

FÍSICA CLÁSICA Y COMPUTACIÓN

Vicente Moret Bonillo

IEEE Senior Member

Apuntes Doctorado

2005

© Vicente Moret Bonillo

FÍSICA CLÁSICA Y COMPUTACIÓN

Las líneas que siguen pretenden razonar sobre las similitudes de la computación y de la física. La computación puede ser tratada, al menos teóricamente, como un sistema físico más y, desde luego, todas las limitaciones que las leyes fundamentales de la física imponen a los sistemas físicos aplican igualmente a la computación. Dicho de otro modo: trataremos de estudiar algunos aspectos que a mí me parecen interesantes de la computación desde la óptica de la física, con los métodos y las herramientas de la física. Razonaremos por analogía, y buscaremos hipótesis, algunas de ellas ciertamente agresivas, para tratar de encontrar dónde están los límites teóricos de la computación, y qué consecuencias pueden derivarse de la existencia de dichos límites. No hablaremos de lo que es computable o no, sino de la computación en sí misma: No seguimos los planteamientos de Gödel, sino los de Shannon, Bennet y Feynman.

De este último, Feynman, copiamos también su estilo y su forma de plantearse problemas. Y tengo que decir que he aprendido mucho tras la lectura de sus obras. Brillante y original, Feynman me ha enseñado que lo importante es trabajar sobre cualquier tema con el único objetivo de comprenderlo. Como él solía decir "... de acuerdo, no he sido el primero pero por lo menos lo entiendo".

Para ser completamente honrado: estas líneas han sido escritas para tratar de consolidar en mi cabeza gran parte de las reflexiones que Feynman publica en su libro "Conferencias sobre computación", texto en el que se basa gran parte del material que desarrollamos aquí.

Hablaremos algo de la física de la información, de termodinámica, de la teoría cinética de los gases, del principio de indeterminación y de la mecánica cuántica, para tratar de explicar -por ejemplo- aspectos relacionados con la energía mínima de una computación, o cómo podemos medir la cantidad de información de un mensaje, o dónde están los límites teóricos del aprendizaje. Ninguna de tales cuestiones es original: muchos otros -mucho más informados y mucho más listos que yo- se las han planteado y las han resuelto, pero a mí me ha parecido divertido volver sobre ellas y, de paso, intentar entenderlas.

Algunos de los temas que trataremos van a requerir un esfuerzo nada desdeñable de abstracción. Esto es siempre interesante, ya que suele fomentar discusiones animadas, que espero se produzcan. Otros, por el contrario, serán muy sencillos de conceptualizar. Tan sólo propondremos cambios de enfoque y puntos de vista algo diferentes de los habituales. También será inevitable hacer algo de matemáticas, pero poco... lo estrictamente necesario para poder trabajar con las ideas que iremos desarrollando aquí.

Y ya sin más –como decía Lewis Carroll- comencemos por el principio y cuando lleguemos al final paremos.

1. BUSCANDO AL BIT

¿Qué es un bit? El diccionario de la Real Academia de la Lengua Española (RAE) define el término bit del siguiente modo: “Unidad de medida de información equivalente a la elección entre dos posibilidades igualmente probables”. La definición, estrictamente cierta, no deja de ser un poco espesa –al menos si uno se la encuentra de sopetón-. Perdamos un instante en analizarla: ... unidad de medida de información... (hasta aquí todo va bien)... equivalente a la elección entre dos posibilidades igualmente probables... (aquí la cosa ya no va tan bien). El que dos posibilidades sean igualmente probables está bastante claro. Por ejemplo, si tenemos A y tenemos B –y sólo tenemos A y B-, la probabilidad de A, $p(A)$, es igual a la probabilidad de B, $p(B)$. Consideremos que $X \equiv \{p(A) = p(B)\}$, es decir denotamos por X a la circunstancia de que los sucesos A y B son equiprobables. Según esto, la definición de bit que propone la RAE podría traducirse del siguiente modo:

BIT \equiv Elección $\{A,B/X\}$

Para mí, la dificultad de la definición de la RAE reside en el término “Elección”... La interpretación es confusa.

Otra definición que encontramos, esta vez en el Diccionario de Computación publicado por McGraw-Hill en 1991, establece que un bit es una “unidad de información que equivale a una decisión binaria o a la designación de uno de los dos valores o estados

posibles e igualmente probables de un medio empleado para almacenar o transferir información”. Esta definición, que es casi la misma que la de la RAE (aunque cambia el término –elección- por el de –decisión-), menciona ya explícitamente el vínculo del bit con la información y con los medios empleados para almacenarla o transferirla. En mi opinión sigue siendo una definición ambigua y poco precisa.

Una segunda acepción del término bit, que aparece en el diccionario arriba mencionado, establece que “bit es una unidad adimensional de la capacidad de almacenamiento que expresa la capacidad de almacenamiento como el logaritmo en base 2 del número de estados posibles del dispositivo”. Aquí entramos ya en el terreno de las definiciones “por decreto” (perfectamente válidas, por otra parte). La única pega que encontramos es ...¿a qué dispositivo nos estamos refiriendo? Por otra parte: ¿por qué el logaritmo en base 2 y no otro? ¿Hay alguna justificación o es por convenio?... Volveremos más adelante sobre esta cuestión.

Dejando ya las fuentes documentales académicas, consulté a varios de mis compañeros –todos ellos profesionales de las ciencias de la computación- sobre cómo definirían ellos el concepto de bit. Les rogué también que trataran de olvidar las definiciones convencionales y me diesen su propia definición. A continuación transcribo algunas de sus respuestas:

- Representación matemática de los dos estados posibles de un interruptor – encendido y apagado-, que utiliza el sistema de numeración en base 2 por poseer éste el mismo número de estados -0 y 1-
- Unidad mínima (de información) de un alfabeto de dos símbolos
- La unidad más pequeña de información en una máquina
- Dígito o guarismo binario (base 2), i.e., 0 ó 1
- Unidad de numeración binaria que puede tomar dos valores asociados, 0 ó 1

- La mínima cantidad de información que puede ser almacenada y transmitida dentro de un sistema informático
- Es un dato con dos posibles valores (o informaciones, según se quiera ver)

Todas las definiciones anteriores, incluidas las académicas, pueden ser englobadas en tres categorías: (a) aquéllas que enfatizan aspectos conceptuales, (b) las que se centran en la idea de unidad de información, y (c) las que resaltan el carácter binario del bit.

Para mí está claro que el bit existe por definición, del mismo modo que el color rojo es rojo porque los científicos se han puesto de acuerdo en que de tal longitud de onda del espectro de radiación electromagnética a tal otra, el color es precisamente rojo. Lo curioso del caso es que la gente de la calle ya sabía lo que era el color rojo mucho antes de que los científicos se pusieran de acuerdo. Así, debe existir algo esencial e intuitivo en los colores de forma que éstos sean reconocidos sin necesidad de recurrir a su definición o a su caracterización formal. Y esto es precisamente lo que vamos a explorar: ¿hay algo que justifique el concepto de bit más allá de su “definición por decreto”?, ¿hay algo que nos permita identificar al bit como una unidad de información?, ¿hay algo que nos permita establecer, sin temor a equivocarnos, la naturaleza binaria del bit?... trataremos de encontrar respuestas a estas cuestiones, pero antes vamos a abrir un pequeño paréntesis.

En casi todas las ciencias es posible razonar de dos formas: desde los datos hacia las conclusiones –lo que en inteligencia artificial, por ejemplo, configuraría un proceso hacia adelante o progresivo-, o desde las hipótesis hacia los datos –lo que, otra vez, en inteligencia artificial configuraría un proceso hacia atrás o regresivo-. En ambos casos es necesario el conocimiento de que disponemos sobre el dominio, lo que cambia es la forma de utilizarlo. Así, en un razonamiento progresivo partimos de un conjunto de datos y utilizamos nuestro conocimiento para obtener conclusiones válidas. Por ejemplo, si yo sé que mi conocimiento Θ está constituido por 5 axiomas:

$$\Theta = \{(1), (2), (3), (4), (5)\}$$

donde:

- (1) $A \rightarrow B$
- (2) $B \text{ y } C \rightarrow D$
- (3) $D \text{ y } E \text{ y } F \rightarrow H$
- (4) $A \text{ y } Z \rightarrow V$
- (5) $V \text{ y } X \text{ y } G \rightarrow W$

y además sé que el conjunto de los datos de mi problema es:

$$\Delta = \{A, C, E, F\}$$

entonces podemos concluir B, D y H sin más que aplicar los axiomas (1), (2) y (3) sobre mi conjunto de datos Δ . Ésta es una forma de razonar progresivamente. Por el contrario, si yo quiero razonar de forma regresiva tengo que asumir una hipótesis de trabajo, digamos D, y utilizar mis axiomas para buscar información que, a través de un proceso evocativo, y contando con las fuentes de información oportunas, me permita confirmar mi hipótesis. En nuestro ejemplo, para confirmar D necesito utilizar el axioma (2), pero en el antecedente de (2) figuran B y C. C está en Δ , pero B no, y lo necesito para confirmar mi hipótesis inicial de trabajo D. B se convierte en una nueva hipótesis de trabajo. Para confirmar B utilizo el axioma (1) que necesita el dato A para poder ser aplicado. Pero A está en Δ , por lo que (1) es aplicable y B es cierto. Una vez sabemos que B es cierto ya sabemos que todos los elementos del antecedente de (2) son ciertos, por lo que demostramos que D es cierto. D deja de ser una hipótesis de trabajo para convertirse en un hecho demostrado. Nótese que el conjunto de axiomas Θ y el conjunto de datos Δ son iguales en ambos casos, pero el resultado final no es el mismo. Así, si Ψ es la “información” final que obtengo después de razonar con mis datos y con mis axiomas:

$$\Psi_{\text{progresivo}} = \{A, C, E, F, B, D, H\}$$

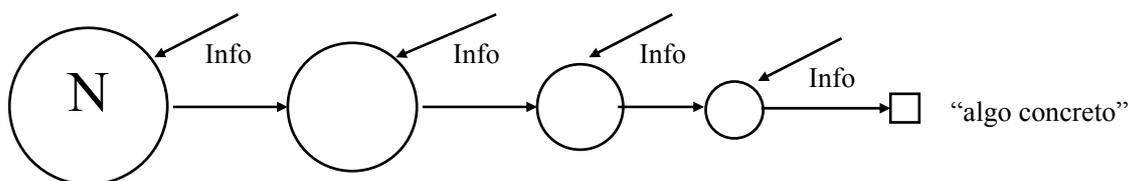
$$\Psi_{\text{regresivo}} = \{A, C, E, F, B, D\}$$

La diferencia no está en la información utilizada, sino en la forma de utilizar esta información. Algo parecido -aunque sólo sea de lejos- ocurre en física cuando, por ejemplo, estudiamos electromagnetismo. Uno puede empezar con Coulomb y llegar a las ecuaciones de Maxwell, o postular las ecuaciones de Maxwell y llegar a demostrar la ley de Coulomb. En este caso también razonamos de forma distinta.

... pero ¿qué tiene que ver esto con el bit? Pues seguramente nada. Lo único que he querido resaltar es que la información y los conocimientos pueden utilizarse de forma diferente y producir resultados distintos según nos convenga (y si no se lo creen pregúntenselo a los políticos)... ¡Además es divertido!

Cerremos ahora el paréntesis y volvamos con nuestro amigo el bit. Y supongamos que no tenemos ni idea de lo que es un bit; es decir, que no lo hemos definido. ¿Seremos capaces de justificar de forma razonada su existencia o, incluso, de “demostrarla”?

Iremos un poquito más allá. Supongamos que tenemos un problema del cual inicialmente no sabemos casi nada. Concretaremos un poco ese “casi”: sabemos que tenemos un problema, que el problema tiene solución, y que la solución tiene que ser una de las N soluciones posibles en un universo cualquiera. Como aún no sabemos cuál es la solución de nuestro problema todas las $n_i \in N$ son igualmente probables (cualquier solución posible puede ser la solución). Poco a poco iremos recabando información que será utilizada para descartar algunas de las opciones iniciales de N . A medida utilizemos esta información, más cerca estaremos de la solución. Finalmente, con un poco de suerte, podremos identificarla perfectamente. La figura pretende ilustrar este proceso de “razonamiento”.



Con lo dicho anteriormente ya estamos en condiciones de formalizar la cuestión en términos matemáticos: Sean N posibilidades equiprobables a priori, que puedo expresar con “ n ” símbolos (uno por cada una de tales posibilidades, soluciones candidatas u opciones). Evidentemente debemos utilizar información para reducir la dimensionalidad de N . Supondremos que toda la información es relevante; por lo tanto, cuanta más información utilicemos más opciones inicialmente posibles descartaremos, y menor será N . Definiremos la cantidad de información Ψ al más puro estilo Shannon:

$$\Psi = k \text{Ln}(N)$$

en donde “ k ” es una constante cuyo valor está determinado por la unidad de información empleada.

Ésta es una asunción tan fuerte como la definición por decreto de nuestro amigo el bit, pero ya hemos dicho que no íbamos a inventar nada nuevo, tan sólo queremos razonar de forma distinta.

En la definición de la cantidad de información Ψ el empleo del logaritmo se justifica fácilmente considerando la existencia de dos sistemas independientes (o espacios de soluciones distintos) con N_1 y N_2 sucesos equiprobables respectivamente. Considerando al sistema globalmente, el espacio de soluciones completo será $N = N_1 \times N_2$. La situación es análoga a la que ocurre cuando tenemos 2 universos distintos e independientes, U_1 y U_2 , cada uno de ellos con un conjunto definido de elementos, por ejemplo:

$$U_1 = \{a, b, c, d\}$$

$$U_2 = \{x, y, z, t, v\}$$

Aunque no hace falta, supondremos que cada elemento de U_1 y U_2 tiene una probabilidad dada de ocurrir: $p(a)$, $p(b)$, $p(c)$, $p(d)$ para los elementos de U_1 y $p(x)$, $p(y)$, $p(z)$, $p(t)$, $p(v)$ para los elementos de U_2 . De este modo, la probabilidad conjunta de, por ejemplo “ a ” y “ x ” será:

$$p(a \wedge b) = p(a) \times p(b)$$

Pero sigamos con la justificación del logaritmo. Tenemos que hacer compatible el hecho de que $N = N_1 \times N_2$ con el hecho de que la información es una propiedad aditiva (i.e., la cantidad de información se incrementa a medida que aparecen nuevos elementos de información). Así, puesto que $\Psi = k \text{Ln}(N)$, si desarrollamos la expresión resulta que:

$$\Psi = k \text{Ln}(N) = k \text{Ln}(N_1 \times N_2) = k \text{Ln}(N_1) + k \text{Ln}(N_2) = \Psi_1 + \Psi_2$$

... y ya está. Gracias al logaritmo hemos logrado construir algo coherente y bien estructurado. Dicho de otro modo, de paso nos acabamos de cargar el “decretazo” de Shannon según el cual se define el concepto de cantidad de información.

