Lecture Note:
Probabilities, Energy, Boltzmann & Partition Function

Marc Toussaint

Learning & Intelligent Systems Lab, TU Berlin

August 19, 2024

Probabilities & Energy

Given a density $p(x)$, we call

$$E(x) = -\log p(x) + c \ , \tag{1}$$

(for any choice of offset $c \in \mathbb{R}$) an energy function. Conversely, given an energy function $E(x)$, the corresponding density (called Boltzmann distribution) is

$$p(x) = \tfrac{1}{Z} \exp(-E(x)) \ , \tag{2}$$

where Z is the normalization constant to ensure $\int_x p(x) = 1$, and $c = -\log Z$. From the perspective of physics, one can motivate why $E(x)$ is called "energy" and derive these relations from other principles, as mentioned below. Here we first motivate the relation more minimalistically as follows:

Probabilities are *multiplicative* (axiomatically): E.g., the likelihood of i.i.d. data $D = \{x_i\}_{i=1}^n$ is the product

$$P(D) = \prod_{i=1}^n p(x_i) \ . \tag{3}$$

We often want to rewrite this with a log to have an *additive* expression

$$E(D) = -\log P(D) = \sum_{i=1}^n E(x_i) \ , \quad E(x_i) = -\log p(x_i) \ . \tag{4}$$

The minus is a convention so that we can call the quantity $E(D)$ a *loss* or *error* – something we want to minimize instead of maximize. We can show that whenever we want to define a quantity $E(x)$ that is a function of probabilities (i.e., $E(x) = f(p(x))$ for some f) and that is additive, then it *needs* to be defined as $E(x) = -\log p(x)$ (modulo a constant, where the minus is just a convention):

Proof. Let $P(D)$, $E(D)$ be two functions over a space of sets D, with properties (1) $P(D)$ is multiplicative, $P(D_1 \cup D_2) = P(D_1)P(D_2)$, (2) $E(D)$ is additive $E(D_1 \cup D_2) = E(D_1) + E(D_2)$, and (3) there is a mapping $P(D) = \frac{1}{Z} f(E(D))$ between both. Then it follows that

$$P(D_1 \cup D_2) = P(D_1)\ P(D_2) = \tfrac{1}{Z_1} f(E(D_1))\ \tfrac{1}{Z_2} f(E(D_2)) \tag{5}$$

$$P(D_1 \cup D_2) = \tfrac{1}{Z_0} f(E(D_1 \cup D_2)) = \tfrac{1}{Z_0} f(E(D_1) + E(D_2)) \tag{6}$$

$$\Rightarrow \quad \tfrac{1}{Z_1} f(E_1) \tfrac{1}{Z_2} f(E_2) = \tfrac{1}{Z_0} f(E_1 + E_2) \ , \tag{7}$$

where we defined $E_i = E(D_i)$. The only function to fulfill the last equation for any $E_1, E_2 \in \mathbb{R}$ is the exponential function $f(E) = \exp(-\beta E)$ with arbitrary coefficient β (and minus sign being a convention, $Z_0 = Z_1 Z_2$). Boiling this down to an individual element $x \in D$, we have

$$p(x) = \tfrac{1}{Z} \exp(-\beta E(x)) \ , \quad \beta E(x) = -\log p(x) - \log Z \ . \tag{8}$$

□

Partition Function

The normalization constant Z is called partition function. *Knowing the partition function means knowing probabilities instead of only energies.* This makes an important difference:

Energies $E(x)$ are typically assumed to be known (in physics: from first principles, in AI: can be pointwise evaluated for each x). However, the partition function Z is typically apriori unknown and evaluating it is a hard global problem: The number Z is a global property of the function $E(\cdot)$, while the energy $E(x)$ of a single state x can be evaluated from properties of x only.

To appreciate the importance of the partition function, consider we want to sample from a distribution $p(x)$ and we generate random samples x. For each x we can evaluate $p(x)$ and decide whether it is "good or bad", e.g., whether we reject it in rejection sampling with probability $1-p(x)$. Now consider the same situation, but we are missing the partition function Z. That is, we only have access to the energy $E(x)$, but not the probability $p(x)$. Sampling now becomes much harder, as we have no absolute reference about which energy $E(x)$ is actually "good or bad". Samplers in this case therefore need to resort to relative comparison between energies at different positions, e.g., before and after a random step and use the Metropolis Hasting ratio to decide whether the step was "good or bad". However, such methods are local: the chance to randomly step from one mode to a distant other mode may be very low (very long mixing times). If you knew the partition function, you would have an absolute scale to estimate the probability mass of each mode/cluster, and the *total* probability mass of all the modes you have yet found; but without the partition function there is no way at all to tell whether out there there might be other modes, other clusters with even lower energies that might, in absolute terms, attract a lot more probability mass. In this view, knowing the partition function is fundamentally global information about the absolute scaling of probabilities.

Concerning the word "partition function": We clarified that with knowing Z we know probabilities $p(x)$ instead of only energies $E(x)$. We can say the same for partitions of the full state space (e.g. modes, or energy levels): If you want to know how much probability mass $p(i)$ is in each partition i, you need the partition function (w.r.t. i). In statistical physics, partitions are often defined as states with same energy, and the question of how populated they are is highly relevant. This is where the word has its origin.

Energy & Probability in Physics

The set D in the proof above is a set of i.i.d. random variables. In physics, the analogy is a system composed of particles: Each particle has its own state. The full-system state is combinatorial in the particle states and full-system state probabilities multiplicative. Axiomatically, each particle also has a quantity called energy which is additive. Now, an interesting question is why that particular quantity that physicists call "energy" has a one-to-one relation f to probabilities – details can be found under the keyword "derivation of canonical ensemble" (e.g. Wikipedia), but two brief comments may clarify the essentials:

First, the one-to-one relation between energy and probability (in physics) only holds in the thermodynamic equilibrium. Second, systems (such as the canonical ensemble) have a fixed total energy, which is the sum of energies of all particles. Perhaps one can imagine that if one particle has particular high energy (e.g. almost all energy), there is not much energy left for the other particles, which means that they need to populate low energy states. If only discrete energy levels exist, one can count the combinatorices of how many full-system states exist depending which energy levels are populated – in the previous example the combinatorics is low because many particles are confined to low energy states, showing that the probability for this one particle to populate such a high energy level is low. In general, computing these combinatorics explicitly leads to the so-called Boltzmann distribution $p(i) \propto \exp(-\beta E_i)$ of a particle to be in energy level i: Lower energy states have higher probability, intuitively because for limited total energy more particles can be in these states, leading to higher combinatorics of microstates. In other words, in the thermodynamic equilibrium all feasible microscopic full-system states are equally likely, but the probability of one particle of the ensemble to have energy E_i goes with the Boltzmann distribution.

The coefficient β tells us how exactly energies translate to probabilities, and should intuitively depend on the total energy of the full system: if the full system has little total energy (is “cold”), higher energy states should become less likely. In physics, the factor is $\beta = \frac{1}{k_B T}$ with temperature T and the Boltzmann constant k_B which translates the system temperature to the average (thermal) particle energy $k_B T$. That is, the word “temperature” roughly means average particle energy, and in AI is often a freely chosen scaling factor of energies.