3 Entropy

Three approaches to entropy are compared: thermodynamic (macro physics), statphysical (micro physics) and informational (mathematics).

3a A framework

Now we are in position to return to the notions introduced tentatively in Sect. 1 and treat them in the framework of ideal physical systems in general.

Let Ω and μ be as in Sect. 1a.[1] Given measurable functions $f, g : \Omega \to \mathbb{R}$, we consider the linear space L of all linear combinations $h = \alpha f + \beta g$ ($\alpha, \beta \in \mathbb{R}$), and the subset of all h that satisfy (1a1) or (1a2).

3a1 Exercise. (a) If $\mu(\Omega) < \infty$ then this subset is a linear subspace;

(b) if $\mu(\Omega) = \infty$ then this subset is a cone without 0.

Prove it.

The same holds for any finite-dimensional linear space L. We assume that the cone has non-empty interior, and denote its interior by K. If $\mu(\Omega) < \infty$ then $K = L$; otherwise K is a cone without 0, and $L = K - K$. We also assume that L does not contain constant functions (except for 0, of course).

We introduce $\Lambda : L \to (-\infty, \infty]$ by

$$\Lambda(f) = \ln \int \mathrm{e}^f \, \mathrm{d}\mu$$

and note that $(-K) \subset \mathrm{Int}\{\Lambda < \infty\}$; by 2g2 (generalized to n dimensions), Λ is infinitely differentiable on $(-K)$.

[1] That is, μ is a finite or σ-finite positive measure on Ω.

If $h \in K$ and $g \in L$ then $h \pm \varepsilon g \in K$ for ε small enough (think, why). Theorem 1b4 gives as $[g|h]$, provided that $h \ne 0$ (and therefore $h \ne \text{const}$). Moreover, Sect. 2j tells us that

$$[g|h](a) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \Lambda(\lambda h + \varepsilon g)$$

whenever

$$a = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \Lambda(\lambda h + \varepsilon h)\,;$$

here

$$\begin{array}{ll} \lambda \in (-\infty, +\infty) & \text{if } \mu(\Omega) < \infty; \\ \lambda \in (-\infty, 0) & \text{if } \mu(\Omega) = \infty. \end{array}$$

In a more physical style, we let $\beta = -\lambda$ and $u = a$:

$$[g|h](u) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \Lambda(-\beta h + \varepsilon g)$$

whenever

$$u = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \Lambda(-\beta h + \varepsilon h)\,;$$

here and henceforth

$$\begin{array}{ll} \beta \in (-\infty, +\infty) & \text{if } \mu(\Omega) < \infty; \\ \beta \in (0, \infty) & \text{if } \mu(\Omega) = \infty. \end{array}$$

On the other hand,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \Lambda(-\beta h + \varepsilon g) = \langle g, \operatorname{grad} \Lambda(-\beta h) \rangle\,;$$

here $\operatorname{grad} \Lambda(-\beta h)$ is treated as a linear functional on L, that is, a vector of the dual space L^*,

$$\operatorname{grad} \Lambda(-\beta h) \in L^*\,.$$

Thus,

$$[g|h](u) = \langle g, \operatorname{grad} \Lambda(-\beta h) \rangle$$

whenever

$$u = \langle h, \operatorname{grad} \Lambda(-\beta h) \rangle\,.$$

We see that $x_{h,u} = \operatorname{grad} \Lambda(-\beta h)$ is the equilibrium macrostate of the system. Indeed, conditionally, given $h^{(n)} \approx u$, we have[1] $g^{(n)} \approx \langle g, x_{h,u} \rangle$ (recall Sect. 1c).

[1] For large n, with high probability, as before. . .

Similarly to Sect. 2j we introduce the set

$$T = \{\operatorname{grad}\Lambda(f) : f \in (-K)\} \subset L^* .$$

Using a basis $(f_1, \dots, f_d)$ of the linear space L we may treat L and L^* as $\mathbb{R}^d$.

