CHAPTER 8

# Learning Gibbs States: High Temperature

We have so far explored algorithms for learning completely general quantum states via quantum state tomography, as well as algorithms for efficiently learning the expectation values of many observables in a quantum state via shadow tomography and its variants. The quantum state tomography algorithms circumscribe our ability to learn about general quantum states, although even for modest system sizes the algorithms are completely impractical, and in fact fundamentally so. On the other hand, the comparatively efficient algorithms for learning quantum observables capitalize on the fact that there are often specific, structured sets of observables that are of interest to us. This relationship between structure and efficiency of learnability will continue to be a persistent theme.

In this chapter, we turn again to learning quantum states, but of a more structured kind. Specifically, we consider Gibbs states, which describe quantum systems at finite temperature. Indeed, essentially all material objects we commonly interact with are (approximately) at finite temperature, and so Gibbs states are particularly natural. Almost all condensed matter experiments operate at (approximately) finite temperature, so even in laboratory settings more removed from ordinary experience, finite temperature states are salient.

We begin with a discussion of Gibbs states and their properties, and then turn to an initial strategy for learning the parameters of a Gibbs state via quantum measurements.

## 1. Some history

While we do not intend to give a detailed account of thermodynamics and its quantum counterpart, we can at least marvel at its conceptual innovation and technical power. The subject of **thermodynamics**, and its cousin **statistical mechanics**, were developed over the course of the 19th century, in large part instigated to build a theory of steam engines that were essential to the industrial revolution. In this way the subjects were highly practical: there was a great need to make engines more efficient, and to understand what aspects of an engine's design were essential or extraneous.

Instead of focusing on engines, let us examine a slightly different thread of the history. In the 17th and 18th centuries, there was progress on understanding "ideal gas laws", namely how a gas' temperature, pressure, and volume are related in simple circumstances. The interrelations were empirically observed to be strikingly simple, which is surprising since we now know that gases are a complex system of interacting particles. We should note, however, that the discoverers of the ideal gas laws did not subscribe (or at least did not fully subscribe) to the 'atomic hypothesis', and so had a rather different physical picture of gases than we now have. By the mid-19th century when the atomic hypothesis was back in vogue

and properties of gases were used industrially for designing engines, the founders of statistical mechanics articulated an interesting puzzle: if a gas is made of atoms, perhaps interacting in a complex manner, then why should gross properties (like temperature, pressure, and volume) be so simply related?

Here is where they made a surprising conceptual move. The standard practice of Newtonian physics is to write down the equations for every particle in a system and track their dynamics and interactions; this is simply too complicated to carry out for a realistically-sized gas, and especially in the 19th century. Instead of appealing to exact dynamical laws, the founders of statistical mechanics reasoned that it would instead be sensible to describe large systems in a *statistical* manner, that we will partially explicate shortly. Then, for appropriate physical observables, the microscopic details may wash out, giving accurate predictions for gross quantities. Amazingly, this works beautifully, and among myriad successes provides a first principles derivation of the ideal gas laws.

An analogy may be helpful for the uninitiated. It is well-known, by the central limit theorem, that the statistical distribution of a sum of i.i.d. random variables converges to a Gaussian, characterized by its mean and variance; all of the other microscopic details wash away in the appropriate limit. What the founders of statistical mechanics did was figure out a type of 'central limit theorem' for *Hamiltonian dynamics*. To be sure, their arguments were not rigorous (and over the intervening century-and-a-half there has been much effort to make the arguments rigorous), but are empirically correct. That is to say, even if mathematics cannot yet verify all of their arguments in complete generality, Nature has definitively demonstrated their correctness.

The pioneering work of James Clerk Maxwell, Ludwig Boltzmann, and Josiah Willard Gibbs on statistical mechanics in the mid-to-late 19th century was ultimately synthesized into quantum mechanics in the 1930's, in large part by John von Neumann and Lev Landau. An information-theoretic articulation of statistical mechanics was later emphasized by Edwin Thompson Jaynes, based on the crucial and pioneering work of Claude Shannon in the mid-to-late 1940's. The perspective of our discussion below is most indebted to von Neumann and Jaynes.

## 2. Gibbs states and their properties

Before delving into the 'derivation' of a finite-temperature quantum state, it is first worthwhile to re-examine our understanding of information-theoretic entropy, due to Shannon. For a probability distribution $\vec{p} = (p_1, ..., p_n)$, its entropy is given by

$$S[\vec{p}] = -\sum_i p_i \log_2(p_i)\,,$$

where we have temporarily decided to use the base two logarithm; we will later turn back to the natural logarithm. Let us ask: what does the entropy of a probability distribution *mean*? While we have used the classical entropy, and many of its mathematical properties, so far in our analyses, we have not until now stared into its soul.

