Guía 3: Entropía e información

Termodinámica Avanzada - 2°C 2020

En estas notas nos vamos a concentrar en resolver los ejercicios 1 y 5b de la guía 3. Dichos ejercicios tratan de procesos que involucran la hipótesis Markoviana y por lo tanto la idea será, primero, hablar un poco sobre dichos procesos como así también listar y recordar los conceptos y ecuaciones que nos serán necesarias para estos problemas (y que servirán para los demás ejercicios de la guía también).

Índice

---

1. Probabilidades y Markov

1.1. Probabilidades, procesos estocásticos y Markov

Para hablar de *procesos de Markov*, primero debemos situarnos en un concepto más amplio que es el de **procesos estocásticos**. Dichos procesos son aquéllos donde las variables son aleatorias; he aquí que muchas veces a estos procesos también se los conoce como **procesos aleatorios**. En general las variables dependen de un parámetro real, donde en honor a Chronos vamos a considerar que se trata del tiempo t. De esta forma, denotaremos como

$$p(x_1, t_1) \tag{1}$$

a la densidad de probabilidad de que la variable aleatoria $X(t)$ tome el valor x_1 a tiempo t_1 (fijo). Asimismo,

$$p(x_1, t_1; x_2, t_2) \tag{2}$$

será la densidad de probabilidad de que la variable aleatoria $X(t)$ tome el valor x_1 a tiempo t_1 y el valor x_2 a tiempo t_2 (con t_1 y t_2 fijos). Así, en general, denotaremos a la probabilidad **conjunta** para n tiempos como

$$p(x_1, t_1; x_2, t_2; \ldots; x_n, t_n) \tag{3}$$

Algunas propiedades que cumplen estos procesos son las siguientes:

(Leonardo) Permutación

$$p(x_1, t_1; x_2, t_2; \ldots; x_n, t_n) = p(x_{i_1}, t_{i_1}; x_{i_2}, t_{i_2}; \ldots; x_{i_n}, t_{i_n}) \tag{4}$$

donde $i_1, i_2, \ldots, i_n$ es una permutación de $1, 2, \ldots, n$

(Rafael) Propiedad de compatibilidad (*marginalización*)

$$p(x_1, t_1; \ldots; x_j, t_j) = \int dx_{j+1}\, p(x_1, t_1; \ldots; x_j, t_j; x_{j+1}, t_{j+1})\,, \quad j < n \tag{5}$$

(Miguel Ángel) Probabilidad condicional

$$p(x_1, t_1; \ldots; x_n, t_n) = p(x_{k+1}, t_{k+1}; \ldots; x_n, t_n | x_1, t_1; \ldots; x_k, t_k)\, p(x_1, t_1; \ldots; x_k, t_k) \tag{6}$$

donde usamos la convención de que $P(A|B)$ es la probabilidad de que ocurra A **dado que** ocurrió B.

(Donatello) Memoria del proceso

- *Proceso completamente aleatorio* $\longrightarrow$ no posee memoria:

$$p(x_1, t_1; \ldots; x_n, t_n) = p(x_1, t_1) \ldots p(x_n, t_n) \tag{7}$$

  o, análogamente

$$p(x_n, t_n | x_1, t_1; \ldots; x_{n-1}, t_{n-1}) = p(x_n, t_n) \tag{8}$$

  Es decir que en este tipo de procesos, las variables son independientes.

- *Proceso de Markov* $\longrightarrow$ las variables se ven influenciadas por la variable contigua

$$p(x_n, t_n | x_1, t_1; \ldots; x_{n-1}, t_{n-1}) = p(x_n, t_n | x_{n-1}, t_{n-1}) \tag{9}$$

A partir de esta última propiedad, y para el caso de un proceso de Markov, es interesante notar que podemos escribir la probabilidad conjunta $p(x_1, t_1; \ldots; x_n, t_n)$ en término de las probabilidades para 1 y 2 tiempos, $p(x, t)$ y $p(x, t|x', t')$ respectivamente. De esto trata el problema 1 de la guía que es el que veremos a continuación.

