CMPTG 5 - Statistical Physics of Neural Networks

Sidharth Kannan

April 2025

1 Lecture 1

1.1 Introduction

This course is about the interplay between *statistical mechanics* and *machine learning*, and so it would likely be apt to begin by defining those two terms.

Statistical mechanics is a branch of physics that emerged in the mid-19th century to provide a microscopic account of thermodynamic phenomena. It deals primarily with the problem of modeling large collections of particles, so numerous that even if we knew the fundamental, *microscopic* laws that determined each particle's motion, it would simply be impossible to carry out the computation that would tell us their behavior. How do we proceed then? Well, it turns out that often, we are not interested in the specific behavior of each atom, but rather we care about large, *macroscopic* properties of the entire system. For example, if I have a container of neon gas, and I want to know what will happen to it if I heat it, I am not so concerned with the specifics of the motion of each atom, but rather the aggregate behavior of the whole volume. Making some simple assumptions, it turns out that there is actually quite a lot that we can say about the aggregate behavior of particles, even when we cannot speak to their individual behaviors.

Machine learning is a very interdisciplinary field, but its primary concern is how to build algorithms or models that, given some examples, termed a *dataset*, can *learn* to perform some downstream task. For example, given a set of images of hand written digits, learn a computer program that can identify handwritten digits. The specific type of machine learning that has really taken the world by storm these past few years is called "deep learning." The name derives from the structure of the models under consideration.

1.2 Logistics and Syllabus

The rough plan for the course is to divide it into two sections. The first, I will title *Learning as Inverse Thermodynamics*. The goal for this part of the course is to understand how we can formulate the learning problem as the reverse of the time evolution of a physical system. In particular, we will discuss two classes of machine learning models: *energy based models* and *diffusion* models.

The second part of the course will be about *field theories of neural networks*. In this part, we will focus more on the theory of neural networks, and how ideas from statistical mechanics, that were first used to understand the interactions of gas particles can be applied to complex systems of interacting neurons. Despite this being a theory course, throughout, I will try to highlight the practical relevance of what we are discussing.

Tentative Schedule of Topics:

1. *Machine Learning as Inverse Thermodynamics*
    (a) Equilibrium stastical mechanics and energy based models
        i. Introduction to statistical mechanics, the Boltzmann Law
        ii. Binary stochastic neurons, Ising Machines, and optimization
        iii. The Boltzmann Machine
    (b) Generative diffusion models and non-equilibrium thermodynamics
        i. Score-based modeling
        ii. Stochastic processes, Brownian motion, and Langevin dynamics
        iii. The Itô calculus, stochastic differential equations, and diffusion
        iv. Fokker-Planck equations and flows
2. *Field Theories of Neural Networks*
    (a) Intro to field theories of neural networks, mathematical preliminaries
    (b) Deep linear networks
    (c) Understanding oversmoothing in Graph Neural Networks
    (d) Neural Tangent Kernels and neural network Hamiltonians

The background that I will be assuming is a basic proficiency in machine learning, and a strong grasp of probability. That is to say that I won't be going over things like gradient descent, back-propagation, multilayer perceptrons, etc. in much detail. I also will assume basic knowledge of linear algebra, probability, and calculus, to the level of a typical machine learning class. We are covering some very advanced topics at the very cutting edge of machine learning and generative AI. As such, we will use probability and calculus in particular at a higher level than a typical intro ML class.

A brief note about logistics. This course is 2 units, P/NP. As such, the workload will be minimal. In order to achieve a passing grade, you just need to attend every week. Every week or every other week, I will also post some useful exercises. These will either be calculations, proofs, or simple programming exercises. They are entirely optional, and are just intended as a learning tool to help solidify your understanding of concepts that may have been unclear. If you do happen to miss a class, you can make that up by doing some number of the exercises from that week.

I have LaTeXed notes that I will post every week for those unable to attend. If you have questions about this stuff, you can either talk to me after class, or email me at skannan@ucsb.edu.

With all that out of the way, let's get started!

1.3 Natural Computing

The first thing I would like to do is provide a high level intuition for why we would expect that physics would have anything to say at all about the learning problem.

Let's start by formulating the learning problem in more formal terms. The simplest class of model is 1D linear regression. In a linear regression problem, our model is something of the form

$$y = mx + b \tag{1}$$

and given a set of data points, we want to find the value of m, b, so that our line most accurately fits the data. One way that one might do this is with gradient descent. We define some *loss function* that we wish to optimize, and then we compute the gradient of that loss with respect to our two parameters. For example, we could use the MSE loss.

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = \sum_i (y_i - (mx_i + b))^2 \tag{2}$$

And we determine our model parameters, m and b by starting them at some random values, and then computing the gradient of the loss function with respect to each of them, and following the negative gradient down.

