1 Gibbs measures: what and why?

We will be studying *Gibbs measures*, which are probability measures with the form $\mu(x) \propto e^{f(x)}$ for some function f on the outcome space Ω (which can be an arbitrary finite set in this talk—but continuous settings are possible too). Astute readers will notice that every probability measure which assigns nonzero probability to each outcome is of this form (and all probability measures are of this form if we allow f to take the value $-\infty$). So the first question is: why is this a useful way to write a probability measure?

1.1 The Gibbs variational principle

Notice that samples from μ are more likely to have a higher value of f. Often, f will represent some objective function, which should be large on good values of x, so this is a good reason to use a measure proportional to $e^{f(x)}$. But plenty of other distributions would also have this property, so why this particular form? Well, the Gibbs measure is in some sense the most "generic" probability measure satisfying this property.

What do we mean by this? Consider the following functional of a probability distribution ν:

$$\mathcal{F}_f(\nu) = \mathbb{E}_\nu[f] + \mathrm{Ent}(\nu);$$

the first term is the expecation of $f(X)$ when $X \sim \nu$, and the second is the entropy of ν,

$$\mathrm{Ent}(\nu) = -\sum_x \nu(x) \log \nu(x).$$

The functional $\mathcal{F}_f$ is (a slight modification of) the "free energy," but this name is unimportant for us; we will just aim to maximize the functional. For ν to make $\mathcal{F}_f$ large, it must make f high on average, while also having a high entropy. Making $\mathbb{E}_\nu[f]$ large is self-explanatory, but the high-entropy term is a "genericity" condition. Think of this as akin to how the uniform distribution maximizes the entropy alone. We are searching for the analog of the uniform distribution which takes an objective function f into account.

It turns out that our Gibbs measure $\mu \propto e^f$ maximizes this functional. Indeed, for any ν,

$$\begin{aligned}
\mathbb{E}_\nu[f] + \mathrm{Ent}(\nu) &= \sum_x \nu(x)(f(x) - \log \nu(x)) \\
&= \sum_x \nu(x) \log \frac{e^{f(x)}}{\nu(x)} \\
&\le \log \sum_x e^{f(x)} \qquad (*)
\end{aligned}$$

using the concavity of log at step $(*)$. The last expression is $\mathcal{F}_f(\mu)$, as we will see shortly. First, let's give a name to the sum inside the log in the last step. We'll call it the partition function and denote it by Z_f. Notice that it's the normalization constant for μ, so that $\mu(x) = \frac{e^{f(x)}}{Z_f}$. So we have

$$\begin{aligned}
\mathbb{E}_\mu[f] + \mathrm{Ent}(\mu) &= \sum_x \mu(x) \left(f(x) - \log \frac{e^{f(x)}}{Z_f} \right) \\
&= \sum_x \mu(x) \log Z_f \\
&= \log Z_f.
\end{aligned}$$

Both of the previous two displays imply that

$$\mathcal{F}_f(\nu) \le \mathcal{F}_f(\mu)$$

for any ν, and the only inequality is Jensen's inequality $(*)$, which is an equality exactly when $\frac{e^{f(x)}}{\nu(x)}$ is constant. So indeed, μ maximizes $\mathcal{F}_f$,and the maximal value is $\log Z_f$. This is the *Gibbs variational principle.*

1.2 Connection to statistical physics

You might object that the functional $\mathcal{F}_f$ is somewhat arbitrary. Why not weight the two terms unequally? Perhaps we want to put a greater emphasis on achieving a high value of f, or conversely perhaps we want to emphasize the entropy term. Since the first term, $\mathbb{E}_\nu[f]$, is linear in f, reweighting the terms of the free energy could be achieved simply by multiplying f by some number; let's do this but also switch notation to match the literature.

More precisely, let's consider $f(x)$ as of the form $f(x) = -\beta H(x)$, where $H(x)$, called the *Hamiltonian* represents the *energy* of the outcome x, and the minus sign is there because we typically want to minimize energy, not maximize it. The functional becomes

$$\mathcal{F}_f(\nu) = -\beta \mathbb{E}_\nu[H] + \mathrm{Ent}(\nu),$$

so in order to maximize $\mathcal{F}_f$ we want to minimize the average energy and maximize the entropy, and the parameter β controls how much we care about each of these objectives. Since we want $f = \beta H$ to be unitless, β must have units of inverse energy. However, since energy and temperature are related by a constant multiple (Boltzmann's constant), β is usually called the "inverse temperature" instead. So for low β (high temperature), the entropy term is more important and the measure is closer to the uniform distribution, and for high β (low temperature), the energy term is more important and the measure is closer to being concentrated on the low-energy states.

1.3 Example: Ising model

Let $G = (V, E)$ be a connected graph. Our state space will be $\Omega = \{\pm 1\}^V$, so each vertex gets a "spin" which is either up or down. We'll take

$$f(x) = \beta \sum_{u \sim v} x_u x_v,$$

so that each pair of neighboring spins would like to be aligned, and β controls this desire. For low β, the entropy takes over, and the spins are approximately independent coin flips. For high β, the spins are mostly aligned to all being $+1$ or all being -1 so the correlation is very high and the spins are very far from being independent.

