Temas avanzados de termodinámica y física estadística − 1er. cuatrimestre de 2013

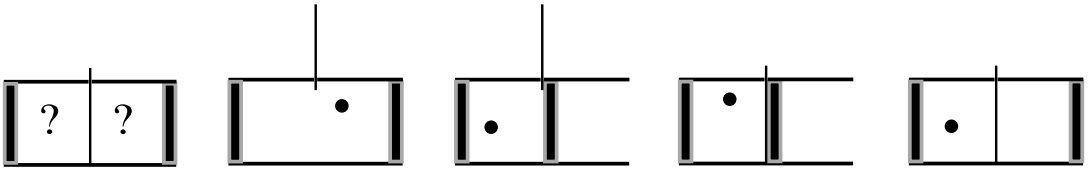
Guía 6: *Entropía e información*

1. 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**.

estado **0** estado **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 $S = S_\mathrm{r} + k\log(V/V_\mathrm{r}) + c\,k\log(T/T_\mathrm{r})$, 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) Encuentre el cambio total de entropía en el sistema formado por la memoria y el reservorio térmico durante este proceso de borrado de un bit.

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.

2. En Teoría de la Información, la entropía contenida en un mensaje se define como

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

donde $p(r)$ es la probabilidad del mensaje r, entre M mensajes posibles. Notar que la entropía no está referida a un mensaje en particular, sino a un ensamble de mensajes posibles. A mayor entropía mayor es la incertidumbre antes de recibir un mensaje, y entonces mayor es la información transmitida por cada mensaje. La base en la que se calcula el logaritmo define la unidad de medida de la información. En este contexto suele usarse el logaritmo en base 2 y la entropía se mide en *bits*.

(a) Graficar $x\log x$ en el intervalo $\overline{0,1}$. ¿Qué pasa con S cuando uno de los mensajes ocurre con probabilidad 1? Interpretar esto en términos de la frase "chocolate por la noticia".

(b) Demostrar que S es no negativa y que es máxima cuando todos los mensajes tienen igual probabilidad, $p(r) = 1/M$. *Sugerencia:* que hay un extremo es fácil; demostrar que es un máximo es más difícil. Las dos cosas pueden hacerse en un solo paso usando la llamada desigualdad de Jensen, de la teoría de funciones convexas. Si Φ es continua y convexa, entonces

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

(Es fácil entender esta desigualdad si se piensa a $\Phi(x)$ como la función que define la coordenada y de cierto arco convexo, con una distribución de masas unidad sobre el arco y cuyas coordenadas x son las cantidades a_k. La desigualdad dice que el centro de masa está *dentro* del arco, lo que es bastante intuitivo; Callen, §17-1.)

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 sólo se extiende a los M_N mensajes de longitud N. 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, $p(o|z,e,t) = 0$, $p(n,o,s,o,c,o,m,i,o|c,i,t,a,d,o,\ ) = 1$.

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 $p(x_1 = a, x_3 = b)$.

c) 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.

d) 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. Ahora para $N > n$ vale

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

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

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

f) 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 las probabilidades condicionales que ligan hasta $n + 1$ caracteres consecutivos, y que es exacta si vale (1), la llamaremos s_{n+1}.

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. El ítem anterior dice como calcular s_{n+1}. 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.

Es interesante considerar la entropía asociada a los lenguajes humanos. Una fuente típica de un lenguaje sería un escritor. Un escritor puede modelarse (algunos más que otros) como un proceso que genera caracteres 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* 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 (que no será una fuente infinita de caracteres, pero se le parece mucho). 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 de qué palabra se tra

*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.

ta.

g) La aproximación más elemental para la entropía corresponde a asumir que cada caracter tiene la misma probabilidad, lo que también implica que no hay correlaciones. Como se ha visto en los primeros ítems, 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.

h) 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[†], 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.)

i) 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 $s_2 \sim 3.32$.)

j) La siguiente aproximación es truncar la estadística en grupos de tres caracteres, y calcular s_3. (Shannon estima en este caso $s_3 \sim 3.1$.) Dependiendo de la computadora, la estimación de $p_3(x_1, x_2, x_3)$ mediante el censo de todas las ternas de caracteres puede llevar de unos segundos a varios minutos; se recomienda hacer pruebas con los archivos más cortos.