As a warm-up, suppose your friend flips a fair coin and hides it in their hand. You would surmise that the probability of heads is $\frac{1}{2}$ and the probability of tails is likewise $\frac{1}{2}$. Let us ask: how many bits of information do you expect to learn, once the state of the coin is revealed to you? Well, in this case, with probability

$\frac{1}{2}$ if heads revealed you learn $-\log_2(\frac{1}{2}) = \log_2(2) = 1$ bit of information, and with probability 1/2 if the tail is revealed you likewise learn $-\log_2(\frac{1}{2}) = \log_2(2) = 1$ bit of information. Thus the expected number of bits you learn is 1. Indeed, the entropy of $\vec{p} = (\frac{1}{2}, \frac{1}{2})$ is $-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$. By a similar argument, you can convince yourself that if your friend flips $n$ unbiased coins, then when the outcomes are revealed to you, you in expectation learn $n$ bits of information; this is because the entropy of the uniform distribution $\vec{p} = (\frac{1}{2^n}, \frac{1}{2^n}, ..., \frac{1}{2^n})$ is $n$.

Let us try one more example. Suppose we have a three-outcome experiment, where the outcomes have probabilities $\vec{p} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. Then, upon learning the outcome, the expected number of bits we learn is $-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{4}\log_2(\frac{1}{4}) - \frac{1}{4}\log_2(\frac{1}{4}) = \frac{3}{2}$, which is the entropy of $\vec{p}$. Even though we have been considering probability distributions which interface nicely with powers of 2, more general distributions have the same interpretation.

In summary, we have the interpretation:

*If we sample $i$ from $\vec{p}$, the number of bits we expect to learn when the sample is revealed to us is $S[\vec{p}]$.*

If we use the natural logarithm for the entropy, where $S[\vec{p}] = -\sum_i p_i \log p_i = -\log(2)\sum_i p_i \log_2 p_i$, then the amount of information we learn is counted in 'nats', which is the natural log version of 'bits'. The entropies in different bases for the logarithm are evidently equal up to constants of proportionality.

Our ordinary-language interpretation of entropy is suggestively written: it tells us that (classical) entropy may be regarded as contingent on the observer and their knowledge of a system. For instance, if your friend knows how to toss a coin the same way every time, they can exactly predict that it will land on heads, and therefore when they see 'heads' they will learn 0 bits of information. On the other hand, if you are unaware that the toss is sure to be heads and instead surmise that the coin toss is fair, you will expect to learn 1 bit of information upon seeing the outcome. In this way, our formulation sneaks in a Bayesian viewpoint: your prior and your friend's prior may be different, and so entropy depends on the knowledge of the individual.

To this end, let us consider the following question. Suppose we have a classical system where each configuration $i$ has an associated energy $E_i$, and we know that when we measure the energy at different times it is on average equal to $E$. If the system is composed of many interacting particles, we cannot in practice keep track of the configuration or detailed dynamics of the system; so in the spirit of statistical mechanics, let us describe the system by a probability distribution $\vec{p}$ over configurations $i$. That is, we say that the probability of configuration $i$ is $p_i$. Then which probability distribution $\vec{p}$ is natural to choose? One desired property is that $\mathbb{E}_{i\sim\vec{p}}[E_i] = \sum_i p_i E_i = E$, since we would like the distribution to be consistent with our empirical observation that the energy of the system is $E$ on average.[1] This property is very much not sufficient to uniquely pin down $\vec{p}$, so we need some other criteria as well. To this end, consider the following: suppose we let $\vec{p}$ be the probability distribution of *maximum entropy* with average energy $E$. In other words, we adopt a principle of *maximum ignorance* in which we stipulate that, if we were to learn the state of the system, then it would be maximally surprising to us

[1] This is actually a subtle point, which we will return to later.

(i.e. we would learn the maximum possible number of bits) provided that we already knew the average energy was $E$. Before taking a step back to decide if this is a good idea, let us pursue the stated maximization. We consider the maximization of the cost function

$$\mathsf{C}[\vec{p}, \beta, \lambda] = S[\vec{p}] + \beta \left( E - \sum_i p_i \, E_i \right) + \lambda \left( 1 - \sum_i p_i \right) \tag{33}$$

where we enforce $\mathbb{E}_{i \sim \vec{p}}[E_i] = E$ via a Lagrange multiplier $\beta$ and $\sum_i p_i = 1$ via a Lagrange multiplier $\lambda$. We will also look for maximizers for which $p_i \geq 0$, which we can pin down uniquely. Solving the saddle point equations $\frac{\partial \mathsf{C}}{\partial \beta} = 0$ and $\frac{\partial \mathsf{C}}{\partial \lambda} = 0$, as well as $\frac{\partial \mathsf{C}}{\partial p_i} = 0$ for all $i$, we find