If we get a bit more specific, this gives one of the simplest examples of a phase transition. Let's take G to be the d-dimensional torus with side-length n, so that there are n^d vertices. It turns out that there is some critical value of β, called β_c, depending on d, which separates two phases of behavior. For $\beta < \beta_c$, as n increases, the spins at two distant vertices in the graph get less and less correlated, while for $\beta > \beta_c$, there is some minimum level of correlation between the spins at any two vertices of the graph, no matter how far.

For β near β_c, the structure of the Gibbs measure can be very difficult to describe beyond some basic bounds on the correlations as just mentioned. However, for very small β the structure is very well-approximated by independent coin flips at each vertex of the graph, so the Gibbs measure is actually pretty simple. Additionally, for very large β, while the spins are highly correlated, the Gibbs measure actually approximately splits into two measures, one concentrated on mostly $+1$ configurations, and the other concentrated on mostly -1 configurations. These are also called the $+1$ and -1 "phases" of the model respectively, so really the phase transition described above is 2nd order: it is a phase transition between the model having only one phase and having two phases. The measures corresponding to the two phases in the very large β regime are again actually quite simple; they are well-approximated by a sea of their main spin with small perturbations of the opposite spin that are relatively easy to describe.

In general, Gibbs measures which feature a phase transition get very difficult to describe near the critical point. But there are certain critical structures which appear frequently such as SLE and GFF. This is a huge topic, so perhaps someone else can give a more in-depth talk (or a few) about this.

Also, if the interactions between the spins are themselves random, we obtain a *spin glass*. There are myriad questions about spin glasses: do they have phase transitions? How many? What kind of behavior do they exhibit in their various phases? This is a huge area of modern research, so it would be great to have a few talks on this topic as well.

2 What to do with Gibbs measures?

Our broad and vague themes this semester are *approximation*, *inference*, and *sampling*. Let's discuss how each of these things work with Gibbs measures.

2.1 Approximating Gibbs measures in general

In general, a Gibbs measure $\mu \propto e^f$ may be quite complicated, even if the function f is computable. The difficulty comes with computing the partition function Z_f, which is needed to do any probability with μ. So we often want to approximate μ by a simpler measure ν which is easier to compute.

In general, the quality of an approximation ν to μ can be measured using the KL-divergence of ν with respect to μ, which is defined as

$$\mathrm{KL}(\nu \| \mu) = \mathbb{E}_\nu \log \frac{\nu}{\mu} = \sum_x \nu(x) \log \frac{\nu(x)}{\mu(x)}.$$

This is frequently used as a measure of distance between ν and μ, although it is not a metric in the standard sense of the word. Since we are looking for a "nice" measure ν which approximates μ, we might try to minimize $\mathrm{KL}(\nu||\mu)$ over some set of "nice" measures.

If $\mu \propto e^f$ is a Gibbs measure with partition function Z_f, then we have

$$\begin{aligned} \mathrm{KL}(\nu \| \mu) &= \sum_x \nu(x) \log \frac{\nu(x) Z_f}{e^{f(x)}} \\ &= \sum_x \nu(x) \log \nu(x) - \sum_x \nu(x) f(x) + \log Z_f \\ &= -\mathcal{F}_f(\nu) + \log Z_f. \qquad (**) \end{aligned}$$

So, since Z_f is a constant, minimizing the KL-divergence is equivalent to maximizing the Gibbs functional $\mathcal{F}_f$. Recall that μ itself is the maximizer of $\mathcal{F}_f$, over all possible probability measures on the outcome space. the point is that we can get a nice approximation by maximizing $\mathcal{F}_f$ over some nice subcollection of probability measures. If this collection is rich enough, then the approximating measure ν should retain some relevant features of μ.

2.2 Mean-field approximation

One particular "nice" property to have is independence. Suppose our state space Ω is of the form $\Omega_1 \times \cdots \times \Omega_n$. This is the case, for example, in the Ising model, where we took $\Omega = \{\pm 1\}^V$. The Gibbs measure $\mu = e^f$ is a product measure (i.e. a sample has independent coordinates) exactly when f splits as

$$f(x) = f_1(x_1) + \cdots + f_n(x_n).$$

Of course, this is not the case for functions f we care about, such as in the Ising model. Nevertheless, we may wish to approximate $\mu = e^f$ by a product measure. The *mean-field approximation* of μ is the product measure which maximizes $\mathcal{F}_f$, among all product measures on Ω. This is a bit different than the "Naïve" mean-field approximation you may have seen before, where the interactions with neighbors are replaced by an interaction with the "mean field," meaning the average of all vertices in the graph, essentially giving an interaction structure of the complete graph.

There are strong connections between this definition and the Naïve mean-field approximation, though. For instance, for the Ising model on the d-dimensional torus that we discussed before, the Naïve mean-field approximation gives the Curie-Weiss model, which is essentially an Ising model on the complete graph with n^d vertices, with a different temperature $\beta' = \frac{d\beta}{n^d}$. Of course, this model does not have independent spins.

