Statistics 291: Lecture 2 (January 25, 2024)

Free Energies and Moment Computations

Instructor: Mark Sellke

Scribe: Jarell Cheong Tze Wen

1 Free Energies: A Quick Introduction

We begin this class by introducing free energies. We introduce some general definitions before specializing to our setting. In general, we would be endowed with the data of a probability space (Ω, μ), a bounded and measurable function $H \colon \Omega \to \mathbb{R}$, the "hamiltonian" or energy function, and finally an inverse temperature β. With these objects, we can make the following three definitions.

Definition 1.1 (Partition function)**.** The *partition function* Z is the function defined by the integral

$$Z(\beta) := \int e^{\beta H(x)}\, d\mu(x).$$

Definition 1.2 (Free energy)**.** The *free energy* F is the function defined by $F(\beta) := \frac{1}{\beta} \log Z(\beta)$.

Definition 1.3 (Gibbs measure)**.** The *Gibbs measure* is defined as the following Radon-Nikodym derivative:

$$d\mu_\beta(x) := \frac{e^{\beta H(x)}\, d\mu(x)}{Z(\beta)}.$$

We can reason about why F is called a free energy through the physical interpretation where Ω specifies the default states of nature, H specifies the energy of a state of nature, and the Gibbs measure specifies the distribution that nature finds itself in at inverse temperature β. Recalling, from basic chemistry, that Gibbs free energy is the amount of useful work a system can perform, we can in fact see that F has a similar, but slightly different, interpretation. Indeed, F is the "Helmholtz" free energy, defined by the identity

$$F(\beta) = U(\beta) - TS(\beta) \tag{1}$$

from thermodynamics, where $U(\beta) = \int H(x)\, d\mu_\beta(x)$ is the average energy of the system, $T = 1/\beta$ is the temperature, and $S(\beta)$ is the entropy of the system. By entropy, we mean the Kullback-Leibler divergence:

Proposition 1.4. *If $S(\beta)$ is set to be the Kullback-Leibler divergence $D_{\mathrm{KL}}(\mu_\beta \,\|\, \mu)$, then (1) holds.*

Proof of Proposition 1.4. The desired result follows form plugging in the definition of Kullback-Leibler divergence and performing some algebraic manipulation as follows:

$$D_{\mathrm{KL}}(\mu_\beta \,\|\, \mu) := \int \log\left(\frac{d\mu(x)}{d\mu_\beta(x)}\right) d\mu_\beta(x) = \log Z(\beta) - \int \beta H(x)\, d\mu_\beta(x) = \beta(F(\beta) - U(\beta)). \qquad \square$$

In thermodynamics or physical chemistry, one usually starts with some axiomatic definition of a physical system that does not include the state space definition. Instead, one considers rectangular prism blocks of gas interacting with neighboring blocks, and from there, the definitions found above can naturally arise.

The main idea is that one has differential expressions for free energy of the form

$$\begin{aligned} dU &= T\,dS - P\,dV, \\ dF &= -S\,dT - P\,dV, \end{aligned}$$

where P represents pressure and V represents volume. Using the product rule changes which variable goes where, and this corresponds to changing the relevant state variables one is tracking. From this, one could view the Helmholtz free energy naturally as useful work at a fixed temperature T and volume V. This is not too relevant for our class, but it's good to know that the things we're thinking about have real meaning.

Meanwhile, in statistics, we have the following interpretation instead: μ can be viewed as some prior distribution, maybe for a signal we're trying to estimate, βH can be viewed as a log-likelihood, and μ_β can be viewed as the resulting posterior. Here, β can usually be realized as some signal-to-noise ratio. Hence, understanding the Gibbs measure in statistics amounts to understanding a posterior instead of a physical system, but both of these settings are very natural ones to be in.

In this class, we're going to be interested in a very particular type of state space. We're going to set

$$\Omega = S_N := \{x \in \mathbb{R}^N : \|x\|_2 = \sqrt{N}\}.$$

Then, we usually consider the uniform measure on the sphere, $\mu = \mathrm{Unif}(S_N)$, which is the unique rotationally invariant probability measure on the sphere. The hamiltonian is a random function $H_N\colon S_N \to \mathbb{R}$.

Remark. Expectations over the sphere S_N will usually be written as integrals, with $\mathbb{E} = \mathbb{E}^{H_N}$ reserved for the expectation over the randomness of H_N.