3a2 Example. Let (Ω, μ) be $(\mathbb{R}^2, \gamma^2)$ as in Sect. 1c, and L consist of linear functions $f(\omega) = \langle f, \omega \rangle$ on $\mathbb{R}^2$. Then $K = L = \mathbb{R}^2$;

$$\Lambda(f) = \ln \int \mathrm{e}^{\langle f,\omega\rangle}\, \gamma^2(\mathrm{d}\omega) = \ln \mathrm{e}^{\|f\|^2/2} = \frac12 \|f\|^2 ;$$

$\operatorname{grad}\Lambda(f) = f$ (here $L^* = \mathbb{R}^2 = L$); $T = \mathbb{R}^2$; $x_{h,u} = -\beta h$ for $u = -\beta\|h\|^2$. It conforms to the formula $S_{f,a} = af/\|f\|^2$ of Sect. 1c.

3a3 Example. (Spin 1/2.) Let (Ω, μ) be $\{-1, 1\}$ with the counting measure, and L consist of functions $\omega \mapsto f\omega$, $f \in \mathbb{R}$. Then $K = L = \mathbb{R}$;

$$\begin{gathered}
\Lambda(f) = \ln \int \mathrm{e}^{f\omega}\, \mu(\mathrm{d}\omega) = \ln(\mathrm{e}^{-f} + \mathrm{e}^{f}) ; \\
\operatorname{grad}\Lambda(f) = \Lambda'(f) = \frac{\mathrm{e}^f - \mathrm{e}^{-f}}{\mathrm{e}^f + \mathrm{e}^{-f}} = \tanh f ; \\
T = (-1, 1) ; \\
u = \langle h, \operatorname{grad}\Lambda(-\beta h)\rangle = -h \tanh \beta h .
\end{gathered}$$

For small β we have $\tanh\beta h \approx \beta h$, thus, $u \approx -\beta h^2$ and $x_{h,u} \approx -\beta h = u/h$, the same as in Example 3a2 for $(\mathbb{R}^1, \gamma^1)$. This is why the latter can approximate spin systems at high temperatures (as was promised in Sect. 1c).

As was noted in Sect. 1a, given two systems described by (Ω_1, μ_1) and (Ω_2, μ_2), the combined system is described by the product space $(\Omega_1 \times \Omega_2, \mu_1 \times \mu_2)$, and if they do not interact then $L = L_1 \oplus L_2$, that is, every $h \in L$ is of the form $f \oplus g : (\omega_1, \omega_2) \mapsto f(\omega_1) + g(\omega_2)$. We have

$$\begin{aligned}
\Lambda(h) = \ln \int \mathrm{e}^h \,\mathrm{d}\mu = \ln \iint \mathrm{e}^{f(\omega_1)+g(\omega_2)}\, \mu_1(\mathrm{d}\omega_1)\mu_2(\mathrm{d}\omega_2) = \\
= \ln \biggl(\Bigl(\int \mathrm{e}^{f(\omega_1)}\, \mu_1(\mathrm{d}\omega_1) \Bigr) \Bigl(\int \mathrm{e}^{g(\omega_2)}\, \mu_2(\mathrm{d}\omega_2) \Bigr) \biggr) = \Lambda_1(f) + \Lambda_2(g) ;
\end{aligned}$$

in this sense, $\Lambda = \Lambda_1 \oplus \Lambda_2$. Therefore $\operatorname{grad}\Lambda = \operatorname{grad}\Lambda_1 \oplus \operatorname{grad}\Lambda_2 \in L_1^* \oplus L_2^* = L^*$.

Sometimes a physical system does not seem to be combined, but can be treated as combined; see the next example.

3a4 Example. (Ideal gas.[1]) Let $\Omega = V \times \mathbb{R}^3$ where $V \subset \mathbb{R}^3$ ("container") is a domain, μ is the Lebesgue measure (six-dimensional, restricted to Ω). The Hamiltonian is

$$h(q,p) = \frac{1}{2m}\|p\|^2 + U(q) \quad \text{for } q \in V\,,\ p \in \mathbb{R}^3\,;$$

here m is the mass of the particle, q its coordinate, p its momentum,[2] $\frac{1}{2m}\|p\|^2$ its kinetic energy, and $U(q)$ its potential energy. We may treat q and p as separate systems,[3] and further, we may split the three-dimensional momentum into three one-dimensional momenta.