1.2. Ejercicio 1

Antes de comenzar a resolver el ejercicio, hagamos un pequeño abuso de notación para que no se vuelva densa la lectura. Vamos a llamar

$$(x_j, t_j) \equiv X_j \qquad \Longrightarrow \qquad p(x_1, t_1; x_2, t_2; \ldots; x_n, t_n) \equiv p(X_1, X_2, \ldots, X_n) \tag{10}$$

De esta manera la probabilidad conjunta de n tiempos - por definición - la podemos escribir según

$$p(X_1, \ldots, X_n) = p(X_2, \ldots, X_n|X_1)p(X_1) \tag{11}$$

A su vez,

$$p(X_2, \ldots, X_n|X_1) = p(X_3, \ldots, X_n|X_1, X_2)p(X_2|X_1) \tag{12}$$

Por si esto no es evidente, una forma sencilla de ver esto es la siguiente: si llamamos

$$A = X_3, \ldots, X_n \qquad B = X_2|X_1 \tag{13}$$

entonces reemplazando en la definición de probabilidad condicional

$$p(A|B) = \frac{p(A, B)}{p(B)} \tag{14}$$

obtenemos que

$$p(X_3, \ldots, X_n|X_2|X_1) \equiv p(X_3, \ldots, X_n|X_2, X_1) = \frac{p(X_3, \ldots, X_n, X_2|X_1)}{p(X_2|X_1)} \tag{15}$$

en donde si ordenamos adecuadamente los X_j obtenemos la ecuación (12) 

Usando ahora que el proceso es de Markov

$$p(X_2, \ldots, X_n|X_1) = p(X_3, \ldots, X_n|X_2)p(X_2|X_1) \tag{16}$$

Si reemplazamos en la probabilidad conjunta de n tiempos, ec.(11), obtenemos pues

$$p(X_1, \ldots, X_n) = p(X_3, \ldots, X_n|X_2)\, p(X_2|X_1)\, p(X_1) \tag{17}$$

Repitiendo el proceso, obtenemos que

$$p(X_1, \ldots, X_n) = p(X_4, \ldots, X_n|X_3)\, p(X_3|X_2)\, p(X_2|X_1)\, p(X_1) \tag{18}$$

el cual haciendolo para los n términos pertinentes, llegamos a que

$$\boxed{p(X_1, \ldots, X_n) = p(X_n|X_{n-1})\, p(X_{n-1}|X_{n-2}) \ldots p(X_2|X_1)\, p(X_1)} \tag{19}$$

Asimismo, otra manera de arribar a la misma ecuación puede ser usando la probabilidad condicional al igual que en la ec. (11), pero en vez de separar la variable X_1, separamos la variable X_n (es decir, hacemos el procedimiento inverso). Esto es

$$p(X_1, \ldots, X_n) = p(X_n|X_1, \ldots, X_{n-1})\, p(X_1, \ldots, X_{n-1}) \tag{20}$$

A su vez, hacemos lo mismo para la probabilidad conjunta de $n-1$ tiempos y resulta

$$p(X_1, \ldots, X_n) = p(X_n|X_1, \ldots, X_{n-1})\, p(X_{n-1}|X_1, \ldots, X_{n-2})\, p(X_1, \ldots, X_{n-2}) \tag{21}$$

Haciendo lo mismo para las variables restantes, obtenemos que

$$\begin{aligned} p(X_1, \ldots, X_n) = p(X_n|X_1, \ldots, X_{n-1})\, p(X_{n-1}|X_1, \ldots, X_{n-2}) \ldots \\ \ldots p(X_3|X_1, X_2)\, p(X_2|X_1)\, p(X_1) \end{aligned} \tag{22}$$

Usando que el proceso es de Markov, obtenemos al igual que (19)

$$p(X_1, \ldots, X_n) = p(X_n|X_{n-1})\, p(X_{n-1}|X_{n-2}) \ldots p(X_2|X_1)\, p(X_1) \tag{23}$$

Como útima cuestión, cabe remarcar que para modelar ciertos sistemas, muchas veces si el sistema se comporta de acuerdo a un proceso de Markov se dice que se tiene una *cadena de Markov*. A partir de este último resultado (o a partir de la ec.(19)) es evidente por qué se lo suele denotar así.