Sigamos ahora con nuestras reflexiones. Decíamos hace ya algunas líneas que: “Sean N posibilidades equiprobables a priori, que puedo expresar con “ n ” símbolos (uno por cada una de tales posibilidades, soluciones candidatas u opciones)...” Pero, si yo tengo “ n ” símbolos ¿Cuántos estados equiprobables puedo representar? La respuesta es evidente: $N = 2^n$.

Así, si $n = 3$ (pongamos por caso: A, B, C), los N estados en principio equiprobables que yo puedo representar son:

No A, No B, No C	No A, B, C
No A, No B, C	A, No B, C
No A, B, No C	A, B, No C
A, No B, No C	A, B, C

De este modo, si utilizamos la equivalencia $N = 2^n$ en la expresión definida para la cantidad de información obtenemos que:

$$\Psi = k \text{Ln}(N) = k \text{Ln}(2^n) = k n \text{Ln}(2) = n [k \text{Ln}(2)]$$

Si yo quiero simplificar la expresión y hacer que $\Psi = n$ (recordemos que habíamos definido k como una constante cuyo valor está determinado por la unidad de información empleada), entonces:

$$k \ln(2) = 1 \rightarrow k = 1/\ln(2)$$

Volviendo a la conocida expresión $\Psi = k \ln(N)$, y sustituyendo valores:

$$\Psi = k \ln(N) = \ln(N)/\ln(2) = \log_2(N) = n$$

... et voilà: ¡Acaba de nacer el “bit”!

El que no se crea que $\ln(N)/\ln(2) = \log_2(N)$ puede hacer lo siguiente:

$$\ln(N)/\ln(2) = a \rightarrow \ln(N) = a \ln(2) \rightarrow N = 2^a \rightarrow \log_2(N) = a \log_2(2) = a$$

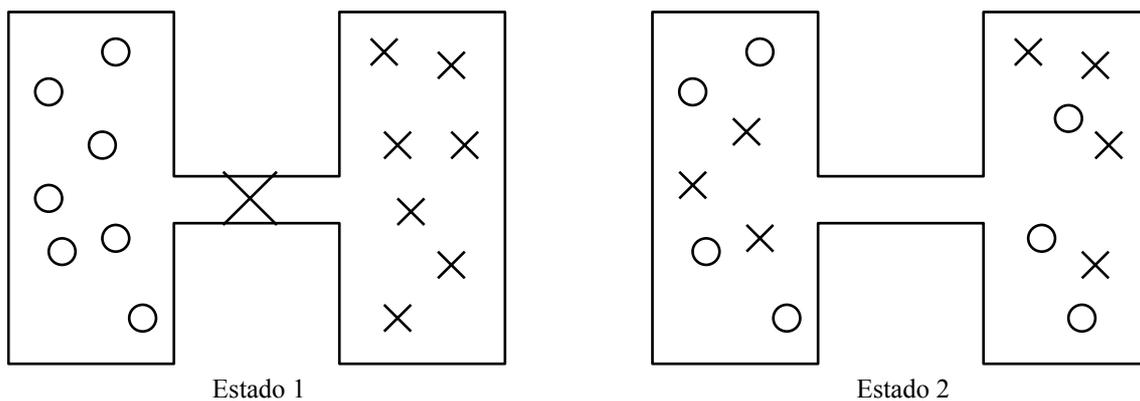
Interesante ...¿verdad?

2. ENERGÍA, ENTROPÍA E INFORMACIÓN

Vamos a plantearnos ahora una cuestión completamente distinta a la anterior, y mucho más cercana a la física tradicional: ¿Existe alguna relación entre la energía, la información y la computación? La respuesta parece trivial. Los que trabajamos con ordenadores –que hoy en día somos casi todos- sabemos que sí. Cuanto más tiempo estamos conectados (supongamos que trabajando) más información manipulamos, más se calienta el equipo, y durante más tiempo funciona el molesto ventilador del ordenador... La energía (en forma de calor, ruido, etc.) está omnipresente cuando realizamos una computación (si no se lo cree intente realizar algo tan sencillo como una suma con el ordenador desconectado y sin baterías). Esta perspectiva energética de la computación (y recordemos que computamos información) está directamente relacionada con el soporte físico de la misma. En efecto, para realizar una computación es preciso el flujo de electrones a través de cables y dispositivos, lo que genera fricción

que se disipa en forma de calor. Para evitar que este calor estropee el equipo debemos refrigerarlo, por lo que instalamos un ventilador que necesita energía para funcionar y que además genera ruido –otra forma de energía-. Además el propio ordenador necesita estar alimentado para poder trabajar con él. ¡Hay energía por todas partes!... pero no es esto lo que nos interesa aquí. Y ya hemos dicho que trataríamos de hablar de la computación en sí misma. Para ello nos acercaremos a las cuestiones energéticas de la computación desde la información que manipulamos¹, y poco a poco iremos profundizando en la cuestión.

Consideremos el sistema representado en la figura. En dicho sistema, que está constituido por dos depósitos conectados entre sí por medio de un tubo que se puede abrir o cerrar con una válvula, hay dos tipos de partículas diferentes y perfectamente distinguibles: partículas (x) y partículas (o). La figura representa al sistema en dos estados distintos. En el estado 1 todas las partículas (o) están a la izquierda, todas las partículas (x) están a la derecha, y la válvula está cerrada. En el estado 2 la válvula está abierta y ambos tipos de partículas (o) y (x) se distribuyen homogéneamente en ambos depósitos. Consideremos a nuestro sistema desde una perspectiva macroscópica: Si el número de ambas partículas es grande... ¿cómo podemos evaluar energéticamente la cantidad de información contenida en ambos estados de nuestro sistema?



Vamos a tratar de responder a esta pregunta a través de la mecánica estadística, también denominada termodinámica estadística. En mecánica estadística se define el concepto de entropía del siguiente modo: “La entropía es una medida del desorden de un sistema,

¹ Obviaremos la diferencia conceptual que existe entre información y dato. No es relevante en nuestra discusión.

igual a la constante de Boltzmann por el logaritmo natural del número de estados microscópicos correspondientes al estado termodinámico del sistema”... ¡Qué curioso! Así que de acuerdo con la definición anterior:

$$S = k \ln(N)$$

expresión en la que S es la entropía, k es la constante de Boltzmann, y N es el número de estados microscópicos correspondientes al estado termodinámico del sistema... veamos ¿a qué me suena esto? Recordamos que la cantidad de información definida por Shannon era:

$$\Psi = k \ln(N)$$

quizás la “k” y la “N” no sean las mismas en ambas expresiones (o sí, en todo caso no nos importa), pero las ecuaciones son idénticas... debe haber algún nexo de unión entre la entropía y la cantidad de información de un sistema. Sigamos razonando: si la entropía es una medida del desorden de un sistema, está claro que el estado 2 está mucho más desordenado que el estado 1. Dicho de otro modo, hay muchos más estados microscópicos posibles en el estado 2 que en el estado 1 –podemos distinguir perfectamente una partícula (x) de una partícula (o), pero todas las partículas (x) son iguales entre sí. Lo mismo podemos decir de todas las partículas (o). Por lo tanto la entropía del estado 1 es mucho menor que la del estado 2. Si N(1) es el número de microestados del estado 1, y N(2) es el número de microestados del estado 2, entonces:

$$S(1) = \text{entropía del estado 1} = k \ln[N(1)]$$

$$S(2) = \text{entropía del estado 2} = k \ln[N(2)]$$

Pero en física no tiene mucho sentido hablar de entropías absolutas; por el contrario, se trabaja con incrementos de entropía. Así:

$$\Delta S = S(2) - S(1) = k \ln [N(2)] - k \ln[N(1)] = k \ln [N(2)/N(1)]$$

En esta formulación subyace la asunción de que partimos del estado 1, abrimos la válvula, y dejamos que el sistema evolucione. De este modo las partículas se

desordenan, y la entropía del sistema aumenta, por lo que ΔS es positivo. Pero analicemos ahora qué ha pasado con nuestra información. Cuando tenemos a nuestro sistema en el estado 1 sabemos muchas cosas... sabemos que todas las partículas (o) están a la izquierda y que todas las partículas (x) están a la derecha. Abramos ahora la válvula y dejemos que el sistema empiece a evolucionar. Al principio del proceso –es decir, justo después de abrir la espita- podemos saber que “casi” todas las (o) están a la izquierda y “casi” todas las (x) están a la derecha, pero ya hay algunas partículas infiltradas en el depósito que no les toca. Estoy perdiendo información. Cuando el sistema alcanza el equilibrio (lo que –más o menos- quiere decir que, macroscópicamente, ya no evoluciona más) he perdido toda la información. De hecho, si con el sistema en el estado 1 alguien me pidiese que sacara del sistema una partícula (x) iría directamente al depósito de la derecha y sacaría una cualquiera sin fijarme, seguro de que le daba la partícula correcta. Si alguien me pidiese lo mismo con el sistema en el estado 2, iría a uno cualquiera de los depósitos, sacaría una partícula, comprobaría que es una (x) y, en caso contrario, devolvería la partícula al sistema y seguiría probando hasta que el azar quisiera darme por fin una maldita partícula (x).

Decir que es el azar quien guía el éxito de nuestra misión con el sistema en el estado 2 es lo mismo que decir que en el estado 2 nuestra información es nula. Por lo tanto, si $\Delta\Psi$ es el incremento de información cuando pasamos de estado 2 al estado 1, $\Psi(1)$ es la información asociada al estado 1, y $\Psi(2)$ es la información asociada al estado 2:

$$\Delta\Psi = \Psi(2) - \Psi(1) = -\Psi(1)$$

puesto que ya hemos dicho que podemos asumir perfectamente que nuestra información sobre el estado 2 es nula.

Recapitulemos ahora: partiendo del estado 1 y dejando que el sistema evolucione de manera espontánea hasta alcanzar la situación de equilibrio del estado 2, se produce un incremento de entropía, pero también se produce una pérdida de información. De acuerdo con nuestra formulación anterior (y olvidándonos de que la “k” y las “N”, pueden no ser las mismas), podríamos decir que la pérdida de información que se produce cuando dejamos que el sistema evoluciones de forma espontánea “equivale” a

un incremento de entropía del sistema. No obstante seremos algo más precavidos y diremos que:

$$[\Delta\Psi = -\Psi(1)] \propto \Delta S$$

expresión que interpretamos del siguiente modo: La pérdida de información que se produce al dejar que el sistema evolucione espontáneamente está relacionada con el aumento de entropía del sistema. Dejamos como ejercicio la comprobación o refutación de la equivalencia antes mencionada.

Podemos encontrar muchísimos ejemplos de lo que acabamos de establecer en la vida cotidiana. Unos más complicados que otros. Yo sugiero que visualicen lo que acabamos de comentar tratando de responder a una misma pregunta en dos escenarios diferentes:

Pregunta: ¿Dónde estarán mis calcetines azules?

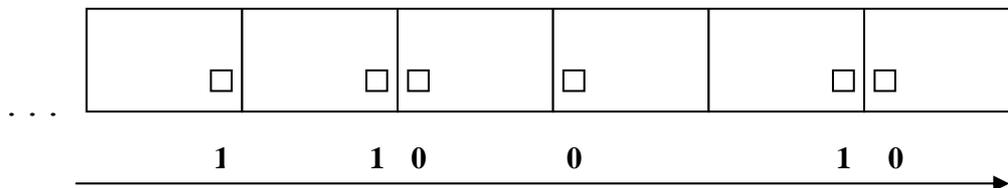
Escenario 1: Estamos buscándolos en casa de mamá

Escenario 2: Estamos buscándolos en nuestro apartamento de estudiante

Hasta ahora hemos tratado de relacionar la computación con la energía, pero la verdad es que no lo hemos conseguido del todo. A lo más que hemos llegado es a establecer un vínculo entre la entropía de un sistema y la información. De todas formas no estamos lejos: hay una función termodinámica, la energía libre, que nos relaciona la energía interna de un sistema con su entropía. Hablaremos algo más adelante sobre esta energía libre y sobre su relación con la computación. Pero antes vamos a plantearnos una interesantísima pregunta a la que Richard Feynman dedicó parte de su tiempo. La cuestión es: ¿Cuánta energía hace falta para llevar a cabo una computación? Volvemos a una cuestión anterior. Si hablamos de la relación entre el soporte físico de la computación y la computación propiamente dicha, ya habíamos visto que la energía aparece por todas partes. Más aún, precisando sobre el término computación, es claro que cuanto mayor sea la velocidad a la que esta computación es efectuada, mayor será la energía disipada en forma de calor. Al respecto, Shannon pretendía resolver el problema de la transmisión de mensajes a través de cables reales, lo cual suponía la existencia de inevitables interferencias con el mundo físico. Esto, a su vez, nos abría la posibilidad de

abordar el problema de la computación desde la física. Pero la cuestión que Feynman se plantea es otra: ¿Cuál es la energía mínima necesaria para llevar a cabo una computación? Según el propio Feynman declara, el problema es ahora de naturaleza estrictamente académica, y necesitaremos realizar un pequeño esfuerzo de abstracción para abordarlo².

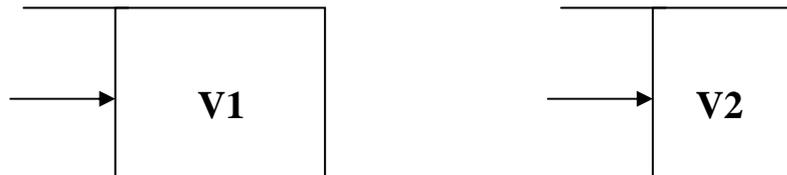
Lo primero que trataremos de hacer es visualizar desde una perspectiva física qué es el contenido de información de un mensaje. Para ello utilizaremos el esquema de la figura, que pretende ilustrar un modelo sencillo de un mensaje que está siendo enviado. En la figura podemos identificar un número, en principio no determinado, de cajas que están pegadas entre sí. En cada caja hay una partícula, y cada una de tales partículas puede estar en la parte derecha de la caja o en la parte izquierda. Por establecer un convenio, diremos que una partícula que ocupe la posición derecha es un bit 1, mientras que una partícula que ocupe una posición izquierda es un bit 0. El sistema global así definido puede considerarse como un montón de partículas que no interaccionan entre sí, y que – además- pueden ocupar distintas posiciones. Pues bien, la teoría cinética de los gases, que trabaja sobre las mismas premisas, podría sernos útil para comprender este modelo de información.