3a5 Example. (One-dimensional momentum.) Let $\Omega = \mathbb{R}$, μ the Lebesgue measure, $h(p) = \frac{1}{2m}p^2$ (the Hamiltonian, not to be changed), and $L = \{\lambda h : \lambda \in \mathbb{R}\}$. We have $L = \mathbb{R}$, $K = (0,\infty)$,

$$\Lambda(\lambda h) = \ln \int \mathrm{e}^{\lambda h}\,\mathrm{d}\mu = \ln \int \exp(\lambda p^2/2m)\,\mathrm{d}p = \ln\sqrt{\frac{2\pi m}{-\lambda}} = \mathrm{const} - \frac{1}{2}\ln(-\lambda)$$

for $\lambda < 0$ (and $+\infty$ otherwise). Thus, $\operatorname{grad}\Lambda(\lambda h) = -\frac{1}{2\lambda}$, that is,

$$x_{h,u} = \operatorname{grad}\Lambda(-\beta h) = \frac{1}{2\beta} \quad \text{for } \beta \in (0,\infty)\,;$$
$$T = (0,\infty)\,;$$
$$u = \langle h, \operatorname{grad}\Lambda(-\beta h)\rangle = \frac{1}{2\beta}\,.$$

We define a quasistatic process as a pair of functions,

$$[0,t_{\max}] \ni t \mapsto h_t \in K \setminus \{0\}\,,$$
$$[0,t_{\max}] \ni t \mapsto \beta_t \in \begin{cases} (-\infty,+\infty) & \text{if } \mu(\Omega) < \infty, \\ (0,+\infty) & \text{if } \mu(\Omega) < \infty; \end{cases}$$

both functions are assumed to be piecewise smooth. Usually we assume $t_{\max} = 1$.

Given a quasistatic process, we define

$$x_t = \operatorname{grad}\Lambda(-\beta_t h_t) \in L^*$$

(the equilibrium macrostate), and

$$u_t = \langle h_t, x_t\rangle$$

[1]Monatomic, classical (Maxwell-Boltzmann).

[2]In fact, $p = m\dot{x}$, but we do not need it.

[3]Not in dynamics, of course, but in equilibrium statistical physics.

(the energy[1]). We split the energy received by the system,

$$u_1 - u_0 = \int_0^1 \bigl(\langle h'_t, x_t \rangle + \langle h_t, x'_t \rangle \bigr) \, \mathrm{d}t$$

into the mechanical part (work) defined by

$$\int_0^1 \langle h'_t, x_t \rangle \, \mathrm{d}t$$

and the thermal part (heat) defined by

$$\int_0^1 \langle h_t, x'_t \rangle \, \mathrm{d}t \, .$$

A quasistatic process is called adiabatic, if

$$\langle h_t, x'_t \rangle = 0 \quad \text{for all } t \, ,$$

and isothermal, if

$$\beta_t = \beta_0 \quad \text{for all } t \, .$$

3b Thermodynamic entropy as adiabatic invariant

Given an initial state $x_0 \in L^*$, can we arrive at an arbitrary $x_1 \in L^*$ by an adiabatic process? Or maybe such x_1 must belong to some surface?

It is easy to guess that the relation $x_t = \operatorname{grad} \Lambda(-\beta_t h_t)$ leads to $\beta_t h_t = -(\operatorname{grad} \Lambda)^{-1}(x_t)$ and so, for every adiabatic process,

$$\langle (\operatorname{grad} \Lambda)^{-1}(x_t), x'_t \rangle = 0 \, .$$

Thus, a function $S : L^* \to \mathbb{R}$ such that $\operatorname{grad} S(x)$ is collinear with $(\operatorname{grad} \Lambda)^{-1}(x)$ for all x, must be an adiabatic invariant, which means, $S(x_t) = \mathrm{const}$ for every adiabatic process.

