Notes for Dad

Peter Conlon

(Dated: January 1, 2024)

Casting aspects of statistical physics and kinetic theory seen through a lense of information theory and exponential families.

CONTENTS

I. SHANNON ENTROPY

Let there be a discretely enumerable space of outcomes of a random variable, labelled by i. Then a random variable, X, is the same idea as associating probabilities $0 \leq p_i \leq 1$ to each outcome i where $\sum_i p_i = 1$.

With respect to a probability distribution, define the 'surprise' of an outcome as $\log \frac{1}{p_i}$, equivalently $-\log p_i$. This definition of suprise intuitively conveys the idea that there is no surprise to entirely certain occurences $p_i = 1$, that surprise is always positive, and the log ensures that the total surprise of independent outcomes is additive, which feels right, since additivity is deeply related to independence.

Define the shannon entropy of a random variable (equivalently, of a distribution) as its expected surprise:

$$H(X) = \sum_i p_i \log \frac{1}{p_i} \tag{1}$$

The entropy is a function of the probabilities defining the distribution, or more informally, the entropy is a property of the distribution itself, which abstracts away the details of the space in which X takes values, and focuses entirely on the intrinsic uncertainty of the distribution over outcomes.

Among all possible distributions on space of W outcomes, the entropy is maximized when $p_i = \frac{1}{W}$, where it takes value $H(X^*) = \sum_i p_i \log W = \log W$.

Unconstrained maximization of entropy gives equal probability on all outcomes, but more frequently we will be interested in constrained maximizations where we maximize the entropy of a distribution, but only within the space of distributions which satisfy certain constraints - typically, that the expectation of some function on sample space takes a certain value.

These constraints will be motivated from the conserved quantities of underlying dynamics of physical systems. Ie, if process of coming to equlibrium characterises the 'forgetting' of initial state, then its important to make a distinction between 'forgettable' degrees of freedom in the dynamics, vs conserved quantities.

The conservation of energy and the locality of physical law, is a fundamental motivation for how temperature arises, representing the thermodynamic conjugate variable for the dynamically conserved energy.

A. Joint Entropy

A basic result is that the entropy of a joint-distribution is less than or equal to the the sum of the entropies of the marginals, with equality only with statistical independence.

Trivial consequence is that given statistical independence, the entropy of N-iid variables is simply $H(X^N) = NH(X)$.

B. Relative Entropy and Continuous Variables

The maximal attainable value of shannon entropy is $H(X^*) = \log W$. Achieved for probabilities $q_i = 1/W$. Note, because the q_i are independent of i we can write it as an average taken on any distribution

$$\log W = -\log q_i = -\sum_i p_i \log q_i \tag{2}$$

Let's define a quantity which is the (positive) shortfall from this maximum attainable value, ie $H * -H(X)$:

$$(-\sum_i p_i \log q_i) - (\sum_i p_i \log \frac{1}{p_i}) \tag{3}$$

which is an example of:

$$D_{\mathrm{KL}}(P\|Q) = -(\sum_i p_i \log \frac{q_i}{p_i}) \tag{4}$$

So maximising entropy of P is the same as minimizing the $D_{\mathrm{KL}}(P\|Q)$ (the divergence from Q to P), where Q is a maximum entropy distribution.

Writing it like this makes it possible to work with continuous variables, because we can only take logs of dimensionless quantities. Writing $\log p(x)$ where $p(x)$ is a probability density with units $[1/dx]$, is essentially ill defined, but $\log \frac{p(x)}{q(x)}$ is ok.

The subtleties here relates to proper measures: in discrete variables, the counting measure is available, but if we were to naively take a continous limit of shannon entropy everything becomes infinite via $\log W$. So really the right way to think about entropy is always relative between two distributions or probability measures.

II. THE FROZEN CLASSICAL GAS

Take a monoatomic ideal gas, very small hard spheres; and as a thought experiment, imagine we can pause the world, much like a photograph takes a snapshot. Then our first object of study will be these photographs.

To be concrete, consider a rectangular container with side lengths L and volume $V = L^d$ where we can imagine working in different dimensions $d = 1, 2, 3, \ldots$.

The parameters are the total number of gas particles, N. If we study unordered sets of photographs (ie, samples), we will discover a simple physical principle - that the total count of the particles is a conserved quantity. This is because the box is a container whose walls are impermeable.

How can we describe the statistics of the particle locations in the photographs?

Using $\mathbf{x}$ to mean a $d-$vector being the position of a particle, then configuration space for the frozen gas of N particles is $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$. And we shall put a probability density over configuration space.

Because it is an ideal gas (there are no interactions between the gas particles, except instantaneous collisions which exchange energy and momentum) then we write the joint density as a simple product of marginals:

$$f_N(\mathbf{x}_1, \ldots, \mathbf{x}_N) \prod_i^N d\mathbf{x}_i = \prod_i^N f(\mathbf{x}_i) d\mathbf{x}_i \tag{5}$$

ie, the particles in the gas are iid.

Analogously to the way the uniform distribution is maximum entropy for a discrete variable, the same argument carries over and maximum entropy distribution is uniform in the volume $V = L^d$.

III. THE MOVING CLASSICAL GAS

We can now go beyond the frozen gas, to consider the total phase space of the system $(\mathbf{x}, \mathbf{p})^N$.

The energy of a particle is $\frac{|\mathbf{p}|^2}{2m}$.