Consideremos un gas cualquiera constituido por N partículas individuales que ocupan un volumen V_1 . Supondremos también que cada partícula es esencialmente libre. Esto quiere decir que las partículas del gas no interaccionan entre sí, lo que – a su vez- significa que el tamaño de las partículas es muy pequeño si lo comparamos con las distancias a las que se encuentran unas de otras (lo cual es estrictamente cierto si trabajamos con presiones suficientemente bajas). Ahora vamos a manipular nuestro gas para tratar de encontrar alguna relación nuestro modelo de gas y el contenido de información de un mensaje.

² En esta discusión seguimos los planteamientos de Feynman publicados en “Conferencias sobre Computación, Crítica, eds., 2003.

Vamos a someter a nuestro gas a una compresión isoterma. Es decir, vamos a disminuir el volumen del gas (por ejemplo mediante un émbolo), manteniendo constante la temperatura del sistema. Esto puede conseguirse sumergiendo a nuestro sistema en un baño a temperatura constante, capaz de absorber la energía disipada durante la compresión, tal y como se ilustra en la figura.



Baño a Temperatura Constante

De este modo pasamos de un volumen 1 a un volumen 2, manteniendo constante la temperatura, pero... ¿Cuánto trabajo necesitamos realizar para comprimir el gas?

Por definición sabemos que:

$$\delta W = F \delta x$$

en donde W es trabajo, F la fuerza realizada durante el proceso, y δx es un desplazamiento pequeño del émbolo empleado para lograr la compresión.

Si decimos que p es la presión del gas, y A es la sección del émbolo, entonces:

$$F = p A \rightarrow \delta W = p A \delta x$$

pero $A \delta x = \delta V$, luego: $\delta W = p \delta V$

Por otra parte, la teoría cinética de los gases establece que:

$$p V = N k T$$

expresión en la que N es el número de partículas del gas, k es la constante de Boltzmann, y T es la temperatura del sistema que, como ya hemos mencionado, permanece constante durante el proceso.

Por lo tanto:

$$p = \frac{N k T}{V} \rightarrow \delta W = N k T \left(\frac{\delta V}{V} \right) \rightarrow W = \int_{V_1}^{V_2} N k T \left(\frac{\delta V}{V} \right) = N k T \ln \left(\frac{V_2}{V_1} \right)$$

Como tras la compresión isoterma el volumen V_2 es menor que el volumen V_1 , el trabajo neto, W , es menor que cero. Esto hay que interpretarlo considerando que el trabajo realizado en la compresión isoterma de un gas lo pierde el sistema.

Es indudable que ha habido una variación energética en nuestro sistema. De hecho, al efectuar la compresión, nuestro gas se ha tenido que calentar como consecuencia del incremento de la energía cinética de las partículas del gas tras la compresión. Si no lo ha hecho es porque hemos sumergido a nuestro sistema en un baño térmico. Sin embargo esto no siempre funciona. Para que funcione, la compresión debe efectuarse muy lentamente, de forma que ambos sistemas –partículas de gas encerradas y baño térmico- se encuentren siempre en una situación de equilibrio térmico. Pero entonces ¿adónde ha ido a parar el trabajo asociado al cambio de estado que supone la compresión isoterma de un gas desde un volumen V_1 hasta un volumen V_2 ?

La termodinámica define una importante ecuación, sobre la que volveremos más adelante, que nos permite relacionar las variables que rigen los cambios de estado de un sistema. Esta ecuación es la siguiente:

$$F = U - T S$$

en la cual el término F representa la energía libre del sistema, U es la energía total del gas –que es la suma de las energías individuales de las partículas del gas, y que no varía a lo largo de todo el proceso-, T es la temperatura del sistema (que en nuestro caso es constante por tratarse de una compresión isoterma), y S es la entropía del sistema –¡Otra vez la entropía!-

Si trabajamos con variaciones muy pequeñas, y dado que T es constante, podemos reescribir la ecuación anterior del siguiente modo:

$$\delta F = \delta U - T \delta S$$

Como $\delta U = 0$ –puesto que U no varía a lo largo del proceso-, $\delta F = - T \delta S$, y es precisamente δF la energía que pierde el sistema, que es absorbida por el baño térmico.

$$\text{De este modo: } \delta F = - \delta W = - N k T \left(\frac{\delta V}{V} \right) = - T \delta S \rightarrow \delta S = N k \left(\frac{\delta V}{V} \right)$$

Integrando ahora la expresión de la entropía entre los límites V1 y V2, que marcan respectivamente los estados inicial y final del proceso de compresión isotérmica, obtenemos que:

$$\Delta S = \int_{V_1}^{V_2} N k \left(\frac{\delta V}{V} \right) = N k \text{Ln} \left(\frac{V_2}{V_1} \right)$$

Nótese que en este caso $\Delta S < 0$, puesto que como resultado de la compresión $V_2 < V_1$. Es decir que el cambio de estado que se produce tras la compresión isotérmica de un gas, se traduce en una disminución en la entropía del sistema³. Naturalmente ello debe ser a expensas de que en algún otro sitio haya un aumento de entropía. Sobre esta nueva cuestión volveremos también algo más adelante.

Con nuestra discusión sobre la compresión isotérmica de un gas hemos conseguido, por una parte, introducir el concepto de energía en nuestros planteamientos; y por otra, expresar la variación de energía como un cambio en la entropía del sistema. Además, un poco más arriba habíamos podido relacionar la pérdida (o ganancia) de información con el aumento (o disminución) de la entropía. Trataremos ahora de afinar un poco estas

³ La expresión obtenida para ΔS corresponde a un proceso reversible. En un proceso irreversible tendríamos que sustituir el signo “=” por el signo “≥”. En la práctica todos los procesos reales son irreversibles, pero se puede conseguir una buena aproximación a la reversibilidad de un proceso procediendo de manera muy lenta, tal y como hemos discutido ya al hablar de la compresión isotérmica de nuestro gas.

ideas, y buscar un vínculo algo más fuerte entre la termodinámica estadística y la información, al mismo tiempo que nos acercamos de nuevo al concepto de bit.

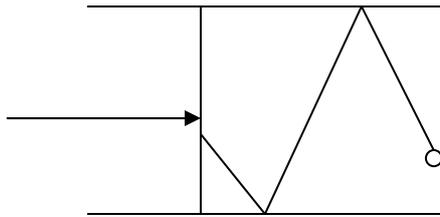
Seguiremos con el modelo de gas encerrado en un recinto sobre el que puede operar un émbolo. Pero ahora supondremos también que nuestro gas está constituido por una única partícula. En tales condiciones conceptos como la temperatura, la presión, el volumen, la energía libre o la entropía, casi pierden su sentido físico... pero hay un truco para que esto no ocurra. Basta con olvidarnos del tiempo y permitir que nuestra única partícula rebote todo cuanto quiera contra las paredes del recipiente que la contiene. Después, lo único que hay que hacer es promediar los valores de cada una de las magnitudes que nos interesan. Un ejemplo puede contribuir a aclarar la cuestión: para inferir⁴ la probabilidad de que al lanzar una moneda al aire salga “cara” (o “cruz”), puedo proceder de dos maneras diferentes:

- Lanzar muchas monedas al aire una sola vez, contar el número de “caras” (o de “cruces”), y dividir este número por el número de monedas lanzadas.
- Lanzar una única moneda al aire muchas veces, contar el número de “caras” (o de “cruces”), y dividir este número por el número de veces que he lanzado la moneda al aire.

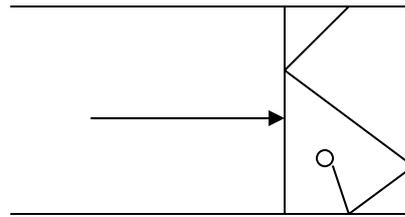
En ambos casos el resultado tendría que ser el mismo. Y es que la naturaleza de las magnitudes termodinámicas nos permite trabajar de igual modo que si estuviésemos lanzando monedas al aire.

Repetiremos ahora nuestro “experimento” de compresión del gas pero incluiremos una nueva limitación: el volumen del gas después de la compresión será exactamente la mitad del volumen inicial. Es decir, $V_2 = \frac{V_1}{2}$ (La figura ilustra esta nueva situación)

⁴ Adviértase que empleo, conscientemente, el término “inferir” en lugar de los términos “calcular” o “evaluar”.



$$V_1$$



$$V_2 = (1/2)V_1$$

Yendo a las expresiones correspondientes, y considerando las limitaciones que acabamos de imponer, resulta que:

$$\Delta F = -k T \text{Ln}\left(\frac{V_1}{2 V_1}\right) = k T \text{Ln}(2)$$

$$\Delta S = k \text{Ln}\left(\frac{V_1}{2 V_1}\right) = -k \text{Ln}(2)$$

Al igual que antes, el estado físico de nuestro gas –constituido ahora por una única partícula- no ha cambiado en absoluto. Sin embargo, vuelve a haber cambios tanto en la energía libre como en la entropía. La pregunta es ahora ¿qué podemos concluir de ello?

Sencillamente, nuestro conocimiento sobre la partícula ha aumentado: hay menos posiciones que la partícula puede ocupar después de la compresión que antes de la compresión. Dicho de otro modo, V_1 es mucho mayor que V_2 y –por lo tanto- en V_2 el espacio de búsqueda es mucho menor. Tal y como ha sido reformulado nuestro problema, la entropía está relacionada ahora con la probabilidad de que el gas se encuentre en la configuración en que se encuentra⁵. Así, si ω es la probabilidad de una configuración determinada:

$$S \approx k \text{Ln}(\omega)$$

Como vemos, cuanto menos sabemos de la configuración de un gas, en más estados puede estar, ω es mayor, y S también es mayor. Recordando la relación a la que

⁵ Si hablásemos de fotografía, por “configuración” podríamos entender algo parecido a una “instantánea” del sistema en la que apareciesen todos los detalles.

habíamos llegado entre información y entropía, y extrapolando un poco estos resultados, volvemos a llegar a la conclusión de que cuanto menos información tenemos sobre un “estado” mayor es su entropía.

Pensemos ahora en una cuestión que pudiera parecer trivial, pero que a lo mejor no lo es tanto: ¿para qué computamos?... Hay muchas respuestas posibles a esta extraña pregunta, pero sin duda estarán de acuerdo conmigo en que al computar algo relacionado con un sistema cualquiera, nuestra información sobre dicho sistema aumenta, y por lo tanto –como hemos visto ya- la entropía asociada debería disminuir.

Este punto de vista nos permitirá llegar a conclusiones muy interesantes sobre el aprendizaje de las máquinas (más concretamente sobre el llamado aprendizaje no supervisado), que trataremos más adelante. Sin embargo, aquí, nos va a servir para empezar a tratar ya la computación –en sí misma-, sobre la que hasta ahora hemos pasado casi de puntillas.

La aproximación de Feynman

Recordando nuestro sencillo modelo de mensaje enviado, definiremos la información del mensaje como una magnitud proporcional a la cantidad de energía libre necesaria para reinicializar el sistema. Esto equivale, en cierto modo, a comprimir cada caja hasta asegurarnos de que la partícula que la ocupa esté en la posición 0. Estamos tratando con un mensaje típico. En este contexto, y en el caso más general, podemos asumir que sobre algunos bits tenemos conocimiento previo, mientras que sobre otros no. En cualquier circunstancia, en el proceso de reinicialización se pueden dar los siguientes casos:

- Si el bit es 0 no hay que hacer nada para reiniciar: $\Delta F = 0$
- Si el bit es 1 hay que hacer trabajo para desplazarlo hacia la posición 0: $\Delta F \neq 0$

La cuestión parece de convenio: ¿No podríamos utilizar una definición alternativa para el proceso de reinicialización? Dicho de otro modo: ¿No sería lo mismo poner todos los bits a 1? ... Esta alternativa introduce una simetría en el sistema, y recordemos que las

simetrías son piezas muy cotizadas en el mundillo de la física. Pero no... en este caso la pieza se nos ha escapado. Y es que el saldo energético sólo sería el mismo si el mensaje contuviera el mismo número de ceros que de unos. Por otra parte, recordemos que reinicializar el sistema equivale a eliminar del mensaje toda la información. Así, con un pequeño esfuerzo de imaginación podemos concluir que sólo se modifica la energía libre del sistema cuando tratamos de reinicializar un bit del cual no sabemos en qué estado se encuentra. Básicamente, esto quiere decir que la información de un mensaje está contenida en los bits que desconocemos... ¡Y para concluir esto tanta historia!

