log p_N(r),$$

donde la suma se extiende a los M_N mensajes de longitud N, y $p_N(r)$ es la probabilidad del mensaje r. Si existe el límite

$$s = \lim_{N\to\infty} \frac{S_N}{N},$$

entonces s es la entropía por caracter asociada a la fuente. Los siguientes ítems analizan lo que sucede cuando la probabilidad de cada caracter depende de un número finito de caracteres inmediatamente anteriores. El caso más simple es una fuente sin correlaciones, donde cada caracter es emitido con independencia de los anteriores. El siguiente caso en complejidad es cuando la probabilidad de emisión de un caracter depende sólo del inmediatamente anterior. Y así siguiendo con fuentes más complicadas, donde la probabilidad de cada caracter depende de los dos, tres o más caracteres anteriores. Por ejemplo, en español, $p(\text{o}|\text{e, n, t, r, o, p, í}) \approx 0$, $p(\text{P, l, a, n, c, k}|\text{F, o, k, k, e, r, -}) \approx 1$.

Para fijar la notación, un mensaje de N caracteres es una $N-$upla ordenada $(x_1, x_2, \ldots, x_N)$. El índice de cada x_i es la posición que ocupa dentro del mensaje. Cada x_i se elige de entre un conjunto de M símbolos distintos. Se asume siempre que el proceso es homogéneo, de modo que la probabilidad de encontrar un dado grupo de n caracteres dentro del mensaje no depende de qué parte del mensaje se mire. Por ejemplo $p(x_1 = a) = p(x_2 = a)$, $p(x_1 = a, x_2 = b) = p(x_3 = a, x_4 = b)$, etc. (Esta hipótesis sólo es razonable para mensajes largos y no muy cerca de los extremos; en algún sentido, se desprecian los efectos de borde.) Así, el índice i puede usarse también para señalar la posición en que aparece un dado caracter dentro de un grupo de caracteres **consecutivos**. Notar que no es lo mismo $p(x_1 = a, x_2 = b)$ que $p(x_1 = a, x_3 = b)$.

   a) Si los caracteres no están correlacionados, demostrar que S_N se escribe en términos de S_1. Encontrar s. A esta s, que se calcula usando sólo $p_1(x_1)$, y que es exacta en el caso en que no hay correlaciones, la llamaremos s_1.

   b) Suponga que el proceso de emisión de caracteres liga la probabilidad de cada nuevo caracter sólo con el anterior. Es decir, es un proceso de Markov en el sentido habitual: basta con conocer las probabilidades $p_2(x_1, x_2)$ y $p_1(x_1)$. Escriba S_N en términos de S_2 y S_1. ¿Cuánto vale s? A esta s, que se calcula usando sólo $p_2(x_1, x_2)$ y $p_1(x_1)$, y que es exacta cuando las probabilidades condicionales sólo ligan 2 caracteres, la llamaremos s_2.

Más generalmente, cada caracter estará ligado a los n anteriores (notar que los caracteres en las posiciones $i \leq n$ están ligados en verdad con un número menor de caracteres, pues el mensaje empieza en $i = 1$). Esta es una extensión natural de los procesos de Markov usuales. Para $N > n$ vale

$$p(x_N|\underbrace{x_1, x_2, \ldots, x_{N-2}, x_{N-1}}_{N-1}) = p(x_N|\underbrace{x_{N-n}, \ldots, x_{N-2}, x_{N-1}}_{n}). \tag{1}$$

c) En analogía con los procesos de Markov ordinarios, escriba cualquier probabilidad conjunta de N caracteres consecutivos usando $p_1(x_1)$ y las condicionales

$$p(x_2|x_1), \quad p(x_3|x_1, x_2), \quad \ldots \quad p(x_N|x_{N-n}, \ldots, x_{N-2}, x_{N-1}).$$

d) Demostrar que es posible escribir S_N en términos de S_{n+1} y S_n. ¿Cuánto vale s? A esta s, que se calcula usando sólo p_{n+1} y p_n, y que es exacta si vale la Ec. (1), la llamaremos s_{n+1}.

e) A partir de la subaditividad fuerte de la entropía, muestre que $s_{n+1} \leq s_n$. Interprete este resultado.

Puede darse el caso de que la probabilidad de cada nuevo caracter dependa siempre de todos los anteriores, sin importar lo extenso que sea el mensaje. Así, en principio, para calcular la entropía s sería necesario conocer las probabilidades conjuntas $p_N(x_1, \ldots, x_N)$ con N arbitrariamente grande. El cálculo de la entropía usando la estadística que incluye a lo sumo grupos de $n + 1$ caracteres, es decir, lo que hemos llamado s_{n+1}, será, como mucho, una aproximación a la s verdadera. Puede demostrarse que bajo ciertas condiciones $s = \lim_{n\to\infty} s_{n+1}$. Así, para ciertas fuentes tiene sentido tratar a s_1, s_2, s_3, etc., como aproximaciones cada vez mejores de s. Además, por lo que hemos visto en el último ítem tenemos $s \leq s_{n+1} \leq s_n$, de modo que cada s_n da una cota más precisa para s.

