Notes on Statistical Mechanics and Thermodynamics relevant to Machine Learning

compiled by D. Gueorguiev, 1/7/2025

# Starting Note

This document does not aim to discuss in depth the complete set of topics covered in the classic Statistical Mechanics books. Rather the interest here is in statistical models and underlying physical mechanisms relevant to certain problems in the Machine Learning field in order to establish a clear link to and to provide useful insights to Machine Learning algorithms of interest.

# Introductory Material

## Classical Ideal Gas

The abstraction ‘Ideal Gas’ and the mathematical model behind it have many applications outside the Physical Sciences. Thus we start our journey into the world of Statistical Mechanics and Thermodynamics with the definition of Ideal Gas.

An ideal gas is different from real gases by the absence of interactions between particles which constitute the gas.

While idea gas is unrealistic model for real gasses it can be thought as a limit scenario for real gasses with sufficiently low densities. The simple model of ideal gas does not capture phase transitions present in real gasses.

## Phase Space of a Classical Gas

Let us consider particles contained in some specified volume. Each particle has a well-defined position and momentum. The position of every particle is represented as a point in some abstract space (*configuration space*) with an axis for every coordinate of every particle. The generalized coordinates of this abstract space are given as

(1)

The momentum of every particle can be represented as a point in momentum space – an abstract -dimensional space (*momentum space*), with axes for every component of the momentum of every particle.

(2)

The complete microscopic state of the system can be described by a point in *phase space* – an abstract -dimensional space with axes for every coordinate and every momentum component for all particles. Phase space is the union of configuration space and momentum space, .

(3)

The *kinetic energy* of the -th particle is given by the usual expression . The *total kinetic energy* of the system is the sum of the kinetic energies of all particles in the system.

(4)

Since, by definition, there are no interactions between the particles in an ideal gas, the potential energy of the system is zero.

(5)

**Distinguishability**:

Particles in the system are considered *distinguishable* so that the exchange of two particles results in a different microscopic sate. Thus every point in phase space represents a different microscopic state.

**Thermodynamic Entropy (Boltzmann definition):**

Let us denote with the thermodynamic entropy of a system – this quantity is associated with particular microstate defined by thermodynamic parameters such as temperature, volume, and energy. is related to the number of microstates which can result in a given microstate with the following expression:

( Boltzmann entropy law )

Boltzmann defines the thermodynamic entropy in terms of a composite system – that is, a system composed of two or more subsystems with some sort of constraint. An example of such composite system would be a volume of gas divided into two smaller volumes (or subsystems) by a partition. The partition acts as a constraint by restricting the number of particles in each subsystem to be constant. The removal of the partition would then allow the system to develop from a less probable macroscopic state to a more probable macroscopic state. The final state, after the composite system had come to equilibrium, would correspond to the most probable macroscopic state. The Second Law of Thermodynamics states that the thermodynamic entropy will be maximized in the equilibrium state. The comparison of these two properties of the equilibrium state led Boltzmann to associate the entropy with the logarithm of the probability of a macroscopic state. The thermodynamic entropy is defined to *within additive and multiplicative constant*.

**Independence of Positions and Momenta:**

The underlying assumptions of the ideal gas model is that the positions and the momenta of the particles are independent.

**Separation of Entropy into Two Parts**

Since the positions and momenta are independent, we can express their joint probability as a product of two functions - one of positions only and the other of momenta only.

(6)

According to Boltzmann’s entropy definition and (6) the total entropy will be expressed as a sum of the contributions of the positions and the momenta. The probability distribution in configuration space, , depends only on the volume and the number of particles, . Consequently, the configurational entropy, depends only on and ; that is, .

The probability distribution in momentum space, , depends only on the total energy, , and the number of particles, . Thus, we write:

(7)

The thermodynamic quantities , , and are referred to as *extensive* parameters (aka *observables*, *variables*) because they measure the amount or extent of something. They are to be contrasted with *intensive* parameters, such as temperature or pressure, which do not automatically become bigger for bigger systems.

**Distribution of Particles between Two Subsystems**

Let us consider a composite system consisting of two boxes (subsystems) containing a total of distinguishable, non-interacting particles. We will name the boxes and , with volumes and . The total volume is . The number of particles in is , with being the number of particles in .

We can either constrain the number of particles in each box to be fixed or allow the numbers to fluctuate by making a hole in the wall separating the boxes. The total number of particles is constant in either case.

We are interested in the probability distribution for the number of particles in each subsystem after the constraint implemented as the impenetrable wall in the middle is removed by making a hole in the middle.

For simplicity we assume that the positions of the particles are mutually independent of each other besides being independent on momentum. The probability distribution of the configuration space can be written as:

(8)

If we further assume that a given particle is equally likely to be anywhere in the composite system, the probability of it being in subsystem is .

Figure: Box and Box with volumes and and number of particles and accordingly

If there are particles that are free to go back and forth between the two subsystems, the probability distribution of is given by the binomial distribution

(9)

with the constraint that .