Trataremos este nuevo planteamiento desde una perspectiva algo más formal. Si Ψ es la cantidad de información de un mensaje, podemos establecer que:

- (a) $\Psi \propto \Delta F(x, y, \dots, t \rightarrow 0, 0, \dots, 0)$
- (b) $\Psi \propto \Delta F(x, y, \dots, t \rightarrow 1, 1, \dots, 1)$

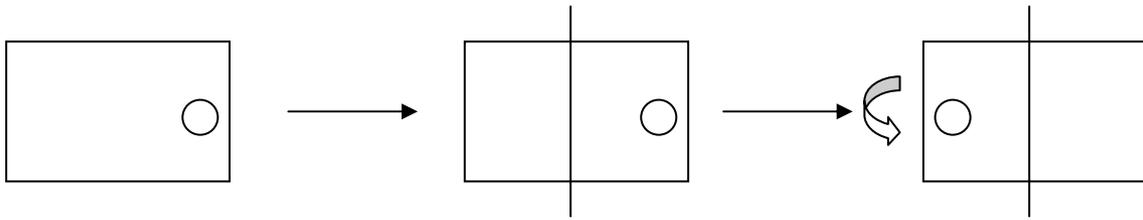
Expresiones que debemos interpretar del siguiente modo: “La cantidad de información de un mensaje es proporcional a la energía libre necesaria para reinicializar el sistema [a “0” –caso (a)-, o a “1” –caso (b)-]”

En general, considerando nuestra discusión previa, y salvo que haya el mismo número de ceros que de unos:

$$\Delta F(x, y, \dots, t \rightarrow 0, 0, \dots, 0) \neq \Delta F(x, y, \dots, t \rightarrow 1, 1, \dots, 1)$$

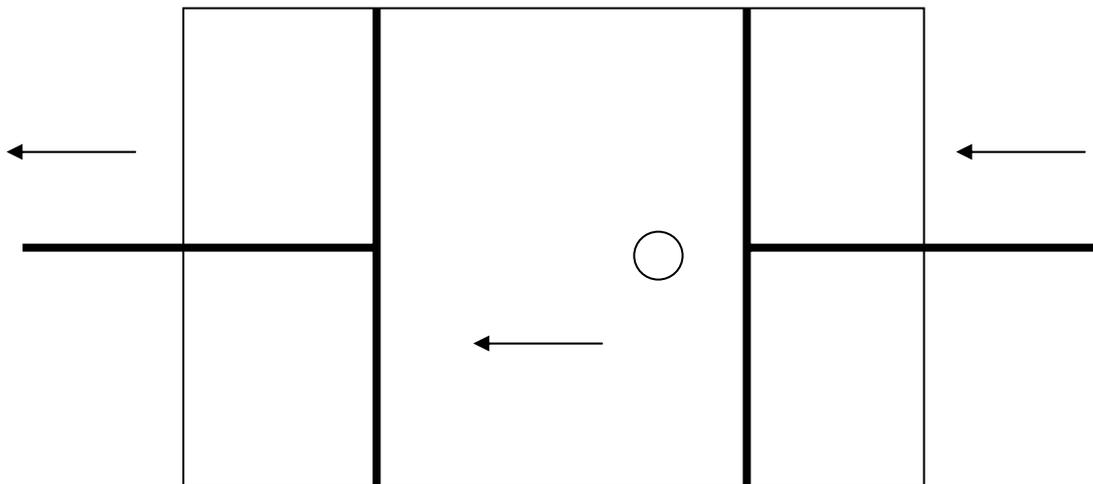
Pero la cantidad de información de un mensaje, Ψ , es siempre la misma independientemente de cómo decidamos reinicializar el sistema. Esta inconsistencia sólo puede resolverse concluyendo que, precisamente, la información del mensaje se encuentra en lo que sobre él desconocemos.

Una consecuencia de los argumentos anteriores es que la energía puesta en juego para reinicializar una celdilla de 1 a 0, es la misma que la que se pone en juego al dejar un 0 como está –es decir, ninguna-. La figura puede contribuir a aclarar un poco la situación.



En este caso consideramos que las cajas son ideales, por lo tanto no consideramos aspectos mecánicos. Además, la energía se refiere sólo a la del contenido del mensaje, y esta energía está relacionada con lo que sabemos sobre el mensaje. De esta forma, el contenido del mensaje depende de las posiciones de las partículas.

Otra manera de verlo se ilustra en la siguiente figura.

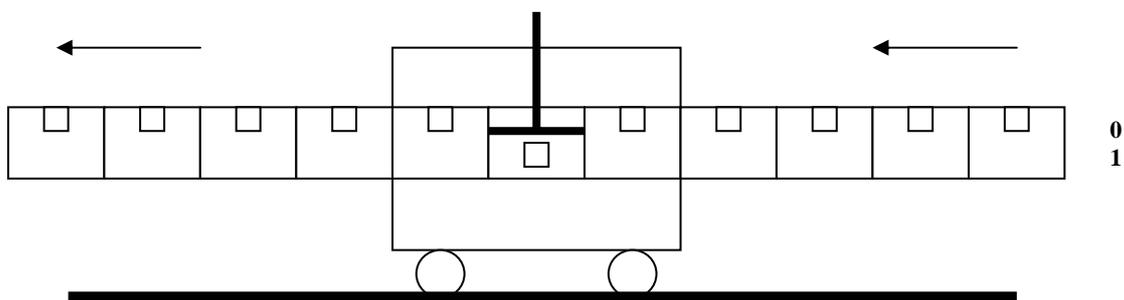


En este caso, como el trabajo sobre un lado se recupera por el otro, el saldo es cero, si el proceso se efectúa suficientemente despacio. El único factor que interviene en F es S , que sólo varía si no sabemos en qué posición está la partícula.

La aproximación de Bennett

También interesado por la energía de la computación en sí misma, Bennett se plantea la posibilidad de utilizar la cantidad de información de un mensaje como combustible. Para ello hay que relacionar dicha información con su poder calorífico; lo que equivale a decir, con la cantidad de energía que podemos extraer de dicho mensaje.

Para ello Bennett propone la siguiente abstracción: supongamos que tenemos una máquina inmersa en un baño térmico. Esta máquina se alimenta de un mensaje: una cinta dividida en celdillas en cada una de las cuales hay una partícula. Supondremos que, inicialmente, todas las partículas de la cinta están en la misma posición (digamos a 0). El mensaje -i.e., la cinta- se introduce en la máquina por un extremo y sale por el otro. Este dispositivo es capaz, al menos teóricamente, de suministrar energía –trabajo útil- que nos permita mover nuestra máquina. Para ello lo único que necesitamos es un pistón de forma que, cada vez que una nueva celda de la cinta que contiene nuestro mensaje está dentro de la máquina, el pistón se introduce en la celda, y reduce el espacio de la celdilla hasta la mitad. En estas condiciones el baño isotérmico se encarga del resto. Recordemos que el sistema inicial está en equilibrio, pero ahora la disminución de volumen de la celda hace que la partícula choque contra el pistón (tiene menos espacio para moverse en las mismas condiciones de temperatura), desplazándolo de nuevo hacia fuera, tratando de restablecerse las condiciones iniciales. El resultado del proceso es un trabajo sobre el émbolo que podría utilizarse para mover la máquina.



Baño Térmico: $T = cte$

Este planteamiento es el mismo que el de Feynman, pero al revés. Aquí el mensaje es quien nos suministra energía, en forma de trabajo sobre el pistón. Suponiendo una cinta con n bits, la energía que podemos obtener es:

$$F = n k T \ln(2)$$

Este es el experimento tal y como lo planteó Bennett, pero pensemos un poco en las consecuencias de dicho experimento... Después de haber pasado el mensaje a través de la máquina, y de haber extraído toda la energía posible ¿cuál es el estado final de la cinta? ¿en qué estado vamos a encontrar a nuestro mensaje inicial? La respuesta, aunque evidente, no deja de ser sorprendente: Ya no hay mensaje (o no sabemos cuál es, que viene a ser lo mismo). Las posiciones de las partículas en la cinta a la salida de la máquina son aleatorias.

Después de haber efectuado su trabajo sobre el pistón, la partícula puede estar en cualquier sitio, y para conocer “exactamente” su posición habría que realizar algún tipo de medida⁶.

Consideremos ahora que, en lugar de estar todos los bits a 0, existe la posibilidad de que algunos de tales bits estén a 1. Para poder obtener ahora la misma cantidad de energía de la cinta que en el caso anterior necesitamos dos cosas: (1) disponer de un pistón maniobrable –o de dos pistones que podamos accionar independientemente, según el bit que entra en la máquina-, y (2) poder conocer cuál es el estado del bit (0 ó 1) que está entrando en la máquina en cada momento, para accionar así el émbolo adecuado. En cualquier otro caso perdemos energía.

Dicho de otro modo, si no sabemos qué bit entra en la máquina, algunas veces el émbolo será empujado por una partícula que está donde tiene que estar, y el sistema generará energía aprovechable. En otras ocasiones, sin embargo, tendremos que trabajar nosotros contra la partícula, consumiendo exactamente la misma cantidad de energía que obtenemos con una configuración favorable. En último extremo, una cinta aleatoria tiene poder calorífico nulo.

⁶ Trataremos sobre la cuestión de la medida más adelante, cuando hablemos sobre los límites teóricos del aprendizaje espontáneo.

La máquina de Bennett produce el efecto justamente contrario al proceso de reiniciación de Feynman. Bennett utiliza un mensaje para producir energía a costa del propio mensaje, y Feynman parte de una cinta aleatoria –i.e., sin mensaje conocido- sobre la cual efectúa un trabajo para obtener una información concreta –i.e., al final todos los bits de la cinta están en la posición 0-

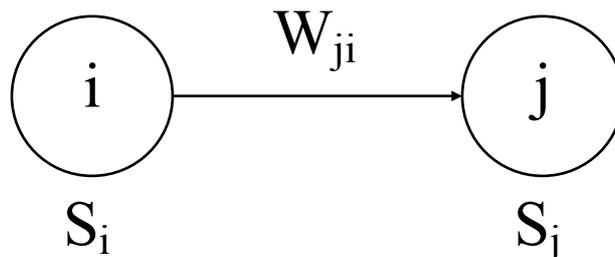
La simetría entre los planteamientos de Feynman y Bennett es evidente. Así, para una cinta con N bits e información Ψ , el poder calorífico de la máquina de Bennett viene dado por la expresión:

$$Poder_calorífico = (N - \Psi)kT \ln(2)$$

Consecuencia de lo anterior es que si extraemos de la máquina de Bennet toda la energía posible -i.e., $\frac{kT \ln(2)}{bit}$ -, al final nos quedamos con información 0.

3. FÍSICA DEL APRENDIZAJE

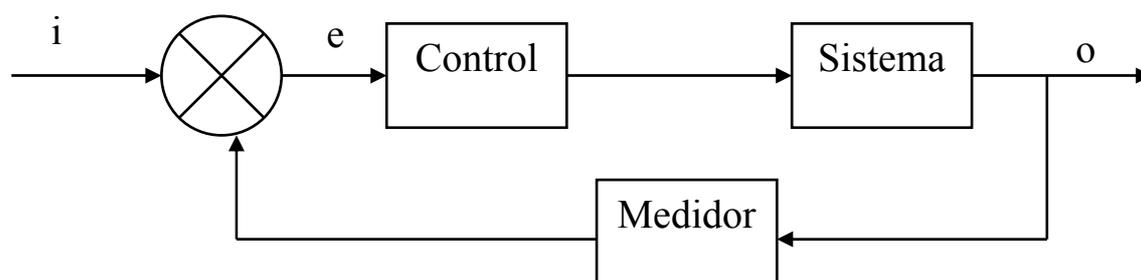
Vamos a abrir ahora un pequeño (o gran) paréntesis en nuestras disquisiciones energéticas sobre la información, para pasar a debatir sobre algunos aspectos relacionados con el aprendizaje. Mantendremos, no obstante, planteamientos basados en conceptos derivados de la física, y utilizaremos como herramienta de “experimentación” esas estructuras de interior oscuro denominadas redes neuronales. Para empezar consideraremos una red neuronal muy sencilla formada por dos únicas neuronas (-i- y -j-). Cada una de tales neuronas se encuentra en un estado S_i y S_j , y ambas están conectadas a través de una sinapsis de peso W_{ji} , tal y como se muestra en la figura.



Uno de los procedimientos de que disponemos para que una red neuronal aprenda es el entrenamiento de la misma. Podemos definir al entrenamiento neuronal de una red como el proceso que modifica los pesos sinápticos de forma que, al final, la aplicación de una serie de entradas produce un conjunto de salidas deseado. Esta definición de entrenamiento neuronal no difiere mucho de lo que en realidad ocurre en los sistemas de control automático clásicos. En estos últimos el sistema que queremos controlar (i.e., sobre el cual queremos obtener un determinado comportamiento), está conectado con un medidor que registra la salida. Esta salida se realimenta y se compara con una referencia. El resultado de esta comparación es una señal de error que es la que alimenta al controlador. Con la señal de error, el controlador elabora una acción de control que es la que se introduce en el sistema para regular su comportamiento.

Los problemas típicos de los sistemas de control son: (1) caracterizar adecuadamente el comportamiento físico del sistema que queremos controlar, (2) diseñar un algoritmo de control adecuado, y (3) lograr la estabilidad del sistema controlado cuando ocurren cambios en la entrada.

Por supuesto, esto no es más que una visión muy simplificada del proceso genérico de control, pero nos sirve para comprender la analogía entre un sistema neuronal y un sistema de control clásico. La figura ilustra lo que acabamos de mencionar sobre los sistemas de control



En cierto modo, el diseño del algoritmo de control adecuado a que antes hacíamos alusión equivale al entrenamiento de una red neuronal. Quizás la diferencia más importante entre ambos sea estructural. Así, mientras que un algoritmo de control está perfectamente caracterizado, en una red neuronal no se sabe muy bien qué es lo que

pasa en su interior: sólo se sabe que funciona bien cuando ante un conjunto de entradas conocido produce las salidas esperadas.

Algunas nociones sobre el aprendizaje neuronal artificial

Nos apartamos ya de las analogías con los sistemas de control para volver sobre la cuestión del entrenamiento neuronal. Al respecto encontramos dos procedimientos genéricos de entrenamiento: el determinístico y el estadístico.

El método determinístico de entrenamiento de una red neuronal no es más que un procedimiento de ajuste de los pesos sinápticos de la red, que se efectúa paso a paso, y que considera:

- Los correspondientes valores de los pesos en un momento dado
- Los valores de las entradas
- Los correspondientes valores de las salidas
- Los valores deseados en las salidas

Frente a los métodos determinísticos, los métodos estadísticos realizan cambios aleatorios⁷ en los pesos sinápticos de la red, y conservan aquellos cambios que suponen una mayor estabilidad del sistema (i.e., que producen menores desviaciones con respecto a la salida deseada). El proceso, globalmente considerado, supone:

- Aplicar un conjunto de entradas y encontrar las salidas correspondientes
- Calcular el error obtenido
- Definir este error como la función objetivo que hay que minimizar
- Seleccionar un peso cualquiera y efectuar un cambio
- Si el sistema se vuelve más estable entonces aceptar el cambio
- Si el sistema no se vuelve más estable entonces descartar el cambio
- Repetir el proceso hasta minimizar la función objetivo

⁷ Tómesese el término “aleatorio” con las debidas reservas.

Un procedimiento estadístico clásico para el entrenamiento de una red neuronal es el método de Boltzmann, que básicamente consiste en lo siguiente:

- Definir la temperatura artificial del sistema que queremos entrenar, T
- Iniciar el proceso de entrenamiento con un valor alto de T
- Aplicar un conjunto de entradas a la red
- Calcular las salidas correspondientes y la función objetivo
- Efectuar un cambio arbitrario en un peso sináptico
- Recalcular la salida de la red neuronal
- Evaluar el correspondiente cambio de la función objetivo
- Si la función objetivo mejora, conservar el cambio efectuado sobre el peso sináptico
- Si la función objetivo empeora, calcular la probabilidad de aceptar el cambio de acuerdo con la ley de distribución de Boltzmann: $P(c) = \exp(-c / kT) = e^{-c/kT}$

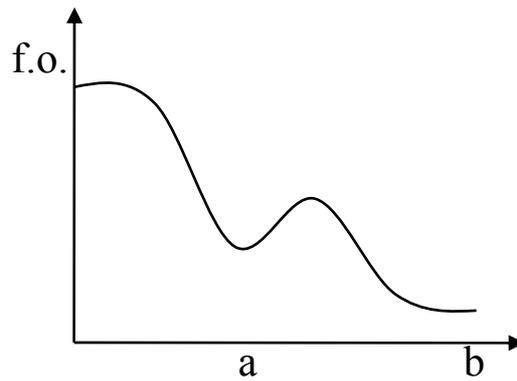
$P(c)$ define la probabilidad de un cambio (c) en la función objetivo, k es la constante de Boltzmann, y T es la temperatura artificial de la red.