$$p_i = \frac{e^{-\beta E_i}}{Z(\beta)} \,, \tag{34}$$

where $Z(\beta) = e^{\lambda + 1} = \sum_i e^{-\beta E_i}$ is a constant such that $\sum_i p_i = 1$, and $\beta$ is chosen such that $\sum_i \frac{e^{-\beta E_i}}{Z} \, E_i = E$. Because the Shannon entropy is a strictly concave function of $\vec{p}$ on the probability simplex, and because the constraints in (33) are linear, any stationary point of the cost function is automatically a global maximum. Moreover, the strict concavity ensures that this maximum is unique (except in degenerate cases where all $E_i$ are equal). Provided that $E$ lies within the feasible range $[\min_i E_i, \max_i E_i]$ and that the partition function $Z(\beta)$ converges, the **Gibbs distribution** (or Gibbs ensemble) $p_i = \frac{e^{-\beta E_i}}{Z(\beta)}$ in (34) is therefore the unique maximum-entropy distribution consistent with the given average energy constraint. The reciprocal of the parameter $\beta$, often denoted by $T := \frac{1}{\beta}$, is called the **temperature** of the system.

Crucially, the Gibbs distribution (34) does not just predict the average energy which we put in 'by hand'; additionally it makes predictions for any other observables we can measure. That is, given some observable which takes value $O_i$ on configuration $i$, the Gibbs distribution makes the prediction that we will measure

$$\overline{O} := \sum_i \frac{e^{-\beta E_i}}{Z(\beta)} \, O_i \,.$$

Empirically, this distribution provides remarkably accurate predictions for observables $O$ that are insensitive to microscopic details; that is, when $O_i \approx O_j$ for configurations $i$ and $j$ that are similar on medium or large scales. The empirical success of the Gibbs distribution is thus far from trivial: it reflects deep underlying principles of statistical mechanics and points to the existence of universality classes governing macroscopic behavior across diverse physical systems.

So why should the Gibbs distribution work so well? Let us revisit some of the ingredients in the optimization that led us to the distribution. First, suppose we call our system of interest $\mathcal{S}$, and also consider an environment $\mathcal{E}$ that couples to our system. If the system and the environment are weakly coupled and come to equilibrium, then their energies are stable on average: the average energy of our system is $E$ and the average energy of our environment is $E_{\mathcal{E}}$, with total conserved energy $E_{\text{tot}} = E + E_{\mathcal{E}}$. Thus the energy of $\mathcal{S}$ can fluctuate about its mean $E$. Imagine, for instance, that $\mathcal{S}$ is described by particles in a box, and that the environment $\mathcal{E}$ is the world outside the box. Particles outside the box can hit the box and confer energy to it; similarly, particles within the box can hit the box and

confer energy to the environment. Importantly, if we characterize our knowledge of what is inside the box, it may initially have little entropy, e.g. if we knew very well the initial state of the system. But due to the chaotic dynamics inside the box, a small lack of knowledge about the system can balloon at later times into a *near-total* lack of knowledge, further facilitated by the interaction of our system with an environment of which we have little knowledge. Moreover, the chaos happens quickly, on time scales smaller than we can measure; thus our empirical average when we e.g. measure the average energy is secretly a time average as well. While it is not surprising, then, that we might *personally* have little knowledge of the system at later times (and thus would assign a large entropy if the state of the system were revealed to us), it is surprising that our *personal* ignorance suggests a distribution which is *predictive* of any coarse measurement.

To generalize the above arguments to quantum systems, we need to define a suitable notion of entropy. One criterion is if we picked a diagonal density matrix

$$\rho = \mathrm{diag}(p_1, p_2, ..., p_d)\,,$$

we would like for the quantum entropy to satisfy

$$S[\rho] = -\sum_i p_i \log(p_i)\,.$$

For a general state $\rho$, von Neumann stipulated the following quantum generalization of the classical entropy, which in his honor, we call the **von Neumann entropy**, or *quantum entropy* (or even merely the *entropy* when the context is clear):

$$S[\rho] := -\mathrm{tr}(\rho \log \rho)\,.$$

Here $\log \rho$ is the matrix logarithm, meaning that if $\rho = \sum_i \lambda_i |\psi_i\rangle\langle\psi_i|$ is the spectral decomposition of $\rho$, then $\log \rho = \sum_i \log(\lambda_i)|\psi_i\rangle\langle\psi_i|$. Accordingly,

$$S[\rho] = -\sum_i \lambda_i \log(\lambda_i)\,,$$

where we are taking $0 \log 0 := 0$. (Or if you want to be fussy, if $\rho$ has zero eigenvalues let $\rho_\varepsilon := (1-\varepsilon)\,\rho + \varepsilon\,\frac{\mathbb{1}}{d}$ and take $\lim_{\varepsilon \to 0^+} S[\rho_\varepsilon]$, which will land you on the "$0 \log 0 := 0$" mnemonic.) The von Neumann entropy enjoys many nice mathematical properties, which we will ourselves enjoy later on.