6. Es interesante considerar la entropía asociada a los lenguajes humanos. Una fuente típica de un lenguaje sería un escritor. Un escritor puede modelarse como un proceso que genera caracteres al azar (algunos escritores más que otros) siguiendo ciertas reglas probabilísticas. Debido a que no se tiene acceso al proceso estocástico fundamental que genera los caracteres, la información sobre las probabilidades y las correlaciones debe estimarse directamente de la obra del escritor. Los archivos que acompañan esta guía consisten en dos versiones de *Moby Dick*, la original y una traducción al español. Para mayor simplicidad, se usan sólo letras minúsculas y se han eliminado todos los caracteres que no forman parte del conjunto de 27 símbolos compuesto por el espacio simple y las letras de la *a* a la *z*. La *ñ* se ha sustituido[1] por la *n*. Además se incluyen dos archivos que consisten únicamente en los mil primeros caracteres de cada versión del libro, útiles para hacer pruebas antes de emprender el análisis del libro completo. El objetivo es estimar la entropía por caracter de Herman Melville. Referida a *Moby Dick* en particular, esta entropía da una medida de qué tan predecible es el texto: si una palabra se corta al pasar de una página a la siguiente, a menor entropía más fácil es adivinar cuáles son las letras que fal-

[1]No se trata tanto de ahorrar un símbolo, sino de evitar los caracteres especiales, que quizá demanden algunas operaciones más de la computadora. Incluso podría reemplazarse la *ñ* por algún número, por ejemplo.

xzyxyz (puede fallar).

a) La aproximación más elemental para la entropía corresponde a asumir que los caracteres están descorrelacionados y que todos tienen la misma probabilidad. Por lo que se ha visto más arriba, esto da la entropía máxima teórica. Bajo esta suposición ni siquiera hace falta examinar *Moby Dick*. Calcule la entropía $s_{\max}$ en bits para el conjunto de 27 símbolos disponibles. Una fuente que produzca caracteres equiprobables proporciona el máximo posible de información por caracter. Los lenguajes naturales están regulados por otros principios.

b) La aproximación de orden uno corresponde a asumir, como antes, que no hay correlaciones pero que los caracteres no son necesariamente equiprobables. Para calcular s_1 sólo importan las probabilidades $p_1(x_1)$. A partir de los archivos suministrados, estime las probabilidades de cada caracter, en inglés y en español, y calcule s_1. Como referencia, según estimaciones del propio Shannon[2], en esta aproximación la entropía del idioma inglés, no de *Moby Dick* en particular, es $s_1 \sim 4{,}03$. (Recordar que se usan logaritmos en base 2.)

c) Evidentemente, la aproximación de orden uno es muy cruda. Estime las probabilidades de pares de caracteres $p_2(x_1, x_2)$ y calcule s_2, es decir, la aproximación para s que tiene en cuenta una memoria de un solo caracter. (Para el idioma inglés, Shannon da el valor $s_2 \sim 3{,}32$.)

d) La siguiente aproximación es truncar la estadística en grupos de tres caracteres, y calcular s_3. (Shannon estima en este caso $s_3 \sim 3{,}1$.)

Si, en lugar de *Moby Dick*, se trabaja con un texto finito generado al azar, usando los 27 símbolos disponibles y sin correlaciones entre caracteres sucesivos, las tres estimaciones de la entropía, s_1, s_2 y s_3, deberían dar valores muy próximos. (Teóricamente deberían coincidir, pero hay que recordar que las probabilidades se infieren a partir de una muestra finita.) Es más, si los caracteres son equiprobables, estos valores deberían ser cercanos al valor teórico $s_{\max}$ de los 27 símbolos.

e) Genere un texto al azar, con distribución uniforme de caracteres, y con el mismo número de caracteres que *Moby Dick*. Compare los tres valores para la estimación de la entropía. (Esto da un método para chequear que sus programas no tengan errores evidentes.)

f) Genere una muestra de texto de 1000 caracteres de acuerdo a las probabilidades p_1, p_2 y p_3 estimadas a partir de las versiones en inglés y en español de *Moby Dick*.

Y una última pregunta.

g) Recordando que la entropía mide la cantidad de bits por caracter, estime la entropía de *Moby Dick* comparando el tamaño de los archivos suministrados con los que resultan de la compresión de esos archivos en formato zip.

[2] C.E. Shannon, *Prediction and Entropy of Printed English*, Bell System Technical Journal **30**, 50 (1951).