The average value of and from the binomial distribution is

, (10)

Thus

(11)

Stirling’s approximation: (12)

Gosper’s form of the Stirling approximation:

(13)

(14)

We substitute the simplified expression of (13) in (14) :

(15)

Let us compute the maximum of allowing to be continuous variable:

(16)

(17)

which becomes

(18)

Thus we get

(19)

Similarly one can write:

(20)

If we compare (19) and (20) to (11) we will notice

and (21)

Thus we conclude the value of corresponding to the maximum of the probability distribution using the Gosper’s form of the Stirling approximation is equal to the mean value .

**Actual Number versus Average Number**

The actual number of particles, , is a property of the system, it is integer and fluctuates over time.

The average number of particles, is part of a description of the system and not property of the system itself.

It is not an integer and is independent of time.

The width of the probability distribution for and therefore the magnitude of the fluctuations of the actual number of particles is given by the standard deviation :

(22)

Note that from (11) and (22) follows that is of order .

Thus the relative width of the probability distribution is given by

(23)

The relative width is proportional to which becomes very small for macroscopic system.

**The Thermodynamic Limit**

The thermodynamics is only valid in the limit of infinite system size. We carry out experiments on finite systems which imply that thermodynamics is just an approximation which works well in large enough real systems. For macroscopic systems containing or more molecules, this means that the statistical uncertainties are of the order or less, which is several orders of magnitude smaller than typical experimental errors.

**Configurational Entropy**

We have shown already that the equilibrium value of is determined by the maximum of the probability distribution given by

with . (24)

Let us introduce a new function

(25)

which will provide additional insights on the composite system.

With (25) in mind we rewrite (24) as

(26)

Observation: Since the logarithm is monotonic function of its argument, the maximum of and the maximum of occur at the same values of and . Using is convenient as it decomposes into three distinct terms.

(27)

The first term on the RHS (in both forms of the equation) depends only on the variables for the subsystem , the second depends only on the variables for subsystem , and the third term depends only on the variables for the total composite system. It will be convenient to define a function

(28)

where and are both (at this point) arbitrary constants.

Let us define the function

(29)

Thus

(30)

The maximum of the function with the usual constraint that , then gives us the location of the equilibrium value of the average number of particles . This comes from (16) which gives the argument maximum of the first term of (29) while the second term is constant.

We have seen that the width of the probability distribution is proportional to . Since the probability distribution is normalized, the value of its peak must be proportional to . This also can be seen from the Gaussian approximation of the Binomial distribution (see eq. (A.7) ) if we substitute with , with and with :

(31)

At the equilibrium values of and , this gives

(32)

This is true because the exponent factor becomes one at the equilibrium value .

Since the function is of order , and , the term in (29) is negligible at equilibrium values of and . Therefore, *in equilibrium*, we have:

(33)

We can now identify as part of the entropy of the composite system that is associated with the configurations or the positions of the particles. will be denoted as the ***total configurational entropy of the composite system*.** The functions and will be denoted as the ***configurational entropies of subsystems and*** .

We have therefore found functions of the variables of each subsystem, such that the maximum of their sum yields the location of the equilibrium values

(34)

subject to the usual constraint that .

**Note 1**: the configurational entropy has the *additivity property* – the contributions to the configurational entropy from each subsystem are simply added to find the configurational entropy of the composite system. Another interpretation of this property is separability since we first derived the entropy of the composite system and were able to separate the expression into the sum of individual contributions from each subsystem.

**Note 2**: The function follows the Boltzmann definition of thermodynamic entropy as specified earlier in this document. This entropy is the logarithm of the probability , within additive and multiplicative constants and is maximized at equilibrium.

**Analytic Approximation for the Configuration Entropy**

The expression for the configurational entropy (28) becomes simpler analytically and hence more useful when we introduce the Stirling approximation for factorial in it.

//TODO: finish this section

# References

[1] [An Introduction to Statistical Mechanics and Thermodynamics, Robert H. Swendsen, 2012](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/books/An-Introduction-to-Statistical-Mechanics-and-Thermodynamics-Swendsen-2012.pdf)

[2] [Statistical Physics by L.D. Landau and E.M. Lifshitz, 2nd Edition, 1970 (orig. 1958)](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/books/LandauLifshitz-StatisticalPhysics_1958.pdf)

# Appendix

## Gaussian Approximation to the Binomial Distribution

Recall the binomial distribution is expressed as

(A.1)

The Central limit theorem tells us that for a fixed value of and large values of , the binomial distribution can be approximated by a Gaussian function. Consider a Gaussian function with mean and :

(A.2)

It is easy to show that the location of the mean and maximum coincide:

(A.3)

The variance is (A.4)

To approximate binomial distribution with Gaussian one all we need to do is to equate the mean and the variance of each of those

(A.5)

(A.6)

(A.7)