1.3. Markov reloaded

Si volvemos a la definición de *Proceso de Markov*, ec.(9), notemos que la misma dice que la variable se ve influenciada por la variable **contigua**. No anterior, sino contigua. A pesar que la ec. (9) significa que inferencias sobre el estado actual del proceso depende solo del pasado inmediato, en realidad también vale lo recíproco: inferencias sobre el estado **actual** del proceso dependen solo del conocimiento de su **futuro** inmediato. Esto es importante notarlo ya que a priori, la ecuación (9) pareciera distinguir una dirección privilegiada del tiempo cuando en realidad no es así. La versión menos filosófica y más matemática sería

$$p(X_j|X_{j+1}, X_{j+2}, \ldots, X_{j+m}) = p(X_j|X_{j+1}) \tag{24}$$

donde implícitamente estamos diciendo que $t_j < t_{j+1} < \ldots < t_{j+m}$. Si pensamos en el **Teorema de Bayes**, esto no debería sorprendernos ya que lo que estamos diciendo - en esencia - es que

$$p(A|B)p(B) = p(B|A)p(A) \tag{25}$$

Un esbozo de la demostración[1] se puede ver a partir de lo siguiente: partiendo del Teorema de Bayes decimos que

$$p(X_1, \ldots, X_n) = p(X_1|X_2, \ldots, X_n)p(X_2, \ldots, X_n) \tag{26}$$

Si utilizamos nuevamente el Teorema de Bayes para escribir la probabilidad condicional del lado derecho y hacemos uso que el proceso es de Markov, llegamos a que

$$p(X_1, \ldots, X_n) = p(X_3, \ldots, X_n|X_2)p(X_1, X_2) \tag{27}$$

Nuevamente utilizando Bayes para escribir la probabilidad condicional $p(X_1, X_2)$ en término de las condicionales, y uniendo con la probabilidad condicional $p(X_3, \ldots, X_n|X_2)$, llegamos a que

$$p(X_1|X_2, \ldots, X_n) = p(X_1|X_2) \tag{28}$$

donde también usamos Bayes para la probabilidad conjunta del lado izquierdo $p(X_1, \ldots, X_n)$.

[1] La demostración explícita queda a gusto de cada quien.

Antes de pasar al otro problema hay un último dato de color - digno de Bolzano - que involucra el uso de este último resultado, y es el siguiente:

"*Si observamos al sistema en t_1 en el estado x_1, y en t_n en el estado x_n donde no conocemos los estados intermedios, entonces podemos inferir el camino que siguió el sistema entre ambos estados de acuerdo a las probabilidades de transición.*"

Dicho y escrito de otro modo, este enunciado es

$$p(X_k|X_1, X_n) = \frac{p(X_n|X_k)p(X_k|X_1)}{p(X_n|X_1)} \tag{29}$$

con $t_1 < t_k < t_n$. La generalización inmediata resulta

$$p(X_2, \ldots, X_{n-1}|X_1, X_n) = \frac{p(X_n|X_{n-1})p(X_{n-1}|X_{n-2}) \cdots p(X_3|X_2)p(X_2|X_1)}{p(X_n|X_1)} \tag{30}$$

2. Entropía, Información y Markov

2.1. Introducción y consideraciones

La clase pasada habíamos definido la entropía asociada a los M mensajes de longitud N (es decir, N caracteres) emitidos por una fuente según

$$S_N = -\sum_{r=1}^{M_N} p_N(r) \log p_N(r) \tag{31}$$

donde $p(r)$ es la probabilidad del mensaje r; notar que r indica la suma sobre los caracteres. A su vez, si en vez de calcular la entropía **por mensaje** queremos calcular la entropía **por caracter**, podemos definir

$$s = \lim_{N\to\infty} \frac{S_N}{N} \tag{32}$$

que es la entropía de cada caracter.