Then, we have the same definitions as before, specializing to our hamiltonian H_N:

$$\begin{aligned} Z_N(\beta) &:= \int e^{\beta H_N(x)}\,d\mu(x), \\ F_N(\beta) &:= \frac{1}{N}\log Z_N(\beta), \\ d\mu_\beta(x) &:= \frac{e^{\beta H_N(x)}\,d\mu(x)}{Z_N(\beta)}. \end{aligned}$$

Observe that in the free energy, there are two differences. First, there is a $1/N$ factor, which makes $F_N(\beta) \le \mathcal{O}(1)$. Second, the $1/\beta$ factor is gone, which eases computation. As a result, the free energy and the original energy no longer have the same units, which makes things a little "unphysical," but this saves us the trouble of using the quotient rule whenever we want to differentiate the free energy to obtain new information.

In a lot of probability and statistical physics settings, one usually takes for granted that the routine thing to do is, given some state space and hamiltonian, to compute the free energy. We will not take this point of view too strongly, and instead we prefer to use the free energy as a tool to obtain the kind of geometric and algorithmic consequences that we are often after.

Example. Consider the "shattering" phenomenon from the previous lecture. This is the scenario where there are little clumps in the Gibbs measure, all very small in terms of diameter and probability mass. These clumps are mostly separated from each other, but in total they cover most of the Gibbs measure.

Upon a sufficiently strong understanding of free energies, we may consider sampling $X \sim \mu_\beta$ from the Gibbs measure, and then we can understand not just the free energy on the sphere, but free energies on all sorts of subsets of the sphere. For instance, we might be able to say that the free energy of some cap on the sphere is larger than the free energy of the band around the cap. If this is the case, we get a "bottleneck" which allows us to deduce that the cap is one of the aforementioned clumps.

2 The p-Spin Hamiltonian

Definition 2.1 (p-spin hamiltonian)**.** The *p-spin hamiltonian* is the hamiltonian $H_{N,p} \colon S_N \to \mathbb{R}$ given by

$$H_{N,p}(x) := \frac{1}{N^{(p-1)/2}} \sum_{i_1,\dots,i_p=1}^{N} g_{i_1\cdots i_p} x_{i_1} \cdots x_{i_p}, \quad x \in S_N.$$

Here, the $g_{i_1\cdots i_p}$ are i.i.d. $\mathcal{N}(0,1)$. We can also write $H_{N,p}(x) = \langle G_N^{(p)}, x^{\otimes p}\rangle$, where $G_N^{(p)}, x^{\otimes p} \in \mathbb{R}^{N^p}$.

Proposition 2.2. *If $x, \tilde{x} \in S_N$, then $H_{N,p}(x), H_{N,p}(\tilde{x})$ are $\mathcal{N}(0,N)$ (with variance N), jointly gaussian, and*

$$\mathbb{E} H_{N,p}(x) H_{N,p}(\tilde{x}) = N \left(\frac{\langle x, \tilde{x}\rangle}{N} \right)^p.$$

Proof of Proposition 2.2. That $H_{N,p}(x), H_{N,p}(\tilde{x})$ are centered and jointly gaussian is clear because linearity preserves centered gaussians and the $g_{i_1\cdots i_p}$ are all centered gaussian. Next, by linearity and independence,

$$\mathrm{Var}[H_{N,p}(x)] = \frac{1}{N^{p-1}} \sum_{i_1,\dots,i_p=1}^{N} x_{i_1}^2 \cdots x_{i_p}^2 = \frac{1}{N^{p-1}} \left(\sum_{i=1}^{N} x_i^2 \right)^p = \frac{1}{N^{p-1}} N^p = N.$$

Since "different $g_{i_1\cdots i_p}$'s do not interact with each other," by an analogous calculation, we deduce that

$$\mathbb{E} H_{N,p}(x) H_{N,p}(\tilde{x}) = \frac{1}{N^{p-1}} \sum_{i_1,\dots,i_p=1}^{N} x_{i_1}\tilde{x}_{i_1} \cdots x_{i_p}\tilde{x}_{i_p} = \frac{1}{N^{p-1}} \left(\sum_{i=1}^{N} x_i \tilde{x}_i \right)^p = N \left(\frac{\langle x, \tilde{x}\rangle}{N} \right)^p. \qquad \square$$

Remark. We can define a symmetrized version of $G_N^{(p)}$, which we call $G_N^{(p),\mathrm{sym}}$, by setting

$$G_{N,i_1\cdots i_p}^{(p),\mathrm{sym}} = \frac{1}{p!} \sum_{\pi \in \mathrm{Sym}(p)} g_{\pi(i_1)\cdots\pi(i_p)}.$$

Then, it is the case that $\langle G_N^{(p)}, x^{\otimes p}\rangle = \langle G_N^{(p),\mathrm{sym}}, x^{\otimes p}\rangle$, and when $p = 2$, $G_N^{(p),\mathrm{sym}}$ is a GOE matrix.