In the case of generative AI, we formulate the problem slightly differently. Instead of trying to learn a *function* that maps an input domain to an output domain, we try to learn a *probability distribution* over some support, along with a method to sample from that probability distribution.

Now let's change tacks a little bit.

Consider a ball inside of a bowl (see Fig. 1.3). Our everyday physical knowledge tells us that the ball will roll down until it is at the bottom of the bowl. Physically, we formulate this as because the ball experiences some force, $\mathbf{F}$, which is the negative gradient of some potential energy function. In one dimension,

$$F = -\frac{dU}{dx} \tag{3}$$

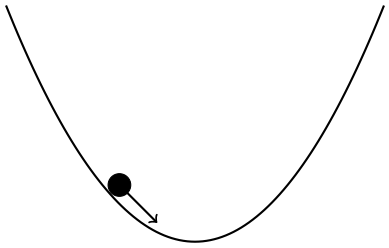

Figure 1: A ball moving in a bowl-shaped potential well.

Now in practice things are a little more complicated than the simple examples I've presented. In time, we will see that for thermodynamic systems, the quantity that is minimized is in fact not the energy, but something termed the *free energy* of the system. For now, though, the point I would like you to take away is that many natural systems, left to their own devices, follow some sort of optimization principle that looks like our gradient descent problem, with the energy function playing the role of the loss.

If we could design a physical system (or simulate one), in which states of low energy corresponded to solutions to our optimization problem, then we could quickly solve a lot of hard problems just by constructing our physical system and letting it settle. In fact, this is the operating principle of a broad class of quantum computers, called quantum annealers (see D-Wave Systems), as well as the thermodynamic computing solutions being built by companies like Extropic, Normal Computing, Ludwig Computing, etc.

1.4 Statistical Mechanics

Now that we understand that natural systems follow optimization principles, I will present a whirlwind tour of some major results of statistical mechanics, and then we will see how these can be used to formulate our first class of machine learning models, the Boltzmann machine. The goal of this section is to come away with an understanding of *free energy*, and how the dynamics of natural systems automatically lead to the minimization of free energy.

Let's start with definitions. As a brief disclaimer, some of these definitions will not be physically rigorous, in that I will be providing definitions that correspond to the things we are interested in in this class, not the most general, physically correct definitions.

1. A *system* is the object(s) under study. For example, the "system" could be a hydrogen atom, with its various orbital energy levels, or something more abstract, like the result of flipping a coin 10 times.

2. A *microstate* is specific microscopic configuration of the system, characterized by the exact positions and momenta (or states) of all constituent particles. For example, in the case of flipping a coin ten times, the outcome HHTTHHTTHT would be a specific microstate.

3. A *macrostate* is a characterization aggregate statistical properties. In the case of a gas, this could be temperature, pressure, volume, and total energy. In the case of our coin, "5 heads, 5 tails" would be one macrostate, while "10 heads, 0 tails" would be another. A macrostate encompasses many possible microstates that share the same macroscopic properties.

4. A *reservoir* is system that is so large that it can exchange particles or energy with another system without any significant change to its macroscopic properties. For example, if the system I am studying is a melting ice cube, I might say that the atmosphere is a reservoir. The ice cube will absorb some energy from the atmosphere, and, in principle, cool the atmosphere a little bit, but the atmosphere is so large and energetic that the amount of energy it loses, and thus the change in its temperature, are negligible.

5. A system is at *equilibrium* when all of its macroscopic properties, like temperature, pressure, volume, magnetization, etc. are constant in time.

The last thing we must mention is the fundamental assumption of statistical mechanics: For an isolated system at equilibrium, *all accessible microstates are equally likely.* To be clear, this is an assumption, not a derivable fact. However, this assumption can be made plausible in many ways, but most convincingly perhaps, by the fact that the physics that stems from it is an accurate descriptor of our world.

Great, now that we have our definitions, let's do some physics. We are going to use some elementary physical principles and rules of probability to derive the "Boltzmann distribution". The argument we follow is due to Schroeder.

Imagine our system is just a single hydrogen atom (with non-degenerate energy levels), and the reservoir is some bath of energy, say a star (Fig. 2).

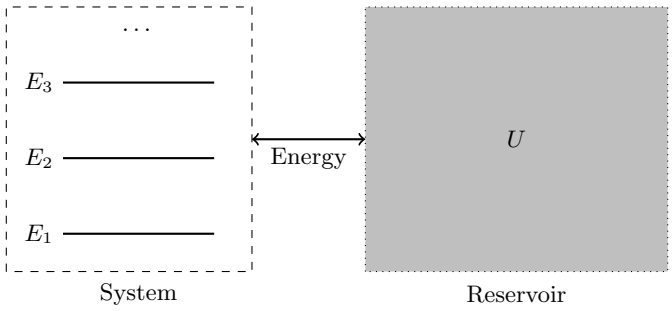

Figure 2: A system with infinitely many energy levels in contact with a thermal reservoir.