Again, the total distribution function can factorize into the product of N one-particle distribution functions.

The total distribution function of one particle in the gas is now $f(\mathbf{x}, \mathbf{p}) d\mathbf{x} d\mathbf{p}$ being interpreted as the probability to find the particle's having a particular position and momentum within a small $d\mathbf{x}d\mathbf{p}$.

Now unlike position, which is constrained by physical barriers, we don't have any physically motivated limit on a maximum momentum. However we do know that momentum can't be too big, because energy is a conserved quantity and $E \sim p^2/2m$, so if the gas has a total energy U then per particle it has an average energy U/N.

We can use the principle of maximum entropy again, to characterise the distribution. Supposing $f(\mathbf{x}, \mathbf{p}) \propto g(\mathbf{p})$ where $g(\mathbf{p})$ is a normalized distribution over $\mathbf{p}$.

Objective is to maximise the entropy $H(g)$ of the normalized distribution $g(\mathbf{p})$ subject to constraint on the distribution to have a specifc average $\mathbb{E}_g[T(\mathbf{p})] = U/N$ where kinetic enery $T(\mathbf{p}) = \frac{\mathbf{p}^2}{2m}$.

This is simple a constrained optimization, so we use the method of lagrange multipliers. Form the Lagrangian:

$$\mathcal{L}(g, \eta) = \underbrace{-\int g(\mathbf{p}) \log g(p) dp}_{H(g)} - \lambda \underbrace{\int g(\mathbf{p}) dp}_{\text{normalization}} + \eta \underbrace{\int T(\mathbf{p}) g(\mathbf{p}) dp}_{\text{constraint}} \tag{6}$$

The signs of the lagrange multipliers λ and η are chosen for later convenience.

Setting the functional derivative $= \frac{\delta \mathcal{L}}{\delta g}$ to 0 to find the extrema:

$$0 = -\log g(p) - 1 - \lambda + \eta T(\mathbf{p}) \tag{7}$$

which can be trivially rearranged to give

$$g(p) = \exp(\eta T(\mathbf{p}) - (\lambda + 1)) \tag{8}$$

Eliminating λ by the normalization condition, to express it in terms of η we can write:

$$g(p) = \exp(\eta T(\mathbf{p}) - A(\eta)) \tag{9}$$

where

$$A(\eta) = \log \int \exp(\eta T(\mathbf{p})) d\mathbf{p} \tag{10}$$

. This is our first example of an exponential family distribution.

The distribution of momenta is a gaussian distribution - because a gaussian distribution of momenta is the maximum entropy distribution, given average energy.

Let's compute $A(\eta)$:

$$A(\eta) = \log \int \exp(\eta \frac{\mathbf{p}^2}{2m}) d\mathbf{p} \tag{11}$$

so by Gaussian integrals in dimension d:

$$A(\eta) = \log\left(\frac{2\pi m}{\eta}\right)^{d/2} \tag{12}$$

or

$$A(\eta) = \frac{d}{2}(\log 2\pi + \log m - \log \eta) \tag{13}$$

The $\nabla_\eta A(\eta) = \frac{d}{2}(-\eta^{-1})$.

That is, the expected value of the energy is linear in $-\eta^{-1}$. This motivates introducing a quantity $\theta = -\eta^{-1}$ which we can call the *temperature*.

IV. CRASH COURSE IN EXPONENTIAL FAMILY DISTRIBUTIONS

Maximum entropy distributions take the form of exponential family distributions. These have many wonderful properties, such as the gradient of the log normalizer with respect to the natural parameters is the expectation of the statistics.

A distribution is exponential family if its density can be written in canonical form:

$$p(x|\eta) = \exp(\eta \cdot T(x) - A(\eta) + k(x)) \tag{14}$$

The η is a vector of natural parameters, the $T(x)$ is a vector of the sufficient statistics, the $A(\eta)$ is the log-normzalizer and $k(x)$ is the carrier measure.

For this crash course, let $k(x) = 0$.

The quantity $A(\eta)$ is the log-normalizer because $A(\eta) = \log Z(\eta)$ where

$$p(x|\eta) = \frac{1}{Z(\eta)} e^{\eta \cdot T(x)} \tag{15}$$

where

$$Z(\eta) = \int e^{\eta \cdot T(x)} dx \tag{16}$$

Exponential family distributions allow to efficiently compute expectations via taking derivatives, even though usually an expectation or variance feels like it should require an integral.

The vector of expectations of the sufficient statisitcs of an exponential family are $\mu = \nabla_\eta A(\eta)$.

This follows easily from the Feynman differentiate under the integral trick:

$$\nabla_\eta A(\eta) = \frac{1}{Z(\eta)} \nabla_\eta Z(\eta) = \frac{1}{Z(\eta)} \int T(x) \exp(\eta \cdot T(x)) dx \tag{17}$$

The log normalizer in general is related to the cumulant generating function.

$$C(s) = A(s + \eta) - A(\eta) \tag{18}$$

Derivatives of $C(s)$ then setting s to 0 evaluate to the same as derivatives of $A(\eta)$, which is why cumulants (including mean, covariance) can easily be derived from derivatives of the log normalizer.

Appendix A: Useful Results

The standard 1 dimenstional gaussian integral.

$$\int_{-\infty}^{\infty} e^{-\frac{a}{2}x^2 + bx} dx = \left(\frac{2\pi}{a}\right)^{1/2} e^{\frac{b^2}{2a}} \tag{A1}$$