However, existence of such S is not at all automatic. For example, there is no non-constant $S : \mathbb{R}^3 \to \mathbb{R}$ such that $\operatorname{grad} S(x, y, z)$ is collinear with (z, x, y).[2]

On the other hand, such S exists in the special case of $(\mathbb{R}^2, \gamma^2)^n$, recall Sect. 1g.

---

[1] Internal energy.

[2] The example is taken from Wikipedia, "Integrability conditions for differential systems".

3b1 Theorem. There exists a nonconstant continuous function $S : T \to \mathbb{R}$ such that $S(x_t) = S(x_0)$ for all t and every adiabatic quasistatic process.

A stronger theorem 3c1 will be proved in Sect. 3d. In Sect. 3c one of such functions S will be singled out and called the thermodynamic entropy.

3b2 Exercise. Find at least one such S (not using 3b1) for each one of 3a2, 3a3 and 3a5.

3c Thermodynamic entropy as rate function

Recall the Fenchel-Legendre transform $\Lambda^* : L^* \to (-\infty, \infty]$ introduced in Sect. 2g:

$$\Lambda^*(x) = \sup_{f \in L} (\langle f, x \rangle - \Lambda(f)) \quad \text{for } x \in L^* \, ;$$
$$\Lambda^*(\operatorname{grad} \Lambda(f)) = \langle f, \operatorname{grad} \Lambda(f) \rangle - \Lambda(f) \quad \text{for } f \in (-K) \, .$$

3c1 Theorem. $\Lambda^*(x_t) = \Lambda^*(x_0)$ for all t and every adiabatic quasistatic process.

The thermodynamic entropy[1] is the function $S : T \to \mathbb{R}$ defined by

$$S(x) = -\Lambda^*(x) \, .$$

It is an adiabatic invariant, which is a macroscopic property. And on the other hand, $nS(x)$ is roughly the logarithm of the number of microstates corresponding to the macrostate x, in the sense of (3c3) below.

3c2 Exercise. Let $f \in (-K)$ and $x = \operatorname{grad} \Lambda(f)$, then

$$\mu^n \{ |f^{(n)} - \langle f, x \rangle| \le \varepsilon_n \} = \exp\bigl(-n\Lambda^*(x) + o(n) \bigr)$$

for every sequence $(\varepsilon_n)_n$ such that $\varepsilon_n \to 0$ and $n\varepsilon_n \to +\infty$.

Prove it.

Taking $h \in K$, $f = -\beta h$ and $u = \langle h, x \rangle$ we get

$$\mu^n \{ |h^{(n)} - u| \le \varepsilon_n \} = \exp\bigl(nS(x) + o(n) \bigr) \tag{3c3}$$

and

$$S(x) = \beta u + \Lambda(-\beta h) \, .$$

---

[1]In physics, $k_{\mathrm{B}} S(x)$ is the entropy per particle, and $n k_{\mathrm{B}} S(x)$ is the entropy of the n-particle system.

3c4 Exercise. Calculate $S(\cdot)$ for each one of 3a2, 3a3 and 3a5.

Answers: 3a2: $S(x) = -\frac{1}{2}\|x\|^2$; 3a3: $S(x) = -\frac{1-x}{2}\ln\frac{1-x}{2} - \frac{1+x}{2}\ln\frac{1+x}{2}$; 3a5: $S(x) = \frac{1}{2}\ln(4\pi e m x)$.

3c5 Exercise. (a) If $\mu(\Omega) = 1$ then $S(\cdot) \le 0$;

(b) If μ is a counting measure then $S(\cdot) \ge 0$.

Prove it.

3d Proving the theorem

The function Λ is infinitely differentiable on $(-K)$, thus, $\operatorname{grad}\Lambda : (-K) \to L^*$ also is infinitely differentiable.

3d1 Exercise. $\displaystyle \frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2}\Big|_{\varepsilon=0} \Lambda(f + \varepsilon g) > 0$ for all $f \in (-K)$, $g \in L \setminus \{0\}$.

Prove it.

Thus, the Jacobian of $\operatorname{grad}\Lambda$ does not vanish on $(-K)$. It follows that the set $T = (\operatorname{grad}\Lambda)(-K)$ is open.