Aunque la cuestión pueda parecer algo confusa, sin embargo es así como aparece en muchos libros. Confío en que pronto podamos ir aclarando cosas.

Desde una perspectiva práctica el método de Boltzmann requiere seleccionar un número al azar (r) de una distribución uniforme entre 0 y 1. Así las cosas:

- Si $P(c) > r$ → conservar el cambio
- Si $P(c) \leq r$ → descartar el cambio

El objetivo de proceder de este modo es permitir que el sistema escape de eventuales mínimos locales, tal y como se muestra en la figura.



Vamos ahora a tratar de formalizar conceptos dejando volar un poco la imaginación. A lo mejor, parte de lo que sigue a continuación no es estrictamente cierto, incluso puede que no tenga ningún sentido, pero ya he mencionado que se trata de razonar por analogía, de divertirse, y de que quede lo más coherente posible. Sigamos, pues.

Consideraciones energéticas sobre el aprendizaje de las redes neuronales

Comenzaremos diciendo que la ley de distribución de Boltzmann es una de las bases del aprendizaje estocástico o desordenado⁸. Por otra parte, una neurona puede ser considerada como una estructura recurrente que opera de manera binaria⁹: (+1 , -1). Ahora bien, en el contexto de un sistema neuronal, una máquina de Boltzmann es una entidad caracterizada por una función de energía E, cuyo valor depende de los estados individuales de cada una de las neuronas de la red:

$$E = -\frac{1}{2} \sum_{i \neq j} \sum_{j \neq i} W_{ji} S_j S_i$$

Donde:

Si es el estado de la neurona “i”

Sj es el estado de la neurona “j”

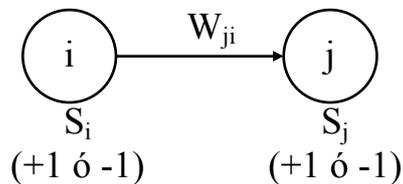
Wji es el peso sináptico de la conexión: i → j

⁸ Esto sí que es cierto.

⁹ El “puede ser considerada...” hace que esto también sea cierto.

Para liar un poco más la cosa, ya que hablamos de analogías ¿no se parece mucho la expresión anterior a la de la energía cinética: $E = \frac{1}{2}mv^2$?

Entraremos ahora en la parte especulativa del asunto, siempre desde una perspectiva formal. Para empezar consideraremos que una sinapsis es excitadora cuando contribuye a disminuir la energía de la red neuronal. Por el contrario, una sinapsis es inhibitoria cuando contribuye a aumentar la energía de la red. Sea ahora una red neuronal muy sencilla, como la de la figura, en la que dos únicas neuronas (i , j) con sus estados correspondientes [S_i (+1 , -1) y S_j (+1 , -1)], están conectadas a través de una sinapsis (W_{ji})



En este contexto podemos asumir que un par de neuronas conectadas están en sintonía si ambas están en el mismo estado. Por lo tanto: $S_i \times S_j > 0$

Análogamente, un par de neuronas conectadas están en disonancia si sus estados individuales son diferentes. En este caso: $S_i \times S_j < 0$

Definamos ahora el nivel de activación elemental de un par de neuronas conectadas como: $\alpha_{ji} = W_{ji}S_jS_i$

De esta forma α_{ji} es una excitación neta cuando:

- El par de neuronas conectadas está en sintonía y la sinapsis es excitadora
- El par de neuronas conectadas está en disonancia y la sinapsis es inhibitoria

Análogamente α_{ji} es una inhibición neta cuando:

- El par de neuronas conectadas está en sintonía y la sinapsis es inhibitoria
- El par de neuronas conectadas está en disonancia y la sinapsis es excitadora

Este planteamiento podría llevarnos a una especie de balance energético de nuestro sistema neuronal. Pero recordemos que estamos tratando de que nuestra red aprenda, y lo estamos intentando desde una perspectiva estocástica utilizando para ello el método de Boltzmann. Con nuestro nuevo planteamiento, el procedimiento anteriormente explicado puede reformularse en los siguientes términos:

- Sea una red neuronal en fase de aprendizaje
- Seleccionar arbitrariamente una neurona j
- Invertir el estado de la neurona j : $S_j \rightarrow -S_j$
- Evaluar la probabilidad de cambio de la red motivada por el cambio de estado de la neurona j de acuerdo con la expresión: $W(S_j \rightarrow -S_j) = \frac{1}{1 + \exp\left(\frac{-\Delta E_j}{T}\right)}$,
 expresión en la cual ΔE_j es la variación energética asociada al cambio de estado, y T es la temperatura de la red neuronal
- Repetir el proceso hasta alcanzar un estado de equilibrio térmico. Dicho de otro modo, repetir el proceso hasta que:
 - o Una nueva modificación (iteración) no modifique la probabilidad de cambio de la red
 - o El sistema no mejore con nuevos cambios
 - o La probabilidad de cambio esté minimizada

Todo esto está muy bien (a mí me gusta). No obstante, pueden surgir algunas cuestiones que pudieran no ser triviales. Por ejemplo:

- ¿Qué es la energía de una red neuronal?
- ¿Qué es la temperatura de una red neuronal?
- En el contexto del aprendizaje ¿qué es una situación de equilibrio térmico?
- ¿Qué quiere decir minimizar la probabilidad de cambio?
- ¿Podemos cuantificar la magnitud de una variación estocástica?
- ¿En qué sentido tiende a evolucionar cualquier sistema físico?

- En definitiva ¿qué es en realidad eso del aprendizaje?
- ¿Existe el aprendizaje espontáneo?
- Si efectivamente existe el aprendizaje espontáneo ¿cuáles son sus límites?

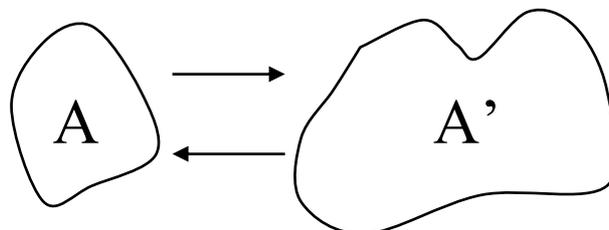
Buscando respuestas desde la física

Algunas de las cuestiones anteriores han sido ya tratadas (o al menos esbozadas, con mejor o peor fortuna, a lo largo de lo que llevamos escrito). Otras no. Para tratar de dar respuesta a los interrogantes planteados recurriremos una vez más a la física. Más concretamente a la teoría cinética de los gases y a la termodinámica.

Asumamos como ciertas las siguientes declaraciones:

- La teoría cinética de los gases establece que la temperatura es una medida de la energía cinética de las partículas de un gas
- Dos sistemas están en equilibrio térmico cuando tienen la misma temperatura
- Como cada partícula de un gas se mueve con una velocidad propia, cada una de tales partículas está en un nivel energético diferente
- A nivel macroscópico, la energía de un sistema es el valor medio de las energías de las partículas de que consta
- A tenor de lo anterior la temperatura es un valor medio de energía

Para acercarnos un poco más a la termodinámica del aprendizaje, consideremos el sistema de la figura, formado por dos subsistemas A y A' que están interaccionando térmicamente.



Claramente, el subsistema A tiene menos partículas que el subsistema A' y, por lo tanto menos grados de libertad. Una pregunta que podemos formularnos ahora es la siguiente:

¿Cuál es la probabilidad P_α de que el subsistema A esté en un estado energético particular E_α ? No nos queda más remedio que seguir con Boltzmann, según el cual la función de distribución de probabilidad es:

$$P_\alpha = \frac{1}{z} \exp\left(-\frac{E_\alpha}{kT}\right)$$

En esta expresión: k es la constante de Boltzmann, T es la temperatura absoluta del sistema, y z es una constante, en principio, independiente de α . Dado que:

$$\sum_\alpha P_\alpha = 1 \rightarrow \sum_\alpha P_\alpha = \sum_\alpha \frac{1}{z} \exp\left(-\frac{E_\alpha}{kT}\right) = 1 \rightarrow 1 = \frac{1}{z} \sum_\alpha \exp\left(-\frac{E_\alpha}{kT}\right)$$

Por lo tanto:

$$z = \sum_\alpha \exp\left(-\frac{E_\alpha}{kT}\right)$$

Llevando este resultado a P_α obtenemos que:

$$P_\alpha = \frac{\exp\left(-\frac{E_\alpha}{kT}\right)}{\sum_\alpha \exp\left(-\frac{E_\alpha}{kT}\right)}$$

Y para otro estado energético cualquiera β :

$$P_\beta = \frac{\exp\left(-\frac{E_\beta}{kT}\right)}{\sum_\beta \exp\left(-\frac{E_\beta}{kT}\right)}$$

Ahora bien, si nuestro sistema está en equilibrio térmico entonces T es constante y, como las sumas anteriores se extienden a todos los niveles posibles de energía, resulta que:

$$\sum_{\alpha} \exp\left(-\frac{E_{\alpha}}{kT}\right) = \sum_{\beta} \exp\left(-\frac{E_{\beta}}{kT}\right)$$

Por lo tanto, dividiendo P_{α} entre P_{β} -Lo que no es más que obtener una relación entre probabilidades de niveles energéticos individuales- se obtiene la expresión siguiente:

$$\frac{P_{\alpha}}{P_{\beta}} = \exp\left\{-\frac{(E_{\alpha} - E_{\beta})}{kT}\right\} = \exp\left\{-\frac{\Delta E}{kT}\right\}$$

Pero ¡un momento! ...¿no habíamos dicho que el método de Boltzmann para el aprendizaje estocástico de una red neuronal se basaba en la ecuación –que por entonces no entendíamos demasiado bien- : $P(c) = \exp(-c/kT)$?

Razonando, por analogía –como hasta ahora-, resulta que la misteriosa “c” no es más que la variación energética del sistema neuronal que está aprendiendo (ΔE), la T neuronal es un parámetro ajustable, y la k neuronal es una constante parecida a la de Boltzmann salvo, tal vez, por sus unidades.

Pero aún podemos estrujar un poco más estas ideas. Recordemos que en el proceso de aprendizaje de una red neuronal, la probabilidad de cambio de la red motivada por el cambio de estado de la neurona j se podía evaluar de acuerdo con la expresión:

$$W(S_j \rightarrow -S_j) = \frac{1}{1 + \exp\left(\frac{-\Delta E_j}{T}\right)}$$

Al respecto, si prescindimos de la k (al fin y al cabo no es más que una constante totalmente prescindible), y recuperamos la expresión obtenida para $\frac{P_{\alpha}}{P_{\beta}}$, obtenemos:

$$W(S_j \rightarrow S_i) = \frac{1}{1 + \left(\frac{P_\alpha}{P_\beta}\right)}$$

Resultado (por cierto, de dudosa utilidad) que demuestra que la probabilidad de un cambio durante un proceso de aprendizaje neuronal es precisamente eso, un concepto probabilístico... Fascinante.

Seguimos buscando respuestas

Ya hemos dado respuesta a algunas de las preguntas que nos formulábamos en relación con el aprendizaje neuronal. Para tratar de contestar al resto de las cuestiones planteadas vamos a necesitar de nuevo el concepto de energía libre de un sistema termodinámico y, aprovechando que ya sabemos un poco más de física de los gases, redefiniremos el concepto de acuerdo con la siguiente expresión:

$$F = -kT \ln(z) = -kT \ln \sum_{\alpha} \exp\left(-\frac{E_{\alpha}}{kT}\right)$$

en la que todos los términos son ya conocidos. Consideremos ahora nuestro sistema anterior, el formado por dos subsistema A y A' que están intercambiando energía en forma de calor. En este contexto sabemos que, por la segunda ley de la termodinámica, la entropía del sistema (S) puede definirse, en función de la energía media, del siguiente modo:

$$S = k \ln(z) + \frac{\bar{E}}{T} = k \ln \sum_{\alpha} \exp\left(\frac{E_{\alpha}}{kT}\right) + \frac{\bar{E}}{T}$$

$$\bar{E} = \sum_{\alpha} P_{\alpha} E_{\alpha} = \text{Energía_media}$$

$$\text{Claramente: } TS = kT \ln(z) + \bar{E}$$

Pero recordemos la expresión de la energía libre de un sistema: $F = U - TS$

Si $U = \overline{E}$ es, como ya hemos señalado, la energía media del sistema calculada sobre todos los estados energéticos posibles, entonces:

$$TS = kT \ln(z) + \overline{E} = U - F \rightarrow F = -kT \ln(z)$$

Lo verdaderamente importante de todo esto es el concepto de energía libre, que relaciona macroscópicamente la energía media de un sistema con la entropía, lo que supone el matrimonio perfecto entre el primero y el segundo principio de la termodinámica. Así:

- Un sistema tiende espontáneamente a minimizar su energía
- Un sistema tiende espontáneamente a maximizar su entropía

Por lo tanto, cualquier sistema debería evolucionar espontáneamente hacia el mínimo de energía libre... Esto no parece ir a favor del aprendizaje espontáneo –todos los que hemos tenido que aprender algo alguna vez sabemos el esfuerzo (energía) que cuesta estudiar, y lo rápidamente que se desordena nuestra mesa de trabajo (entropía)-. Lo ideal sería no gastar energía alguna estudiando, y dejar que la mesa esté todo lo desordenada que le dé la gana, pero entonces no aprenderíamos nada¹⁰. De todas formas el aprendizaje espontáneo no está del todo prohibido. Volveremos sobre esta cuestión un poco más adelante.

Otro de los temas que habíamos dejado planteado, pero no resuelto, tenía que ver con la naturaleza de la entropía: ¿Qué tipo de variable es S?