Remark. In fact, $H_{N,p}$ is a centered gaussian process on S_N.

Remark. $H_{N,p}$ has a rotationally invariant distribution, i.e. if $A_N^\top A_N = A_N A_N^\top = \mathbb{I}_N$, then $\tilde{H}_{N,p}$, defined by

$$\tilde{H}_{N,p}(x) := H_{N,p}(A_N x),$$

has the same distribution as $H_{N,p}$ as functions $S_N \to \mathbb{R}$. After all, covariance is rotationally invariant, plus centered gaussian processes are determined by their covariance.

Remark. The scaling factor in $H_{N,p}$ is good because morally, S_N has $\exp(\Theta(N))$ amount of space, and also this scaling yields the operator norm of a random tensor.

Proposition 2.3. *For all p, there exists $C = C(p) > 0$ so that with probability $1 - e^{-N}$ (extremely high probability), we obtain the inequality*

$$\max_{x \in S_N} |H_{N,p}(x)| \le CN.$$

Proof of Proposition 2.3. We instead bound the larger quantity

$$\bar{\mathcal{M}} := N^{-(p-1)/2} \max_{x^{(1)},\dots,x^{(p)} \in S_N} \langle G_N^{(p)}, x^{(1)} \otimes \cdots \otimes x^{(p)} \rangle. \tag{2}$$

Let $\epsilon = 1/10p$, and let $\mathcal{N}_\epsilon \subseteq S_N$ satisfy:

- $\epsilon\sqrt{N}$-net: if $x \in S_N$, then there exists $\tilde{x} \in \mathcal{N}_\epsilon$ such that $\|x - \tilde{x}\|_2 \le \epsilon\sqrt{N}$.

- $|\mathcal{N}_\epsilon| \le (10/\epsilon)^N$.

It is well-known that such a $\mathcal{N}_\epsilon$ exists. For a construction, just pack radius $\epsilon\sqrt{N}/3$ balls into the sphere. The cardinality bound follows from considering the volume of the sphere. Then, define

$$\bar{\mathcal{M}}_\epsilon := N^{-(p-1)/2} \max_{\tilde{x}^{(1)},\dots,\tilde{x}^{(p)} \in \mathcal{N}_\epsilon} \langle G_N^{(p)}, \tilde{x}^{(1)} \otimes \cdots \otimes \tilde{x}^{(p)} \rangle.$$

Lemma 2.4. *With probability* $1 - e^{-N}$*, we have* $\bar{\mathcal{M}}_\epsilon \le CN/2$.

Proof of Lemma 2.4. We simply employ a union bound. This is more general than Proposition 2.2, but it is also true, by a very similar computation, that each

$$\langle G_N^{(p)}, \tilde{x}^{(1)} \otimes \cdots \otimes \tilde{x}^{(p)} \rangle \in \mathcal{N}(0, N).$$

Therefore, for sufficiently large C,

$$\mathbb{P}[\bar{\mathcal{M}}_\epsilon \ge CN/2] \le \left(\frac{10}{\epsilon}\right)^{pN} e^{-C^2 N/10} \le e^{-N}. \qquad \square$$

Lemma 2.5. *With probability 1 (i.e. almost surely), we have* $\bar{\mathcal{M}} \le 2\bar{\mathcal{M}}_\epsilon$.

Proof of Lemma 2.5. We employ the triangle inequality. Fix $x^{(1)},\dots,x^{(p)} \in S_N$ attaining $\bar{\mathcal{M}}$, and round

$$x^{(j)} \to \tilde{x}^{(j)} \in \mathcal{N}_\epsilon,$$

i.e. pick the $\tilde{x}^{(j)} \in \mathcal{N}_\epsilon$ such that $\|x^{(j)} - \tilde{x}^{(j)}\|_2 \le \epsilon\sqrt{N}$ that we know to exist by the construction of $\mathcal{N}_\epsilon$. Then,

$$|\langle G_N^{(p)}, x^{(1)} \otimes \cdots \otimes x^{(p)} - \tilde{x}^{(1)} \otimes \cdots \otimes \tilde{x}^{(p)} \rangle| \le \sum_{j=1}^{p} |\langle G_N^{(p)}, \tilde{x}^{(1)} \otimes \cdots \otimes \tilde{x}^{(j-1)} \otimes [x^{(j)} - \tilde{x}^{(j)}] \otimes \cdots \otimes x^{(p)} \rangle|,$$

and the latter sum is bounded above by $p\epsilon\bar{\mathcal{M}} \le \bar{\mathcal{M}}/2$ (with this final inequality by our choice of ϵ). □

Now, the desired result follows immediately from Lemma 2.4, Lemma 2.5, and some manipulation. □

Remark. The proof above yields $C \le \mathcal{O}(\sqrt{p \log p})$. This is true for $\bar{\mathcal{M}}$, but for $\max_{x \in S_N} |H_{N,p}(x)|$, the sharp constant is in fact $\sqrt{\log p}$.