However, it does split perfectly into two measures, which each have independent spins, corresponding to the $+1$ and -1 phases of the model. In the $+1$ phase measure, the spins are independent with some mean $+m$, and in the -1 phase measure, the spins are independent with mean $-m$. These two measures are the optimizers of $\mathcal{F}_f$ among all product measures. It turns out that m satisfies $m = \tanh(2d\beta m)$, but again this is just the beginning of the story; this topic would make for a good standalone talk.

2.3 Variational inference

In general, inference is the task of determining hidden parameters from observed samples. To make things more concrete, suppose we have a family of distributions P_θ, for θ in the parameter space Θ. For some unknown θ, we observe a sample $X \sim P_\theta$. The goal is to estimate the unknown θ.

In the Bayesian framework, we start with a simple prior distribution on Θ, let's call it μ_0. Then, upon observing the samples X, we can get a posterior distribution μ_1 on Θ using Bayes's rule:

$$\mu_1(\theta) = \frac{\mu_0(\theta) P_\theta(X)}{\mathbb{E}_{\theta \sim \mu_0}[P_\theta(X)]}.$$

We assume that $\mu_0(\theta)$ is easily computable for each θ, and the same for $P_\theta(X)$ since we have a fixed sample X. However, computing the expectation in the denominator, which is the overall marginal $P(X)$ of X, may not be too easy since the integral computation may be tough so computing $\mu_1(\theta)$ in general is not so easy.

So we would like to find an approximation from a class of "nice" distributions, by minimizing $\mathrm{KL}(\nu \| \mu_1)$ for ν in this nice class. But computing $\mathrm{KL}(\nu \| \mu_1)$ for any ν actually does require us to know μ_1, i.e. to know the marginal $P(X)$, which is precisely what we couldn't compute. However, since μ_1 is a Gibbs measure

$$\mu_1(\theta) \propto e^{f_X(\theta)} \qquad \text{with} \qquad f_X(\theta) = \log(\mu_0(\theta) P_\theta(X)),$$

the same algebra goes through as in $(**)$ and we obtain

$$\mathrm{KL}(\nu \| \mu_1) = -\mathcal{F}_{f_X}(\nu) + \log Z_{f_X}.$$

Here Z_{f_X} is the normalization constant, which is the same as $P(X)$. So minimizing the KL-divergence is the same as maximizing $\mathcal{F}_{f_X}$, but in computing the value of $\mathcal{F}_{f_X}(\nu)$ for any particular ν, we *do not* have to compute $P(X)$: recall that the formula

$$\mathcal{F}_{f_X}(\nu) = \mathbb{E}_\nu[f_X] + \mathrm{Ent}(\nu)$$

only depends on the function f_X, which we can compute, and the measure ν itself, and we are assuming that we can compute the above quantities since ν is nice enough, whatever that means.

In this area of statistics, the quantity $\log P(X) = \log Z_{f_X}$ is called the "evidence," because if it is higher, then X is more typical and so should give a better idea of what the hidden θ is. Additionally, since $\mathrm{KL}(\nu \| \mu_1) \geq 0$, the equation $(**)$ tells us that $\mathcal{F}_{f_X}(\nu)$ is a lower bound on the evidence. For this reason, $\mathcal{F}_{f_X}$ is called the Evidence Lower BOund, or ELBO for short. There's a lot more to say here about which types of "nice" distributions allow one to maximize the ELBO in practice while still being a useful surrogate for a posterior distribution, and also how these calculations go in the real world. As usual, this would be a good topic for another talk.

2.4 Sampling via Markov chains

Instead of finding a surrogate distribution as we have examined in the previous two sections, another way to approximately sample from μ is to find a Markov chain $\{x_t\}_{t=0}^\infty$ which has μ as its stationary distribution. It turns out this can be done without knowing the normalization constant Z_f. For instance, we can use the Metropolis algorithm, which works, at a high level, as follows:

1. Starting at $x_t \in \Omega$, *propose* the next state y from a simple distribution, which typically depends on x_t.
2. Compute the *ratio* of probabilities $\alpha = \frac{\mu(y)}{\mu(x_t)} = \exp(f(y) - f(x_t))$, which does not depend on Z_f.
3. Sample $U \sim \mathrm{Unif}[0,1]$; if $U < \alpha$ we *accept* and set $x_{t+1} = y$. otherwise we *reject* and set $x_{t+1} = x_t$.

In the Ising model, for example, in step 1 we can take y to be a configuration where some uniformly random spin in x_t is rerandomized. Under reasonable assumptions about the simple distribution in this step, this Markov chain does indeed have stationary distribution μ (and in fact is often reversible with respect to μ).

Now the question becomes: how long should we run this Markov chain to get a reasonable sample? This is another huge topic that could easily be the source of multiple talks this semester. There are many different methods analogous to the Metropolis algorithm (e.g. Langevin dynamics), as well as tons of ways to analyze each one (e.g. log-Sobolev inequalities), and various things that can be proved (e.g. cutoff) in order to better understand these chains.