3d2 Exercise. Prove that $\operatorname{grad}\Lambda : (-K) \to L^*$ is one-to-one.

Thus, $\operatorname{grad}\Lambda : (-K) \to T$ is bijective, and the inverse function $(\operatorname{grad}\Lambda)^{-1} : T \to (-K)$ is infinitely differentiable on the open set T.

3d3 Exercise. For every $f \in (-K)$ and $x \in T$,

$$\langle f, x \rangle \le \Lambda(f) + \Lambda^*(x)\,,$$

and the equality holds if and only if $x = \operatorname{grad}\Lambda(f)$.

Prove it.

3d4 Exercise. If $f \in (-K)$ and $x = \operatorname{grad}\Lambda(f)$ then $f = \operatorname{grad}\Lambda^*(x)$.

Prove it.

(The same holds for $\operatorname{Int}\{\Lambda < \infty\}$ instead of $(-K)$, but we do not need it.)

Proof of Theorem 3c1. We have $x_t = \operatorname{grad}\Lambda(-\beta_t h_t)$, therefore $-\beta_t h_t = \operatorname{grad}\Lambda^*(x_t)$, and

$$\frac{\mathrm{d}}{\mathrm{d}t}\Lambda^*(x_t) = \langle \operatorname{grad}\Lambda^*(x_t), x_t' \rangle = \langle -\beta_t h_t, x_t' \rangle = -\beta_t \langle h_t, x_t' \rangle = 0$$

since the process is adiabatic. □

3d5 Remark. For every quasistatic process, for all t,

$$\frac{\mathrm{d}}{\mathrm{d}t} S(x_t) = \beta_t \langle h_t, x_t' \rangle\,.$$

3e Informational entropy

Given a finite probability space (Ω, P) one may ask, how many points in Ω^n are needed in order to form a set of probability close to 1.

3e1 Theorem. Let (Ω, P) be a finite probability space, and

$$H(P) = -\sum_{\omega\in\Omega} p(\omega) \ln p(\omega) \, ,$$

where $p(\omega) = P(\{\omega\})$. Then

(a) there exist $A_n \subset \Omega^n$ such that $P^n(A_n) \to 1$ and $\frac{1}{n} \ln |A_n| \to H(P)$;[1]

(b) if $A_n \subset \Omega^n$ satisfy $P^n(A_n) \to 1$ then $\liminf_n \frac{1}{n} \ln |A_n| \ge H(P)$.

Consider the random variable $f : \Omega \to \mathbb{R}$ defined by $f(\omega) = -\ln p(\omega)$. Clearly, $\mathbb{E} f = H(P)$.

3e2 Exercise. There exist $B_n \subset \Omega^n$ such that $P^n(B_n) \to 1$ and $\inf_{B_n} f^{(n)} \to H$, $\sup_{B_n} f^{(n)} \to H(P)$.

Prove it.

3e3 Exercise. Prove that $\limsup_n \frac{1}{n} \ln |B_n| \le H(P)$.

3e4 Exercise. If $A_n \subset \Omega^n$ satisfy $P^n(A_n) \to 1$ then $\liminf_n \frac{1}{n} \ln |A_n \cap B_n| \ge H(P)$.

Prove it.

Theorem 3e1 follows immediately.

By definition, the entropy of P is $H(P)$.

If P is the uniform distribution on Ω then $H(P) = \ln |\Omega|$. Also, $H(P_1 \times P_2) = H(P_1) + H(P_2)$.

More generally, if μ is a (finite or σ-finite) measure on Ω and ν a probability measure on Ω absolutely continuous w.r.t. μ, then the differential entropy is defined by

$$H_\mu(\nu) = -\int_\Omega \Big(\ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \Big) \, \mathrm{d}\nu = -\int_\Omega \Big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \Big) \, \mathrm{d}\mu \, .$$

Similarly to 3e1,

(a) there exist measurable $A_n \subset \Omega^n$ such that $\nu^n(A_n) \to 1$ and $\frac{1}{n} \ln \mu(A_n) \to H_\nu(\mu)$;

(b) if measurable $A_n \subset \Omega^n$ satisfy $\nu^n(A_n) \to 1$ then $\liminf_n \frac{1}{n} \ln \mu(A_n) \ge H_\nu(\mu)$.