Si los mensajes tienen N caracteres, entonces $r = (x_1, \ldots, x_N)$ y por lo tanto podemos reescribir la sumatoria según

$$\sum_{r=1}^{M_N} \quad \longrightarrow \quad \sum_{x_1=1}^{M} \sum_{x_2=1}^{M} \cdots \sum_{x_N=1}^{M} \tag{33}$$

$$p_N(r) \quad \longrightarrow \quad p_N(x_1, x_2, \ldots, x_N) \tag{34}$$

Solamente para tratar de entender un poco mejor todos los índices M, N, r, etc, veamos algunos casos sencillos: consideremos que se emiten mensajes con $N = 1$ y $N = 2$ caracteres. Las entropías resultan entonces

$$S_1 = -\sum_{x_1=1}^{M} p_1(x_1) \log p_1(x_1) \tag{35}$$

$$S_2 = -\sum_{x_1=1}^{M} \sum_{x_2=1}^{M} p_2(x_1, x_2) \log p_2(x_1, x_2) \tag{36}$$

A partir de estas ecuaciones es evidente lo siguiente:

— El subíndice de la probabilidad p_k hace alusión a la cantidad de caracteres x_k. Esto quiere decir que p_1 indica la probabilidad de **un solo** caracter, como así análogamente p_2

indica la de **dos** caracteres. En cierto modo es redundante, ya que $p(x_1, \ldots, x_k)$ consistiría en la probabilidad de los k caracteres y se da por entendido. Para no cargar la notación, de ahora en más omitiremos el subíndice.

— De forma similar, el subíndice de la entropía S_k denota la entropía asociada a un mensaje de k caracteres. Como en el primer caso tenemos un solo caracter x_1, entonces $S \equiv S_1$. Asimismo, al tener dos caracteres x_1, x_2, entonces $S \equiv S_2$.

Lo importante a notar acá es que la entropía (y la probabilidad) **no depende** de **cuáles** son los caracteres, sino solamente de la **cantidad** de caracteres. Esto es,

$$S_3 = -\sum_{x_1,x_2,x_3} p(x_1, x_2, x_3) \log p(x_1, x_2, x_3) = -\sum_{x_i,x_j,x_k} p(x_i, x_j, x_k) \log p(x_i, x_j, x_k) \tag{37}$$

Este resultado que a priori pareciera no decir mucho, es clave a la hora de **marginalizar**:

$$\sum_{x_1,\ldots,x_N} p(x_1, \ldots, x_N) \log p(x_i, x_j, x_k) = \sum_{x_i=1}^{M} \sum_{x_j=1}^{M} \sum_{x_k=1}^{M} p(x_i, x_j, x_k) \log p(x_i, x_j, x_j) \tag{38}$$

y por lo tanto este término no es otra cosa más que S_3.

Para tratar de entender y ver un poco mejor esto veamos el caso más sencillo y más general posible: vamos a considerar que los caracteres no están correlacionados; es decir, consideramos un proceso *completamente aleatorio* de emisión de mensajes, en donde la probabilidad se podrá escribir entonces como

$$p(x_1, \ldots, x_N) = p(x_1) \ldots p(x_N) \tag{39}$$

y veamos cómo se escribe la entropía S_N. Las cuentas y los pasos explícitos para llegar a la expresión de la entropía quedan como ejercicio (haganlo, es para entregar ), sin embargo no es muy difícil elucubrar cuál será el resultado. Dado que las probabilidades son independientes y que la entropía no depende de qué caracter estemos viendo sino solo de la cantidad de caracteres, si tenemos N términos que son independientes entonces

$$S_N = N S_1 \tag{40}$$

La continuación inmediata de este resultado es estudiar qué pasaría si en vez de que los caracteres no estén correlacionados, ahora hay correlación de a pares. La parte b del ejercicio 5 trata de esto, y es estudiar qué sucede con la entropía donde ahora consideramos que hay correlación entre los caracteres de acuerdo a un proceso de Markov.