Si en lugar de *Moby Dick*, se trabaja con un texto finito generado al azar, usando los 27 símbolos disponibles, 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, en la práctica, las probabilidades se infieren a partir de una muestra finita.) Más aún, si los caracteres son generados con distribución uniforme, estos valores deben ser cercanos al valor teórico $s_{\max}$ de los 27 símbolos.

k) 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. También podría ver qué pasa si genera un texto al azar donde cada símbolo ocurra con una frecuencia promedio igual a la que se observa en el libro.)

Los resultados anteriores dependen de la estimación que se haga de las probabilidades $p_1(x_1)$, $p_2(x_1, x_2)$ y $p_3(x_1, x_2, x_3)$. Para 27 símbolos, el número máximo de pares de caracteres es 729, y el de ternas es 19 683, aunque, naturalmente, no todas las combinaciones son lícitas dentro de un idioma determinado. En un texto como *Moby Dick*, con cerca de un millón de caracteres y una muestra equiparable de pares y ternas de caracteres, es razonable esperar que tanto $p_1(x_1)$ como $p_2(x_1, x_2)$ puedan estimarse correctamente. En el caso de las ternas de caracteres, la proporción entre resultados posibles y el número de muestras es la misma que en una encuesta por "sí" o por "no" hecha sobre 100 personas. El

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

resultado fácilmente podría variar si se encuestaran 1000 o $10\,000$ personas. (En verdad, el número de ternas registradas en inglés con una frecuencia mayor a 1 por millón es ~ 8000, sin contar el espacio en blanco como caracter; *Moby Dick* contiene unas 5000 ternas diferentes, contando el espacio en blanco.)

l) Considere la versión en inglés. Sea N el número total de caracteres y $L(n)$ la fracción del texto que comprende los primeros n caracteres, de modo que el libro completo es $L(N)$. Grafique en función de n los valores de $s_{1,2,3}$ correspondientes a $L(n)$; es decir, estimando las probabilidades sólo a partir de las versiones parciales del libro. No es necesario que n vaya de uno en uno. Tratándose de un libro con un millón de caracteres puede tomarse n de mil en mil, redondeando N en un múltiplo de mil. (También puede graficar $s_{1,2,3}$, para secciones disjuntas de, digamos, 50 000 caracteres, para ver si hay alguna tendencia que aparte el texto de la homogeneidad estadística.)

Idealmente, a medida que aumenta n las funciones s_i calculadas en base a $L(n)$ deberían acercarse a valores constantes. De la rapidez con que esto ocurre depende el grado de confianza en la estimación de las probabilidades a partir de la muestra finita representada por *Moby Dick*, y en la hipótesis de homogeneidad estadística.[‡].

m) Otra medida para saber si el tamaño de la muestra es suficiente para hacer estadística consiste en graficar, en función de n, el número de tipos de letras, pares y ternas de caracteres con probabilidad distinta de cero. Con n de unas pocas decenas el número de letras rápidamente alcanza el máximo de 27. En cambio, la variedad de pares o ternas no necesariamente alcanzará un valor límite en la muestra acotada que representa el libro. Grafique en función de n el número de pares y ternas de letras con probabilidad distinta de cero. ¿El autor agota rápidamente las posibilidades o se observa un crecimiento sostenido a lo largo del libro? ¿Tendría sentido calcular, basándose en este solo libro, el valor de s_4? (Este punto puede considerarse a la luz de lo dicho más arriba acerca del número y frecuencia de las ternas en inglés.)

n) A modo de comparación, repita los dos ítems anteriores para un texto de la misma longitud N que *Moby Dick* pero generado al azar, i) usando los mismos 27 símbolos con distribución uniforme, ii) usando sólo las tres letras (a, b, c) con distribución uniforme, y iii) usando sólo las tres letras (a, b, c) con las probabilidades que da Shannon en uno de sus ejemplos[§]:

x_1	$p_1(x_1)$
a	$\frac{9}{27}$
b	$\frac{16}{27}$
c	$\frac{2}{27}$

$p_2(x_1, x_2)$	a	b	c
a	0	$\frac{4}{15}$	$\frac{1}{15}$
b	$\frac{8}{27}$	$\frac{8}{27}$	0
c	$\frac{1}{27}$	$\frac{4}{135}$	$\frac{1}{135}$

z) Por último, 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.

[‡]Un escritor podría usar palabras cada vez más cortas, pasar de un vocabulario complejo o a otro más llano, o, en un caso extremo, volverse completamente predecible, como en la película *El resplandor*.

[§]C.E. Shannon, W. Weaver, *The Mathematical Theory Of Communication*, The University Of Illinois Press, 1964 (pág. 41).