With the von Neumann entropy in hand, the quantum analogue of the classical maximum-entropy construction is immediate. Let $H$ be the Hamiltonian of a finite-dimensional quantum system, and let $\rho$ be a density matrix. Among all density matrices $\rho$ obeying $\mathrm{tr}(\rho H) = E$, we seek the state of maximal entropy $S[\rho] = -\mathrm{tr}(\rho \log \rho)$.

Let us introduce Lagrange multipliers $\beta, \lambda \in \mathbb{R}$ and consider

$$\mathsf{C}[\rho, \beta, \lambda] := -\mathrm{tr}(\rho \log \rho) + \beta(E - \mathrm{tr}(\rho H)) + \lambda(1 - \mathrm{tr}(\rho))\,,$$

where the second Lagrange multiplier enforces that $\rho$ has unit trace; we will see that Hermiticity and positive semi-definiteness will be readily enforced. A first derivative

$$\frac{\delta \mathsf{C}}{\delta \rho} = -\log \rho - \mathbb{1} - \beta H - \lambda\,\mathbb{1}\,.$$

At an interior maximizer (which is full-rank whenever $E \in (E_{\min}, E_{\max})$), $\frac{\delta \mathsf{C}}{\delta \rho} = 0$, and so $\log \rho = -(\lambda+1)\,\mathbb{1} - \beta H$ which implies

$$\rho = e^{-(\lambda+1)}\, e^{-\beta H}.$$

Similar to the classical case, imposing the normalization condition $\mathrm{tr}(\rho) = 1$ identifies the partition function

$$Z(\beta) := \mathrm{tr}\big(e^{-\beta H}\big) = e^{\lambda+1},$$

and thus

$$\rho_\beta := \frac{e^{-\beta H}}{Z(\beta)}\,, \tag{35}$$

which is called a **quantum Gibbs state**. The multiplier $\beta$ is then set by $\mathrm{tr}(\rho_\beta H) = E$, equivalently

$$E(\beta) = \mathrm{tr}(\rho_\beta H) = -\frac{\partial}{\partial \beta} \log Z(\beta).$$

Because $S[\rho]$ is strictly concave on the convex set of density operators and the constraints are linear, any stationary point is the global maximizer; strict concavity further implies uniqueness of $\rho_\beta$ (the maximizer becomes rank-deficient only in the endpoint limits $\beta \to \pm\infty$). As before $\beta := \frac{1}{T}$ is the inverse temperature.

To see the connection with the classical Gibbs distribution, we can diagonalize the Hamiltonian as $H = \sum_a E_a\, \Pi_a$ with projectors $\Pi_a$ onto energy-$E_a$ subspaces (of dimensions $g_a := \mathrm{tr}\, \Pi_a$). Then

$$\rho_\beta = \frac{e^{-\beta H}}{Z(\beta)} = \sum_a \frac{e^{-\beta E_a}}{Z(\beta)}\, \Pi_a, \qquad Z(\beta) = \sum_a g_a\, e^{-\beta E_a}.$$

Thus $\rho_\beta$ is block-diagonal in the energy basis and proportional to the identity on each degenerate eigenspace; in any orthonormal energy eigenbasis $\{|a,\mu\rangle\}_{\mu=1}^{g_a}$ one has diagonal entries $\langle a,\mu|\rho_\beta|a,\mu\rangle = e^{-\beta E_a}/Z(\beta)$, reproducing the classical Gibbs weights for energy eigenstates.

As in the classical case, predictions for observables follow by averaging with $\rho_\beta$:

$$\langle O \rangle_\beta := \mathrm{tr}(\rho_\beta O).$$

The quantum Gibbs state provides accurate results for a physical system when $O$ is a coarse-grained, gross observable, such as those corresponding to energy and pressure. But unlike the classical case, the quantum Gibbs state also provides accurate predictions when $O$ is a local observable. The reason is subtle: in the classical setting, a system is in a definite configuration, not a probabilistic average; as such, measurements of the system only portray a probabilistic average when our measurements are coarse enough in space and time so as to blur the fast, chaotic dynamics of the system. In the quantum setting, if our system becomes entangled with its environment, then the density matrix describing our system can *genuinely* be a probabilistic mixture, i.e. a mixed state density matrix. Thus, even local observables, which would reveal the specific microscopic configuration in a classical system, are accurately described by thermal averages in the quantum case, as the system's reduced density matrix has genuinely lost information about quantum superpositions through its coupling to the environment.

All of the above said, we have ample motivation to design quantum learning algorithms for learning quantum Gibbs states. Doing so will require us to leverage