2.2. Ejercicio 5

Del problema 1, ec.(19), sabemos que

$$p_N(r) = p(x_1, \ldots, x_N) = p(x_1) \prod_{i=2}^{N} p(x_i | x_{i-1}) \tag{41}$$

Por lo tanto la entropía resulta

$$S_N = -\sum_{x_1,\ldots,x_N}^{M} p(x_1, \ldots, x_N) \left[\sum_{i=2}^{N} \log p(x_i | x_{i-1}) + \log p(x_1) \right] \tag{42}$$

Notemos que la entropía está definida en término de las probabilidades **conjuntas** y no de las probabilidades **condicionales**, por lo tanto la idea a continuación es reescribir el término dentro del primer logaritmo para que aparezcan las probabilidades conjuntas y no las condicionales. Esto es,

$$S_N = -\sum_{i=2}^{N}\sum_{\{x\}} p(x_1,\ldots,x_N)\Big[\log p(x_i,x_{i-1}) - \log p(x_{i-1})\Big] - \sum_{\{x\}} p(x_1,\ldots,x_N)\log p(x_1) \quad (43)$$

en donde $\{x\}$ en la sumatoria indica la suma sobre todos los caracteres x_j. Veamos que efectivamente los términos que aparecen en S_N no son otra cosa más que S_1 y/o S_2:

- $-\sum_{\{x\}} p(x_1,\ldots,x_N)\log p(x_1) \quad \longrightarrow \quad$ dentro del logaritmo se encuentra **nada más** que $p(x_1)$, por lo tanto podemos sumar sobre todos los otros caracteres x_j con $j \neq 1$; es decir, vamos a **marginalizar**

$$-\sum_{\{x\}} p(x_1,\ldots,x_N)\log p(x_1) = -\sum_{x_1} p(x_1)\log p(x_1) = S_1 \quad (44)$$

- $-\sum_{i=2}^{N}\sum_{\{x\}} p(x_1,\ldots,x_N)\log p(x_i,x_{i-1}) \quad \longrightarrow \quad$ al igual que antes vamos a marginalizar sobre las variables que no se encuentran dentro del logaritmo

$$-\sum_{i=2}^{N}\sum_{\{x\}} p(x_1,\ldots,x_N)\log p(x_i,x_{i-1}) = -\sum_{i=2}^{N}\sum_{x_i}\sum_{x_{i-1}} p(x_{i-1},x_i)\log p(x_i,x_{i-1}) \quad (45)$$

De acuerdo a la definición de entropía de un mensaje, S_N, esta última igualdad no es otra cosa más que

$$-\sum_{i=2}^{N}\sum_{x_i}\sum_{x_{i-1}} p(x_{i-1},x_i)\log p(x_i,x_{i-1}) = \sum_{i=2}^{N} S_2 = (N-1)S_2 \quad (46)$$

en donde en el última igualdad sumamos sobre i, $i = 2,\ldots,N$

- $\sum_{i=2}^{N}\sum_{\{x\}} p(x_1,\ldots,x_N)\log p(x_{i-1}) \quad \longrightarrow \quad$ marginalizando y haciendo el mismo tratamiento que el caso anterior, llegamos a que

$$\sum_{i=2}^{N}\sum_{\{x\}} p(x_1,\ldots,x_N)\log p(x_{i-1}) = -(N-1)S_1 \quad (47)$$

Juntando estos tres resultados llegamos finalmente a que la entropía asociada a un mensaje de longitud N en el caso en que el proceso de emisión de caracteres sea de Markov es

$$\boxed{S_N = (N-1)(S_2 - S_1) + S_1 = N(S_2 - S_1) - S_2 + 2S_1} \quad (48)$$

La *entropía por caracter* $s \equiv s_2$ resulta entonces

$$s_2 = \lim_{N\to\infty} \frac{S_N}{N} \quad \Longrightarrow \quad \boxed{s_2 = S_2 - S_1} \tag{49}$$

Resumiendo, obtuvimos que las entropías para los casos en que los caracteres no están correlacionados y el caso donde hay correlación de a pares (Markov) resultaban en

$$S_N = NS_1\,, \qquad S_N = (N-1)(S_2 - S_1) + S_1 \tag{50}$$

La generalización restante es considerar el caso donde hay correlación de a n términos. El resultado que generaliza esto (y el cual se reduce obviamente a las ec.(50) para $n = 1$ y $n = 0$ bajo la definición $S_0 = 0$) es

$$S_N = (N-n)(S_{n+1} - S_n) + S_n \tag{51}$$