Para abordar este problema tendremos que recordar las siguientes tres expresiones, en las que todos sus términos deberían ser ya familiares:

$$P_{\alpha} = \frac{1}{z} \exp\left(-\frac{E_{\alpha}}{kT}\right)$$

¹⁰ A veces me pregunto... ¿y qué?

$$\bar{E} = \sum_{\alpha} P_{\alpha} E_{\alpha}$$

$$S = k \ln(z) + \frac{\bar{E}}{T}$$

Trabajemos un poco con estas expresiones. De acuerdo con la definición de P_{α} :

$$\ln(P_{\alpha}) = \ln\left\{\frac{1}{z} \exp\left(-\frac{E_{\alpha}}{kT}\right)\right\} = \ln\left(\frac{1}{z}\right) - \frac{E_{\alpha}}{kT} \rightarrow \frac{E_{\alpha}}{kT} = \ln\left(\frac{1}{z}\right) - \ln(P_{\alpha})$$

Por otra parte:

$$S = k \ln(z) + \frac{\bar{E}}{T} = k \ln(z) + \frac{1}{T} \sum_{\alpha} P_{\alpha} E_{\alpha}$$

$$\frac{S}{k} = \ln(z) + \sum_{\alpha} P_{\alpha} \frac{E_{\alpha}}{kT} = \ln(z) + \sum_{\alpha} P_{\alpha} \left\{ \ln\left(\frac{1}{z}\right) - \ln(P_{\alpha}) \right\}$$

$$\frac{S}{k} = \ln(z) + \ln\left(\frac{1}{z}\right) \sum_{\alpha} P_{\alpha} - \sum_{\alpha} P_{\alpha} \ln(P_{\alpha}) = \ln\left(\frac{z}{z}\right) \sum_{\alpha} P_{\alpha} - \sum_{\alpha} P_{\alpha} \ln(P_{\alpha}) = -\sum_{\alpha} P_{\alpha} \ln(P_{\alpha})$$

Y, evidentemente: $S = -k \sum_{\alpha} P_{\alpha} \ln(P_{\alpha}) \Leftrightarrow z > 0 : z \neq 0$

Puesto que $z = \sum_{\alpha} \exp\left(-\frac{E_{\alpha}}{kT}\right)$, z no puede ser cero porque:

$$\forall \alpha, E_{\alpha} \geq 0 : k = cte \geq 0 : T \geq 0$$

$$\text{Si } z = 0 \rightarrow \forall \alpha : \exp\left(-\frac{E_{\alpha}}{kT}\right) = 0 \rightarrow \exp\left(\frac{E_{\alpha}}{kT}\right) = \infty \rightarrow E_{\alpha} = \infty$$

La expresión obtenida para S , nos muestra que la entropía es una magnitud de carácter probabilístico. ¡Pues vaya una cosa! Eso es una trivialidad –no me lo creo-. Ya lo

sabíamos –eso sí que me lo creo-. La diferencia está en que ahora lo hemos demostrado, lo cual refuerza muchos de los argumentos esgrimidos cuando discutíamos cuestiones anteriores relacionadas con los bits, la energía de la computación o el aprendizaje. Una ventaja añadida es que este planteamiento suele fomentar la aparición de acaloradas discusiones, precisamente este es uno de los principales objetivos de este –llamémosle- trabajo.

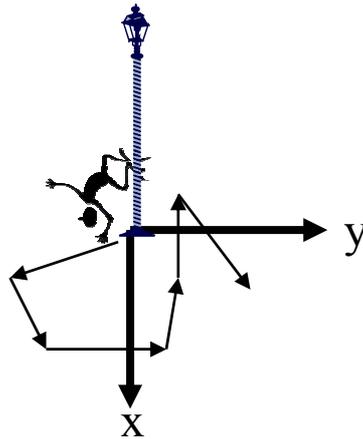
Pero todavía no hemos agotado el tema del aprendizaje desde la óptica, un poco lejana ya, de la física. Decíamos un poco más arriba que la ley de distribución de Boltzmann es una de las bases del aprendizaje estocástico o desordenado. Con el método de Boltzmann introducíamos cambios aleatorios en nuestro sistema neuronal, con el sano y loable propósito de que la red aprendiese algo. Se me ocurren ahora dos preguntas sobre las que debatir:

- ¿Podemos predecir el cambio producido por un proceso estocástico?
- ¿Cómo cuantificamos la magnitud de una variación estocástica?

No nos queda más remedio que sumergirnos en las leyes del caos, o leyes del comportamiento estadístico, pero que no cunda el pánico: el agua apenas nos va a mojar los tobillos. Plantearemos la cuestión desde la perspectiva de Gamow¹¹.

Situémonos en el siguiente escenario: Estudiante, madrugada del jueves, noche de farra, algunas (varias) copas de más, farola providencial en la que nuestro amigo está apoyado (casi recostado) tratando de no perder demasiado el equilibrio. En un momento de lucidez nuestro amigo piensa que ya tiene bastante, y decide marcharse a casa. Con paso más bien vacilante (uno a la derecha, tres a la izquierda, dos adelante, uno hacia atrás,...) inicia su particular peregrinación. ¡Sinceramente esperamos que lo consiga, aunque no parece muy probable que el viernes pueda ir a clase de 9!

¹¹ G.Gamow, “1, 2, 3,..., infinito”



Estamos ante un proceso claramente estocástico sobre el cual vamos a tratar de saber algo. Por ejemplo, después de un número determinado de pasos ¿podemos saber a qué distancia está nuestro amigo de la farola que le servía de soporte?

Si tomamos la base de la farola como el origen de un sistema de dos coordenadas ortogonales, y llamamos R a la distancia del borrachín al farol después de N pasos zigzagueantes, podemos aplicar el teorema de Pitágoras, según el cual:

$$R^2 = (x_1 + x_2 + \dots + x_n)^2 + (y_1 + y_2 + \dots + y_n)^2$$

Así, si la farola tiene de coordenadas $(0, 0)$, en un instante cualquiera “ i ” después del inicio del paseo de nuestro querido amigo, su posición con respecto al origen será (x_i, y_i)

Podemos ahora comenzar a desarrollar expresiones. Según las leyes de la aritmética:

$$A = (x_1 + x_2 + \dots + x_n)^2 = x_1^2 + x_1x_2 + \dots + x_2^2 + x_2x_1 + \dots + x_n^2 + \dots$$

$$B = (y_1 + y_2 + \dots + y_n)^2 = y_1^2 + y_1y_2 + \dots + y_2^2 + y_2y_1 + \dots + y_n^2 + \dots$$

En ambas expresiones nos aparecen cuadrados puros y productos mixtos. Pero, dado que estamos hablando de un proceso estocástico, y de acuerdo ahora con las leyes de la estadística, si el número de pasos de nuestro amigo (N) es grande, los productos mixtos tienen la misma probabilidad de ser positivos que negativos y, por lo tanto tienden a anularse.

Por esta razón:

$$A = (x_1 + x_2 + \dots + x_n)^2 = x_1^2 + x_2^2 + \dots + x_n^2 = N\bar{x}^2$$

$$B = (y_1 + y_2 + \dots + y_n)^2 = y_1^2 + y_2^2 + \dots + y_n^2 = N\bar{y}^2$$

En estas expresiones hemos sustituido los cuadrados puros concretos por los valores medios de las proyecciones de cada paso sobre el eje X y sobre el eje Y. Ello supone que la longitud del paso de nuestro amigo es más o menos constante; aunque para Ns verdaderamente grandes esta última asunción es irrelevante.

Por lo tanto, según lo anterior:

$$R^2 = N(\bar{x}^2 + \bar{y}^2) \rightarrow R = \sqrt{N(\bar{x}^2 + \bar{y}^2)}$$

Pero... ¿podemos ahora estimar cuál es la proyección media de los pasos de nuestro amigo sobre los ejes correspondientes? Pues sí que podemos. Veamos cómo.

Dado que el sistema es ortogonal, y puesto que el movimiento más bien vacilante de nuestro amigo es estocástico, los valores medios sufren una desviación de 45° en cualquier dirección. De este modo:

- Proyección sobre el eje X: $R_x = |\vec{R}| \cos(\alpha) = \frac{\sqrt{2}}{2} |\vec{R}|$
- Proyección sobre el eje Y: $R_y = |\vec{R}| \text{sen}(\alpha) = \frac{\sqrt{2}}{2} |\vec{R}|$

Como el paso normal de una persona es de unos 70 – 80 centímetros:

$$\sqrt{(\bar{x}^2 + \bar{y}^2)} \approx 1 \rightarrow R \approx \sqrt{N}$$

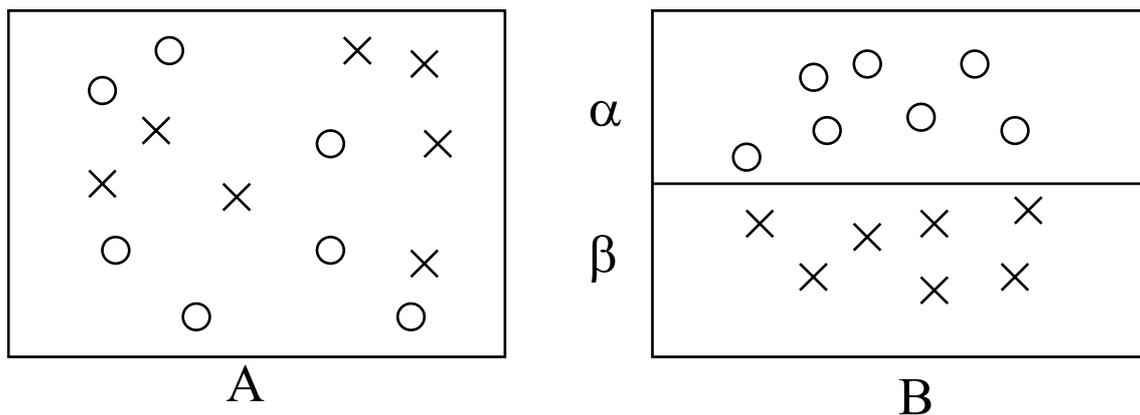
Este resultado se interpreta del modo siguiente: si el proceso es estocástico, la distancia a la que se encuentra nuestro vacilante amigo de la farola, después de un número grande

de pasos, es independiente de la dirección; tan sólo depende del número de pasos dados (más bien de la raíz cuadrada del número de pasos, lo que sugiere que para llegar por el camino más corto a un sitio es mejor estar sobrio)

¿Seríamos capaces ahora de llevar este resultado al modelo de aprendizaje neuronal estocástico que hemos discutido hace ya algún tiempo?... Dejaremos esta interesante cuestión como elemento de debate.

Clasificación, aprendizaje, y límite teórico del aprendizaje espontáneo

Los conceptos de clasificación y de aprendizaje están estrechamente relacionados a través de la entropía, sobre la que tanto hemos especulado. La figura puede contribuir a visualizar lo que queremos decir.



Claramente, clasificar algo nos obliga a minimizar la entropía del sistema. En la figura, la probabilidad de la situación A –en donde nada está clasificado- es mucho mayor que la entropía de la situación B –en la que las partículas (x) están en un sitio, y las partículas (o) están en otro-. Es el momento de recordar nuestros comentarios previos sobre la relación existente entre entropía y probabilidad para comprobar que, efectivamente, clasificar obliga a minimizar la entropía. Pero además –y ésta es una cuestión algo más intuitiva sobre la que también hemos hablado-, cuando somos capaces de clasificar un sistema –en este caso partículas (x) y partículas (o)-, hemos aprendido algo sobre dicho sistema. Así, una vez clasificadas las partículas –este

argumento ha sido ya utilizado-, sabemos perfectamente dónde tenemos que buscar para encontrar una partícula (x) o una partícula (o).

Ahora bien, estamos en un escenario muy concreto. Si nos fijamos, la situación B se caracteriza porque hay un tabique que separa a las partículas (o) y (x). Es más, el tabique es quien realmente nos permite materializar la clasificación de las partículas. Pero... ¿qué pasaría si quitásemos el tabique clasificador? Pues nada que no hayamos visto ya. El sistema, libre de restricciones, y obedeciendo a las leyes de la termodinámica, se desclasificaría solo y evolucionaría espontáneamente hacia la situación A, de máximo desorden.

No obstante habíamos definido el movimiento de las partículas de un sistema de este tipo como algo caótico y desordenado; en definitiva, como un proceso estocástico. Siguiendo con este razonamiento ¿no podría ocurrir que, por casualidad, el sistema se clasificase espontáneamente él solito? Ello supondría que, espontáneamente, todas las partículas (x) se fueran hacia un extremo del recipiente, y las partículas (o) se fueran hacia el otro extremo. La respuesta es que, en principio, no hay nada que impida que esto ocurra. Simplificaremos un poco el escenario para tratar de verlo un poco mejor.

Consideremos una habitación típica, de esas que comparten dos estudiantes afortunados –con dos camas de 80 centímetros, dos armarios de baratillo en los cuales reina la entropía más absoluta, y dos puestecillos de trabajo –elementos estos últimos, habitualmente poco utilizados-. Supongamos que la habitación tiene unas dimensiones regulares, digamos $5m \times 4m \times 2.5m = 50m^3 = 5 \times 10^7 cm^3$

A una temperatura “normal”, esta habitación contiene unos 5×10^4 gramos de aire o, lo que es lo mismo, unas 10^{27} moléculas de aire. En este nuevo escenario la pregunta ahora es: ¿cuál es la probabilidad de que, espontáneamente, todas estas moléculas decidan situarse en una de las esquinas superiores de la habitación provocando –indefectiblemente- la muerte por asfixia de nuestros dos estudiantes?

De acuerdo con lo ya visto, es fácil comprobar que si llamamos α a este particular estado de las moléculas de la habitación, la probabilidad de α es:

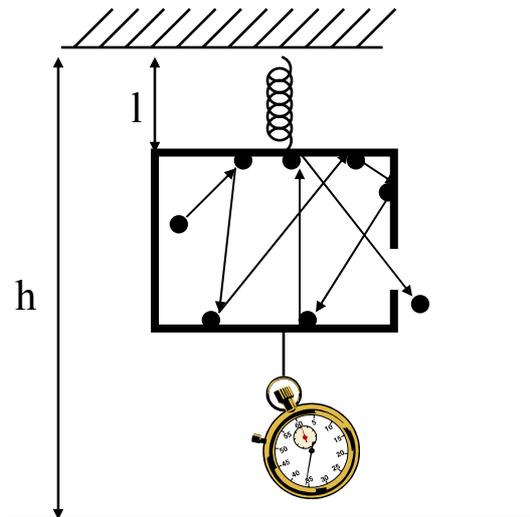
A Einstein el tiro le salió por la culata, ya que Bohr demostró sin ninguna discusión que el dispositivo ideado por Einstein obedecía a las leyes de la mecánica cuántica. En su demostración Bohr utilizó la teoría de la relatividad que el propio Einstein había publicado algunos años antes. A nosotros este experimento nos servirá para especular sobre los límites que la física impone al aprendizaje espontáneo.