Exercise. A proof of Proposition 2.3 can also be obtained through the technique of chaining.

2.1 Addition to Lecture (Used in HW1, Will be Explained in Lecture 5)

We showed above that

$$\bar{\mathcal{M}} \equiv N^{-(p-1)/2} \max_{x^{(1)},\dots,x^{(p)} \in S_N} \langle G_N^{(p)}, x^{(1)} \otimes \cdots \otimes x^{(p)} \rangle \le CN \tag{3}$$

with probability $1 - e^{-N}$. It is not hard to check that

$$\nabla H_N(x) = N^{-(p-1)/2} \nabla_x \langle G_N^{(p)}, x^{\otimes p} \rangle = pN^{-(p-1)/2} \langle G_N^{(p)}, x^{\otimes(p-1)} \rangle$$

and so for $y \in S_N$ with norm $\|y\| = \sqrt{N}$,

$$\langle \nabla H_N(x), y \rangle \le pN^{-(p-1)/2} \langle G_N^{(p)}, x^{\otimes(p-1)} \otimes y \rangle \le p\bar{\mathcal{M}}.$$

This shows that with probability $1 - e^{-N}$, we have

$$\sup_{x \in S_N} \|\nabla H_N(x)\| \le Cp\sqrt{N}$$

for C as in (3).

3 Free Energy Moment Computations

Theorem 3.1. *The inequality* $\lim_{N\to\infty}\mathbb{E}F_N(\beta)\le\beta^2/2$ *holds. Moreover, for small* β *(at most some* β_0*), this is an equality, and for large* β *(at least some* β_1*), this is a strict inequality.*

Providing a full proof of Theorem 3.1 will be the subject of the next class, but we can begin to think about the case of large β, and then work our way towards the case of small β as well. Observe that

$$\mathbb{E}\int e^{\beta H_{N,p}(x)}\,d\mu(x)=\mathbb{E}e^{\beta\sqrt{N}g}=e^{\beta^2N/2},$$

where $g\sim\mathcal{N}(0,1)$. The first equality follows by linearity of expectation since for every $x\in S_N$, $H_{N,p}(x)$ has the same distribution as $\sqrt{N}g$. Therefore, by applying Jensen's inequality on the logarithm, we get the bound

$$\begin{aligned}\mathbb{E}F_N(\beta)&=\frac{1}{N}\mathbb{E}\log Z_N(\beta)\\&\le\frac{1}{N}\log\mathbb{E}Z_N(\beta)\\&=\frac{1}{N}\log\mathbb{E}\int e^{\beta H_{N,p}(x)}\,d\mu(x)\\&=\frac{1}{N}\log e^{\beta^2N/2}\\&=\frac{\beta^2}{2}.\end{aligned}$$

We use $F_N^{\text{ann}}(\beta)$ to denote the "annealed free energy" $\beta^2/2$. Next, for large β,

$$\log Z_N(\beta)\le\beta\max_{x\in S_N}|H_{N,p}(x)|,$$

so altogether, with probability $1-e^{-N}$, we have the bound

$$F_N(\beta)\le\frac{\beta}{N}\max_{x\in S_N}|H_{N,p}(x)|\le C\beta\ll\frac{\beta^2}{2}.$$

Although this is a result about a random variable with high probability, and the previous bound from Jensen's inequality is one of the expectation, we can make these ideas rigorous next class by proving that:

Proposition 3.2. *For all* β*, we have* $|F_N(\beta)-\mathbb{E}F_N(\beta)|\le o(1)$ *with high probability.*

Now, we turn our attention to the case where β is small. For this, we will make frequent use of the second moment method, which is the general principle for some random variable Z that if $\mathbb{E}[Z^2]\approx\mathbb{E}[Z]^2$ in some sense, then $Z\approx\mathbb{E}[Z]$ in some sense. There are a few different versions of this method, some which involve truncation, i.e. computing $\mathbb{E}[Z^2 1_G]$, where G is some "good event." The strong form of this is the following:

Proposition 3.3. *Suppose that as* $N\to\infty$*, we have*

$$r_N(\beta):=\frac{\mathbb{E}[Z_N(\beta)^2]}{\mathbb{E}[Z_N(\beta)]^2}\to1.$$

Then, $Z_N(\beta)\approx\mathbb{E}Z_N(\beta)$ *in the sense that* $\mathbb{P}[|Z_N(\beta)-\mathbb{E}Z_N(\beta)|\ge\frac{1}{2}\mathbb{E}Z_N(\beta)]\to0$ *as* $N\to\infty$.