3e5 Exercise. Prove that $H_{\mu_1\times\mu_2}(\nu_1 \times \nu_2) = H_{\mu_1}(\nu_1) + H_{\mu_2}(\nu_2)$.

---

[1] Here $|A_n|$ is the number of points in A_n.

3f Relation between the two

Assume for now that μ is the counting measure on Ω (which is the case for spin systems). Then the microcanonical ensemble (recall Sect. 2l) may be defined as the uniform distribution on the set $\{|h^{(n)}-u| \le \varepsilon_n\}$ of $\exp(nS(x)+o(n))$ points (see (3c3)); thus, its informational entropy is $nS(x)+o(n)$, and the informational entropy per particle in the limit $n \to \infty$ is $S(x)$, just the thermodynamic entropy.

In general μ is not the counting measure. However, according to quantum mechanics, the physically relevant measure μ of a domain in the phase space Ω is, in some sense, roughly the number of "phase cells" in this domain.[1] Thus, counting measures are more relevant than it may seem, and the differential entropy is "more informational" than it may seem. Having this in mind we return to general measures μ. Still, the differential entropy (w.r.t. μ^n) per particle is $S(x)$ for the microcanonical ensemble.

The canonical ensemble is the measure ν^n, where ν is the tilted measure

$$\nu = \mathrm{e}^{-\beta h - \Lambda(-\beta h)} \cdot \mu = \frac{\mathrm{e}^{-\beta h} \cdot \mu}{\int \mathrm{e}^{-\beta h} \, \mathrm{d}\mu}$$

for given $h \in K$, β and $x = \operatorname{grad} \Lambda(-\beta h)$. Note that $x = \nu|_L$ in the sense that

$$\int g \, \mathrm{d}\nu = \langle g, x \rangle \quad \text{for all } g \in L \, .$$

What can be said about the differential entropy $H_\mu(\nu)$?

The canonical ensemble ν^n is equivalent to the microcanonical ensemble, as explained in Sect. 2l. That is, any macroscopic observable $g^{(n)}$ concentrates around the same value $\langle g, x \rangle$ in both ensembles. Does it mean that $H_{\mu^n}(\nu^n) = nS(x)$? No, since the informational entropy of the microcanonical ensemble is the average of a constant function, while $H_{\mu^n}(\nu^n) = \int n g^{(n)} \, \mathrm{d}\nu^n$ for $g = \ln \frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \beta h + \Lambda(-\beta h)$.

Take $\varepsilon_n \to 0$ such that $\varepsilon_n \sqrt{n} \to +\infty$ and consider $A_n = \{|h^{(n)} - u| \le \varepsilon_n\}$, where $u = \langle h, x \rangle$. Then $\nu_n(A_n) \to 1$ (think, why) and $\mu^n(A_n) = \exp(nS(x)+o(n))$ by (3c3). Therefore $S(x) \ge H_\mu(\nu)$. This is intriguing: are they equal?

Fortunately, it is easy to calculate $H_\mu(\nu)$:

$$H_\mu(\nu) = \int (\beta h + \Lambda(-\beta h)) \, \mathrm{d}\nu = \beta \langle h, x \rangle + \Lambda(-\beta h) = S(x) \, ;$$

the two ensembles have the same entropy,

$$H_\mu(\nu) = S(x) \, .$$

[1] Provided that the domain is much larger than a phase cell; otherwise classical mechanics is not a useful approximation.

3g Hints to exercises

3c2: recall 2f.

3c4: recall 2h6, and (for 3a3) the hint to 2d6.

3c5: use (3c3).

3d1: recall (2e3) and (2e1).

3d2: Λ is strictly convex on every straight line.

3d3: use the definition of Λ^* (and of T).

3d4: Hint: use 3d3.

3e3: $|B_n| \le \exp n \sup_{B_n} f^{(n)}$.

3e4: $|A_n \cap B_n| \ge P^n(A_n \cap B_n) \exp n \inf_{A_n \cap B_n} f^{(n)}$.

Index