Una de las cosas que Einstein no se creía de la mecánica cuántica era el principio de indeterminación de Heisenberg. Este principio establece que, para cualquier partícula en movimiento, es imposible conocer simultáneamente y con absoluta precisión su posición y su cantidad de movimiento. Ello supone que podemos medir exactamente la posición de una partícula. También podemos medir exactamente la velocidad de dicha partícula. Pero ambas, posición y velocidad, al mismo tiempo no pueden ser medidas con total precisión. El simple hecho de medir introduce una distorsión en el sistema, una incertidumbre que puede expresarse mediante la ecuación: $\Delta p \times \Delta s \approx h \approx 10^{-34}$

En esta expresión –en la que hemos prescindido de unidades-, p es la cantidad de movimiento, s es la posición de la partícula, y h es la constante de Planck. La interpretación de esta ecuación es la siguiente: El error que se comete cuando se miden simultáneamente la posición y la cantidad de movimiento de un sistema es del orden de magnitud de la constante de Planck.

La ecuación anterior, sobre cuyos fundamentos no vamos a entrar- no es más que una de las indeterminaciones cuánticas. Hay más, por ejemplo, energía (E) y tiempo (t) también están relacionadas a través de la indeterminación cuántica: $\Delta E \times \Delta t \approx h \approx 10^{-34}$

Sabida es la afición de Einstein por los experimentos teóricos que él mismo diseñaba y que utilizaba como medio para “demostrar” sus hipótesis, o para “rebatir” hipótesis en las que no creía. Enzarzado con Niels Bohr –a quien, por cierto nunca pudo derrotar científicamente- sobre el principio de indeterminación de Heisenberg, y basándose en la segunda de las indeterminaciones cuánticas que hemos reproducido aquí, Einstein diseñó el dispositivo de la figura para convencer a Bohr de que la indeterminación cuántica era, poco más o menos, una majadería.



La idea era la siguiente: supongamos que tenemos una caja en cuyo interior hay una única partícula que está en continuo movimiento, rebotando contra las paredes de la caja una y otra vez. La energía del sistema está determinada por la actividad de la partícula en la caja. También podríamos decir que, si el sistema está sometido a un campo de fuerzas centrales, el peso del sistema es el peso de la caja más el peso de la partícula¹³. La caja del experimento está provista de una pequeña abertura por la cual, eventualmente, puede salir la partícula. Además, la caja está colgada del techo por un muelle cuya elongación –con la partícula dentro- es λ . También hay un reloj, que cuelga de la caja, cuya precisión es total.

Según Einstein, este dispositivo vulnera claramente la indeterminación cuántica. Veamos cómo... Si esperamos lo bastante, llegará un momento en el que la partícula encuentre la abertura de la caja. Este hecho será detectado simultáneamente por el muelle -que se contraerá algo, variando en consecuencia su energía- y por el reloj –cuya precisión es total-. ¡Adiós a la indeterminación cuántica!... ¿Adiós a la indeterminación cuántica?...

Niels Bohr creyó que esta vez su amigo Alberto le había derrotado. Pero Bohr era hombre tenaz y finalmente pudo demostrar con toda claridad que el dispositivo diseñado por Einstein seguía supeditado a las indeterminaciones cuánticas. Porque...

¹³ Obviaremos consideraciones teóricas de mayor calado.

Bohr: ¿Qué pasa cuando la partícula escapa?

Einstein: Ya te lo he dicho, el reloj dice cuándo, y el muelle se acorta.

Bohr: ¿Y qué pasa cuando el muelle se acorta?

Einstein: Pues que la distancia de la caja al techo disminuye.

Bohr: Sí, pero lo hace con una determinada aceleración y a una velocidad determinada con respecto a la situación inicial de reposo ¿no es cierto?

Einstein: Pues sí... claro. ¿Por qué lo dices?

Bohr: No, por nada. Es que hay una teoría por ahí –Relatividad, creo que se llama-, que establece que el tiempo transcurre de forma distinta cuando un sistema está en movimiento que cuando el mismo sistema está en reposo. Vamos, que según esa teoría –Relatividad, creo que se llama-, tu reloj medirá de forma diferente el tiempo cuando: (a) la partícula está en la caja, (b) cuando la partícula está fuera de la caja, y (c) cuando la partícula está saliendo de la caja.

Bohr y Einstein introdujeron las correcciones relativistas al experimento y llegaron a una expresión similar a la ya mencionada $\Delta E \times \Delta t \approx h \approx 10^{-34}$

Pero volviendo a nuestro tema... ¿Qué tiene que ver toda esta historia con la clasificación, el aprendizaje, y los límites teóricos del aprendizaje espontáneo?

Contestaré a la pregunta formulando otra: ¿Podemos diseñar un clasificador más sencillo que el dispositivo propuesto por Einstein para llevar a cabo su experimento en contra de las indeterminaciones cuánticas?

Si la respuesta a mi pregunta es que no, entonces los límites teóricos del aprendizaje espontáneo podrían ser del orden de la constante de Planck.

4. ENERGÍA DE LA COMPUTACIÓN Y COMPUTACIÓN REVERSIBLE

Hasta ahora nos habíamos preocupado de la física de la computación desde una perspectiva semi-cuantitativa. Buscábamos vínculos entre magnitudes termodinámicas y reconocíamos similitudes y analogías entre magnitudes físicas y ciertos aspectos

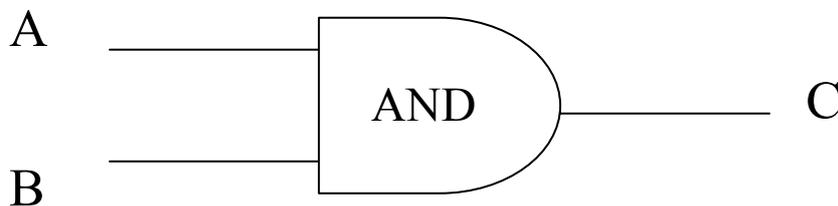
relacionados con la información de un mensaje, con la propia computación, o con el aprendizaje computacional. También propusimos algunos experimentos, más bien abstractos, y pudimos concluir algún resultado interesante –como los obtenidos por Feynman y Bennet con sus máquinas, por ejemplo-

Seguiremos trabajando sobre la energía mínima de una computación, pero con el punto de mira puesto en otro sitio. Vamos a tratar de adelantarnos a los acontecimientos. Y es que, al paso que vamos, no es descabellado suponer que en un determinado lapso de tiempo –imposible de predecir por ahora- la tecnología será capaz de producir ordenadores cuasi-microscópicos en los que el tamaño de sus componentes obligue a tener en cuenta las implicaciones de las leyes fundamentales de la física de lo muy pequeño en la computación.

En este contexto, parece razonable suponer que cualquier proceso computacional requiere un mínimo de energía para poder ser ejecutado. En su versión más sencilla, un proceso computacional constituido por un único paso, la propuesta que podemos considerar inicial es que el mínimo de energía libre necesario para llevar a cabo dicha computación es:

$$F = kT \ln(2)$$

Piénsese, por ejemplo en una puerta lógica AND como la de la figura.



En este caso el “volumen” de la salida es la mitad del de la entrada –hay dos entradas y una única salida-. Con mayor rigor diríamos que el volumen del espacio de fases se ha reducido a la mitad, lo que permite justificar la expresión anterior si recordamos planteamientos ya esbozados anteriormente.

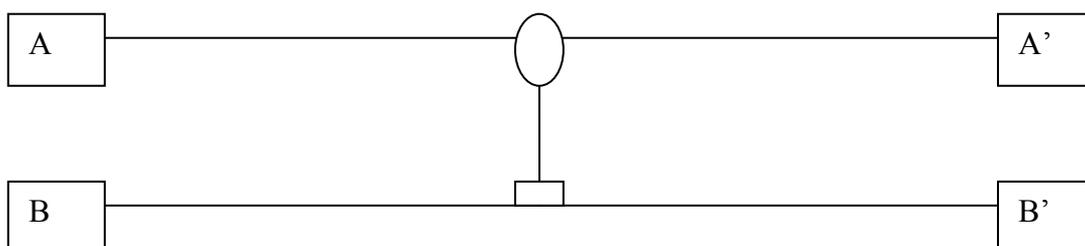
Sin embargo, se puede demostrar que la energía libre puesta en juego en cualquier computación puede ser todo lo pequeña que se quiera. De hecho cualquier computación puede realizarse sin coste de energía libre, pero... (y es un pero muy grande), la condición es que el proceso computacional debe ser infinitamente lento. La situación es muy similar a la que ocurre cuando hablábamos de la entropía y necesitábamos de un baño a temperatura constante para mantener a nuestro sistema en equilibrio térmico.

Sin embargo, lo que acabamos de decir no siempre funciona, de hecho con una puerta AND como la anterior no funcionaría, sólo funcionaría con un tipo especial de ordenadores, a los que denominamos ordenadores reversibles, cuya principal característica es que a partir de la salida siempre podemos conocer la entrada. Al respecto, la puerta lógica AND (y pensemos que los ordenadores realizan fundamentalmente operaciones lógicas) es claramente no reversible. Para comprobarlo basta con examinar su tabla de verdad.

A	B	C = A AND B
0	0	0
0	1	0
1	0	0
1	1	1

Si $C = 0$, tenemos tres configuraciones posibles a la entrada de la puerta, por lo tanto la puerta AND es irreversible.

Sin embargo tenemos otras puertas, aparte de la clásica NOT, que sí son reversibles. Una de tales puertas es la puerta CN que se muestra en la figura.



Esta puerta tiene dos entradas -A y B- y dos salidas A' y B' -. La línea superior es una línea de control, con la particularidad de que la salida A' - es siempre igual a la entrada A -. En la línea inferior, el cuadrado es una operación NOT, pero controlada por la línea superior del siguiente modo:

- Si $A = A' = 0 : B' = B$
- Si $A = A' = 1 : B' = \text{NOT } B$

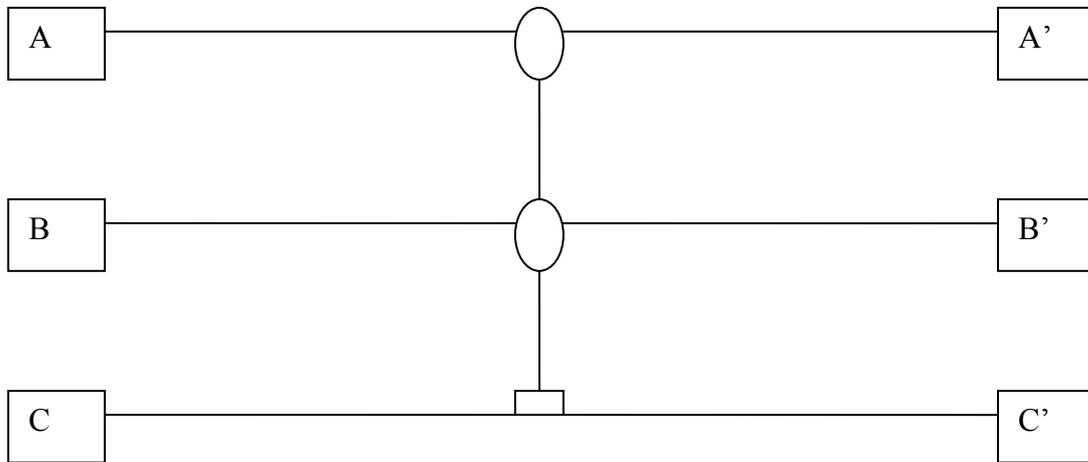
Por lo tanto, la tabla de verdad de la puerta CN es la siguiente:

TABLA DE VERDAD DE LA PUERTA -CN-			
A	B	A'	B'
0	0	0	0
0	1	0	1
1	0	1	1
1	1	1	0

Una cuestión que llama la atención de la puerta CN es que la salida B' es la salida que tendríamos a partir de una puerta XOR alimentada con A y con B. En todo caso, el dispositivo no es el mismo puesto que la puerta CN produce realmente dos salidas. Nótese también que esta puerta es perfectamente reversible ya que, conocida la salida, podemos reproducir siempre la entrada. Otra forma de verlo es ejecutar dos veces seguidas una operación CN, tal y como se muestra en la tabla.

CN – CN					
A	B	A'	B'	A''	B''
0	0	0	0	0	0
0	1	0	1	0	1
1	0	1	1	1	0
1	1	1	0	1	1

Pero esta no es la única puerta reversible –aparte de la NOT clásica- que podemos construir. También tenemos la puerta CCN, es decir, una puerta NOT doblemente controlada. La figura muestra un esquema de dicha puerta.



El funcionamiento de la puerta CCN es muy similar al de la puerta CN. Las dos líneas superiores son las líneas de control, y verifican que $A = A'$ y $B = B'$. El cuadrado de la línea inferior es un NOT, pero controlado por las dos línea superiores, de forma que $C' = \text{NOT } C$ si $A = 1$ y $B = 1$. Consecuentemente, la tabla de verdad de la puerta CCN es la siguiente:

TABLA DE VERDAD DE LA PUERTA –CCN-					
A	B	C	A'	B'	C'
0	0	0	0	0	0
0	0	1	0	0	1
0	1	0	0	1	0
1	0	0	1	0	0
0	1	1	0	1	1
1	0	1	1	0	1
1	1	0	1	1	1
1	1	1	1	1	0

Esta puerta es claramente reversible. Para comprobarlo podemos hacer lo mismo que hicimos con la puerta CN –aplicarla dos veces consecutivas, y verificar si la última salida coincide con la primera de las entradas. Esta situación se ilustra en la siguiente tabla.

CNN – CNN								
A	B	C	A'	B'	C'	A''	B''	C''
0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0	1
0	1	0	0	1	0	0	1	0
1	0	0	1	0	0	1	0	0
0	1	1	0	1	1	0	1	1
1	0	1	1	0	1	1	0	1
1	1	0	1	1	1	1	1	0
1	1	1	1	1	0	1	1	1

Comprobamos que, efectivamente: $\{A, B, C\} = \{A'', B'', C''\}$

Pero esta no es la única ventaja de la puerta CNN ya que, por sí misma, esta puerta constituye un conjunto completo de operadores. Por ejemplo, la puerta AND clásica (que, recordemos, es irreversible), puede obtenerse a partir de la puerta CNN sin más que fijar $C = 0$, y alimentar la puerta con A y B, tal y como demuestra la tabla.