For the purposes of this example, let's pretend that the system can only exchange energy with the reservoir, not particles. If our atom were truly isolated, its energy would be fixed, and so it would remain in the same microstate forever. However, because of the reservoir, energy can be transferred into and out of the system, meaning that there is some distribution of states.

Let's start with the probability of finding the atom in a particular microstate. According to our fundamental assumption, each microstate is equally probable for an *isolated* system, but our system isn't isolated. However, the reservoir-system pair constitutes an isolated system, so let's look at its microstates. The probability of a particular microstate depends on all of the microstates, so let's start by looking at the ratio of probabilities between two states.

$$\frac{\mathcal{P}(E_S = E_i)}{\mathcal{P}(E_S = E_j)} = \frac{\Omega(E_R = U - E_i)}{\Omega(E_R = U - E_j)} \tag{4}$$

Here is a good place to introduce the next major concept of statistical mechanics, *entropy*. Entropy is defined as the logarithm of the number of microstates. It is useful to take the logarithm for a number of reasons that will become apparent, not least because the number of microstates can grow very, very large.

$$S = k \log \Omega \tag{5}$$

Rewriting our ratio in terms of the entropy,

$$\frac{\mathcal{P}(E_S = E_i)}{\mathcal{P}(E_S = E_j)} = \frac{\exp\left(S(U - E_i)/k\right)}{\exp\left(S(U - E_j)/k\right)} = \exp\left(\frac{\Delta S_R}{k}\right) \tag{6}$$

Assuming that $U \gg E_i, E_j$, we can expand around $E_i, E_j = 0$.

For the sake of time, I will not prove the following relation, but it turns out that

$$\frac{1}{T} = \frac{\partial S}{\partial U} \tag{7}$$

and so we can rewrite this identity as

$$\frac{\mathcal{P}(E_S = E_i)}{\mathcal{P}(E_S = E_j)} = \exp\left(\frac{\Delta E}{k} \left.\frac{\partial S}{\partial E}\right|_U\right) = \exp\left(\frac{\Delta E}{kT}\right) = \frac{\exp\left(\frac{E_i}{kT}\right)}{\exp\left(\frac{E_j}{kT}\right)} \tag{8}$$

We assume that $\frac{\partial S}{\partial E} = \frac{1}{T} > 0$, which is to say that temperature is strictly non-negative. There are some systems that can have negative absolute temperatures, but this is beyond the scope of our discussion. Using Eq. 8, we have that

$$\frac{\mathcal{P}(E_S = E_i)}{\exp\left(\frac{E_i}{kT}\right)} = \frac{\mathcal{P}(E_S = E_j)}{\exp\left(\frac{E_j}{kT}\right)} \tag{9}$$

Since the left depends only on E_i, and the right depends only on E_j, we have that it must be a constant, which we will call Z. Then,

$$\mathcal{P}(s) = \frac{1}{Z} e^{-\frac{E(s)}{kT}} \tag{10}$$

where the *partition function*, Z, is given by

$$Z = \sum_i \exp\left(\frac{-E(i)}{kT}\right) \tag{11}$$

In the continuous case, the sum is instead replaced with an integral over states. This is the *Boltzmann Law*, and it is the most important equation in statistical mechanics. From it, we can, in principle, calculate the probability of a thermal system being in any particular state. We see that low energy states are, in fact, more probable. As temperature increases, the distribution becomes more uniform. As temperature decreases, the probability of being in the ground state approaches 1, and the probability of the other states goes to 0. To bring things back to where we started, we now understand why thermodynamic systems approach the state of lowest energy, and so if we can construct a system, such that the energies of desirable states are low, and energies of undesirable states are high, then we can find the optimal solution with high probability, just by sampling from this distribution.

As a quick note, we can generalize the Boltzmann Law to the case where we allow both particle and energy exchange as follows:

$$\mathcal{P}(i) = \frac{1}{Z} \exp\left(\frac{-E(i) - \mu n(i)}{kT}\right) \tag{12}$$

where here we have defined the *electrochemical potential*, μ, to be $\frac{\partial S}{\partial n}$, and $n(i)$ is the number of particles in the system in microstate i.

Some caveats about the Boltzmann Law. What turns out to be the main problem in using this Boltzmann distribution for any kind of computational task is the intractability of the partition function.