Proof of Proposition 3.3. This is a simple application of Chebychev's inequality. As $N\to\infty$,

$$\mathbb{P}\left[|Z_N(\beta)-\mathbb{E}Z_N(\beta)|\ge\frac{1}{2}\mathbb{E}Z_N(\beta)\right]\le\frac{4\,\mathrm{Var}[Z_N(\beta)]}{\mathbb{E}[Z_N(\beta)]^2}=4(r_N(\beta)-1)\to0.$$ □

However, the condition on $r_N(\beta)$ in Proposition 3.3 is far too strong (and too precise), so for us, we'll usually just show that $r_N(\beta) \le e^{o(N)}$ for fixed small β and $N \to \infty$. Surprisingly, this will be enough to get the correct free energy, which is a nice feature of these types of models. Now, we prove this condition for our $r_N(\beta)$.

Definition 3.4 (Overlap)**.** The *overlap* $R\colon S_N \times S_N \to [-1,1]$ is the function $R(x,\tilde{x}) = \langle x,\tilde{x}\rangle / N$.

Being a little loose with the underlying measure μ, we can express

$$\mathbb{E}[Z_N(\beta)^2] = \int_{S_N}\int_{S_N} \mathbb{E}\exp(\beta H_{N,p}(x) + \beta H_{N,p}(\tilde{x}))\,dx d\tilde{x}.$$

The term inside the $\exp(\cdot)$ is $\mathcal{N}(0, 2\beta^2(1+R(x,\tilde{x})^p)N)$, so we can reparameterize in terms of R to get

$$\mathbb{E}[Z_N(\beta)^2] = \int_{-1}^{1} \exp(\beta^2(1+R^p)N)\,d\nu_N(R)$$

for some probability distribution ν_N for R. Moreover, some geometric reasoning about the volume of the cross-sections on the sphere lets us deduce that $d\nu_N(R) \approx (1-R^2)^{N/2}\,dR$, and in fact, it is true that

$$d\nu_N(R) = \alpha_N(1-R^2)^{N/2-1},$$

where α_N is a constant depending on N that is nice in the sense that $\alpha_N \in [1, \mathcal{O}(\sqrt{N})]$. Therefore,

$$\mathbb{E}[Z_N(\beta)^2] = \alpha_N \int_{-1}^{1} \exp N\left(\beta^2(1+R^p) + \frac{1}{2}\log(1-R^2)\right)(1-R^2)^{-1}\,dR.$$

To find the asymptotics of this final integral, we realize that to leading exponential order, it suffices to find the maximum value over R of the expression inside the exponential. In general, this is known as Laplace's principle, and with this we find that as $N \to \infty$,

$$\frac{1}{N}\log\mathbb{E}[Z_N(\beta)^2] \to \max_{-1\le R\le 1}\left\{\beta^2(1+R^p) + \frac{1}{2}\log(1-R^2)\right\}.$$

For small $\beta \le \beta_0$, this maximum is at $R = 0$ (as can be seen visually or with basic calculus), so the maximum value is $\beta^2 = 2F_N^{\text{ann}}(\beta)$, and $r_N(\beta) = e^{o(N)}$ naturally follows. A question now arises about which properties our p-spin model needed to have for this result on $r_N(\beta)$ to hold. For now, this will be answered within the context of spherical spin-glasses. We will consider the mixed p-spin model determined by the following:

Definition 3.5 (Mixed p-spin hamiltonian)**.** A *mixed p-spin hamiltonian* is a function of the form

$$H_N(x) = \sum_{p=1}^{P} \gamma_p H_{N,p}(x),$$

where the $H_{N,p}$'s are p-spin hamiltonians. These are still centered gaussian processes determined by

$$\mathbb{E}H_N(x)H_N(\tilde{x}) = N\xi(R(x,\tilde{x})), \quad \xi(R) = \sum_{p=1}^{P}\gamma_p^2 R^p.$$

As long as $\xi'(0) = 0$ (which is equivalent to $\gamma_1 = 0$), we can follow the argument above to show $r_N(\beta) \le e^{o(N)}$.