CNN → AND					
A	B	C	A'	B'	C'
0	0	0	0	0	0
0	1	0	0	1	0
1	0	0	1	0	0
1	1	0	1	1	1

Observamos que C' sólo cambia su estado inicial $\{C = 0\}$ cuando $A = 1$ y $B = 1$. C' es la salida que obtendríamos con una puerta AND convencional.

Análogamente la puerta CNN nos permite reproducir los resultados de una puerta NAND convencional. En este caso hay que fijar la entrada $C = 1$, y alimentar la puerta CNN como antes. La tabla ilustra el resultado.

CNN → NAND					
A	B	C	A'	B'	C'
0	0	1	0	0	1
0	1	1	0	1	1
1	0	1	1	0	1
1	1	1	1	1	0

También podemos reproducir los resultados de una puerta XOR convencional. En este caso hay que fijar $A = 1$ o $B = 1$, y operar como siempre (aunque ahora las entradas “variables” son B y C). La tabla ilustra el proceso.

CNN → XOR					
A	B	C	A'	B'	C'
1	0	0	1	0	0
1	1	0	1	1	1
1	0	1	1	0	1
1	1	1	1	1	0

Dejamos como ejercicio la demostración de que también puede obtenerse una puerta OR a partir de una puerta CNN.

Pero hay más cuestiones interesantes sobre las puertas reversibles. Hagamos un pequeño análisis. Hasta ahora hemos identificado tres: la puerta NOT clásica, la puerta CN, y la puerta CNN. La puerta NOT acepta una entrada y devuelve una salida, la

puerta CN acepta dos entradas y devuelve dos salidas, y la puerta CNN acepta tres entradas y devuelve tres salidas. Por otra parte, las tres son reversibles. La conclusión que parece necesaria -aunque no tiene por qué ser suficiente- para construir una puerta reversible, es que el número de entradas coincida con el número de salidas.

Pero... ¿por qué podemos decir que una computación reversible se efectúa con coste cero de energía? Feynman argumenta en este sentido ayudándose de una computación especial: La computación copia. Siempre según Feynman, es bastante poco intuitivo que se pueda copiar algo sin gastar nada de energía. Para tratar de demostrarlo, consideraremos dos mensajes idénticos, el original y la copia. En este contexto se pueden dar dos situaciones diferentes:

- Conocemos el mensaje original
- No conocemos el mensaje original

Si conocemos el mensaje original, recordando argumentos anteriormente discutidos, no gastamos energía libre al borrar ni el original ni la copia. Simplemente identificando los bits que no están a 0, basta con “darles la vuelta” –como hacíamos antes-, y ya tenemos ambas cintas reiniciadas con coste 0 de energía libre. La situación es algo diferente cuando no conocemos el mensaje original. En este caso borrar el mensaje original sí que cuesta energía libre, precisamente la asociada a los bits desconocidos. Ahora bien, una vez identificados estos bits, sobre la copia el proceso de borrado vuelve a ser “darles la vuelta”, lo que –como ya sabemos- no cuesta energía libre. La idea que subyace detrás de todo esto es que no hay más información en el conjunto {original y copia}, que en el conjunto {original}.

5. ENERGÍA Y VELOCIDAD DE UNA COMPUTACIÓN

Hasta ahora hemos estado especulando bastante sobre la energía intrínseca involucrada en una computación. Hemos llegado a alguna conclusión interesante; por ejemplo, el coste computacional 0 requiere que la computación se haga infinitamente despacio (lo cual, por cierto, no es muy práctico). Introduciremos ahora una restricción: queremos que nuestra computación se realice en un tiempo finito. La cuestión planteada requiere

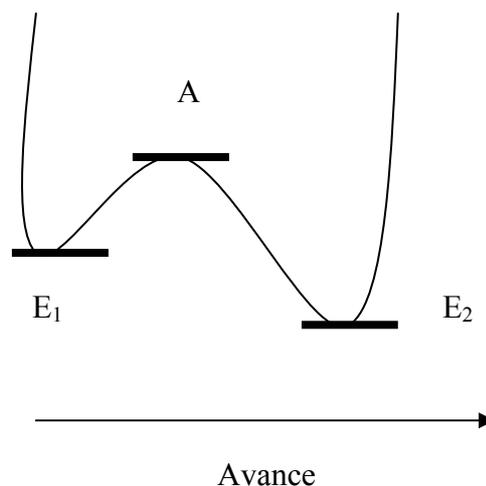
tratar de resolver la siguiente pregunta: ¿cuánta energía libre se necesita para lograr una computación en un tiempo finito? Para ello utilizaremos las nociones que ya tenemos sobre computación reversible.

Ya hemos comentado que cuando un proceso computacional se realiza infinitamente despacio y de forma reversible el coste de energía libre es cero. Pero ahora estamos ejecutando el proceso a una velocidad v . En este caso, se puede demostrar que la energía libre puesta en juego en cada paso computacional es:

$$\frac{F}{\text{paso}} = kT \ln(v)$$

Claramente la energía libre en cada paso disminuye con la velocidad de la computación.

Imaginemos ahora que tenemos un dispositivo en un estado particular con una energía asociada. El sistema puede avanzar o retroceder a un nuevo estado. Este proceso de avance o retroceso de estados implica realizar una computación –en el caso de avanzar a otro estado-, o deshacer una computación –en el caso de retroceder a un estado previo-. Este escenario puede ser representado mediante un diagrama de niveles de energía como el de la figura.



Sea ahora nuestro ordenador en uno de los dos estados E_1 o E_2 , de forma que el sistema puede ir de E_1 a E_2 -lo que supone realizar una computación-, o de E_2 a E_1 -lo que supone deshacer una computación- (podemos visualizar la situación considerando una

partícula –bit- que puede estar en uno cualquiera de los fosos de potencial, E_1 o E_2 , y que puede pasar del uno al otro). En esta discusión asumimos que la energía del proceso disminuye a medida que la computación avanza¹⁴. El nivel energético A (los que se dedican a la cinética química lo saben muy bien), se denomina energía de activación, y en este caso representa la energía que hay que suministrarle al sistema para que se produzca una transición en cualquier sentido.

Tendremos que tener cuidado con las fluctuaciones térmicas, ya que si su energía es mayor que la energía de activación A, nuestro ordenador se moverá de forma aleatoria entre estados permitidos. Esto quiere decir que el dispositivo diseñado puede ir en cualquier sentido. En este contexto podemos calcular las correspondientes tasas de avance y retroceso que, dadas las características energéticas del dispositivo, no tienen por qué ser iguales.

En términos generales, la probabilidad de que el bit de nuestro sistema pase a un estado E_i equivale a la probabilidad de que –por alguna razón- adquiera una energía mayor o igual que la energía de activación A, y caiga en el foso de potencial de energía E_i . De acuerdo con el esquema de la figura anterior, el tránsito $E_1 \rightarrow E_2$ implica un avance computacional, y la energía necesaria para que dicho tránsito se produzca es $(A-E_1)$. Por contrario, para el retroceso computacional $E_2 \rightarrow E_1$ la energía necesaria es $(A - E_2)$.

Pero la mecánica estadística predice que la probabilidad de una transición entre dos estados de energía E_a y E_b es:

$$P_{E_a \rightarrow E_b} = C \exp\left(-\frac{\partial E}{kT}\right)$$

En esta expresión C es un factor relacionado con las fluctuaciones térmicas.

En términos absolutos, para calcular las tasas de avance y retroceso hay que incluir un nuevo factor –X- que depende de ciertas propiedades moleculares, de forma que:

¹⁴ Lo cual no debería sorprendernos si recordamos que $F = U - TS$

$$Tasa(avance) = CX \exp\left[-\frac{(A - E_1)}{kT}\right]$$

$$Tasa(retroceso) = CX \exp\left[-\frac{(A - E_1)}{kT}\right]$$

Como C y X no nos interesan demasiado en nuestra discusión, y puesto que lo que sí nos interesa es averiguar cómo evoluciona nuestra computación, podemos expresar lo anterior como una relación según la expresión:

$$\frac{Tasa(avance)}{Tasa(retroceso)} = \exp\left[\frac{E_1 - E_2}{kT}\right]$$

De este modo nos olvidamos de las molestas fluctuaciones térmicas, de las engorrosas propiedades moleculares, y de la indeseable energía de activación. Además, la última ecuación sólo depende de la diferencia energética entre estados sucesivos, de tal forma que cuanto mayor sea la diferencia energética ($E_1 - E_2$) más rápido será el proceso computacional, y recordemos que habíamos asumido que la energía del proceso disminuía a medida que la computación avanzaba.

Podría ser interesante recordar ahora la expresión general:

$$\frac{F}{paso} = kT \ln(v)$$

en donde v era la velocidad del proceso computacional. Desarrollando expresiones obtenemos que:

$$\ln\left[\frac{Tasa(avance)}{Tasa(retroceso)}\right] = \frac{E_1 - E_2}{kT} \rightarrow kT \ln\left[\frac{Tasa(avance)}{Tasa(retroceso)}\right] = E_1 - E_2$$

Del desarrollo anterior se concluye que la velocidad del proceso computacional no es otra cosa que la relación de tasas de avance/retroceso, y que la energía libre puesta en

juego en cada paso computacional es, precisamente, la diferencia energética entre estados:

$$v = \frac{Tasa(avance)}{Tasa(retroceso)}$$

$$\frac{F}{paso} = E_1 - E_2$$

Analicemos ahora un problema parecido, aunque no idéntico, al que acabamos de plantearnos. La cuestión es: ¿Por qué –y cómo- una computación sigue una dirección determinada? Está claro que para que esta pregunta tenga sentido, nuestro bit –que está en un estado cualquiera E_a - debe poder ir a más de un estado nuevo. Lo que nos importa aquí no es la energía de los estados individuales, sino su accesibilidad desde un estado dado. Dicho de otro modo, el ordenador tiene que poder seleccionar el nuevo estado hacia el que va a evolucionar, no en base a la energía de los nuevos estados, sino en función del número de estados energéticamente equivalentes que son accesibles. Visto así, podemos decir que nuestro ordenador funciona por “difusión”, ya que siempre es más probable avanzar hacia estados más accesibles.

Supongamos que durante un proceso computacional nos encontramos con n_1 estados energéticamente equivalentes entre sí, desde los que podemos acceder en un solo paso a n_2 nuevos estados –también energéticamente equivalentes entre sí-, con las siguientes restricciones:

$$E(n_1) > E(n_2) : n_1 < n_2$$

En este caso:

$$v = \frac{Tasa(avance)}{Tasa(retroceso)} = \frac{n_2}{n_1}$$

Necesitamos ahora volver a un viejo concepto ya casi olvidado, la entropía $S = k \ln \omega$ en la cual ω es la probabilidad de una configuración. Por lo tanto, desarrollando la expresión definida para v , obtenemos que:

$$\frac{F}{\text{paso}} = kT \ln(v) = kT [\ln(n_2) - \ln(n_1)] = (S_2 - S_1)T$$

Es evidente que, de acuerdo con esta expresión, el gasto de energía libre en cada paso computacional equivale a la entropía generada en ese paso multiplicada por la temperatura.

6. CONCLUSIONES

Al escribir estas páginas hemos pretendido ilustrar cómo la física clásica puede ser utilizada para tratar de comprender algunos aspectos interesantes de la computación en general, y –más específicamente- de las ciencias de la computación en particular. Es evidente que algunas de las cuestiones planteadas rozan también la tecnología de la computación o la lógica, por sólo citar dos disciplinas afines a las ciencias de la computación, pero con aproximaciones completamente diferentes: la tecnología de la computación es una materia de marcado carácter práctico, mientras que la lógica es precisamente eso, lógica. Lo maravilloso del asunto es que ambas, bajo el paraguas de las ciencias de la computación, se utilizan para construir ordenadores, cuyo comportamiento, a su vez, puede ser explicado desde la física. En este caso nos hemos ocupado del comportamiento de los ordenadores clásicos y, por lo tanto, hemos tenido que utilizar la física tradicional. No obstante, las discusiones efectuadas sobre las puertas reversibles y sobre la computación reversible abren el camino de la computación cuántica, que será lo siguiente que abordaremos.

Hemos especulado con cierto detalle sobre cuestiones fundamentales de la computación clásica. Para ello hemos utilizado con profusión el razonamiento por analogía. En este sentido, puede que algunos resultados obtenidos no sean rigurosamente ciertos, aunque hayan sido rigurosamente tratados. En cualquier caso los temas planteados se basan en abstracciones más o menos claras, cuyo desarrollo es una excelente gimnasia mental,

que además contribuye a la génesis de animadas discusiones. Y éste ha sido, precisamente, uno de los objetivos fundamentales del trabajo. A fin de cuentas en la Universidad se nos paga para que pensemos.

Algunos principios básicos de la física han tenido que ser introducidos por decreto. Pero esto no debería escandalizar a nadie. Basta con asumirlos y trabajar con ellos para tratar de obtener explicaciones a fenómenos que, aunque son habitualmente manejados por los informáticos, no son habitualmente explicados en las aulas... es la pescadilla que se muerde la cola.

A lo largo de este trabajo nos hemos concentrado en tres cuestiones básicas, una de ellas relativa a la forma de plantear los temas, y las otras dos relativas a los contenidos globales desarrollados. Así, hemos pretendido abordar temas ya sabidos, pero desde puntos de vista diferentes a los habituales. Por otra parte, los aspectos energéticos de la computación –reversible e irreversible-, y los límites teóricos que la física clásica impone a la computación clásica, han sido objeto de constante atención a lo largo de lo desarrollado en este texto. Pero... ¿qué pasará cuando el tamaño de los ordenadores sea tal que las leyes de la física de lo muy pequeño empiecen a tener una importancia grande? ¿cuáles serán las limitaciones que la física cuántica impondrán a la computación? De estos, y de otros temas relacionados, trataremos cuando hablemos de computación cuántica.

7. BIBLIOGRAFÍA

- R. Feynman. Conferencias sobre computación. Crítica, eds., 2003
- G.Gamow. Un, dos, tres,... infinito. RBA, eds., 1993
- I.Pohl, A.Shaw. The nature of computation: an introduction to computer science. Computer Sciences Press, 1981
- C.H.Bennett. Logical reversibility of computation. IBM Journal of Research and Development, 17, 525, 1973
- C.H.Bennett. Thermodynamics of computation. International Journal of Theoretical Physics, 48, 1581, 1982
- E.Fredkin, T.Toffoli. Conservative logic. International Journal of Theoretical Physics, 21, 905, 1982
- S.Haykin. Neural networks: a comprehensive foundation. MacMillan College Publishing Company, 1994