Exercise 1. (a) Combining what you have read from the Preface of E. T. Jaynes' "Probability Theory: The Logic of Science" (Cambridge University Press, 2003) and the observation given below, discuss relationships between the mathematical theory of probability and its applications.

Let us consider a probability space with finite $\Omega = \{1, 2, \cdots, n\}$, $\mathcal{F} = 2^{\Omega}$, and $\mathbb{P} = \{p_i | 1 \leq i \leq n\}$, and a random variable $X(\omega) : \Omega \to \mathbb{R}$. If $X_1, X_2, \cdots, X_m$ are a sequence of i.i.d. X's, then

$$\lim_{m\to\infty} \frac{X_1 + X_2 + \cdots + \cdots X_m}{m} = \mathbb{E}[X]. \tag{1}$$

Furthermore, let a function of the sequence of random variables

$$\nu_k(m) = \sum_{i=1}^{m} 1_k(X_i),$$

in which $k \in \Omega$. Then

$$\lim_{m\to\infty} \frac{\nu_k(m)}{m} = p_k. \tag{2}$$

Eqs. (1) and (2) are known as the *law of large numbers* and the *Borel's law of large numbers*, respectively. You note that the terms on the left-hand-side of the two equations are quantities that widely used on measurements in the real world. The terms on the right-hand-side of the two equations, however, are abstract mathematical concepts in the theory of probability.

(b) Check out what *frequentist* and *Bayesian* approaches are. Please comment on the statement in my notes that in terms of the Gibbs conditioning principle, "*statistical mechanics is a nuanced blend of the frequentist and Beyesian schools.*"

Solution 1. (a) Probability theory as I understand it is a way of making sense of the world through small scale modeling. That is, it gives us a way of analysing systems via observation under certain logical paradigms. Taking the law of large numbers as an example, we might collect some data which we believe to have come from some fixed unobserved process with particular properties. We may assume that each datum is an independent observation under our logical paradigm and this allows us to make judgements or form expectations on how our data may behave. In the case of the LLN, this is just the idea that we can approximate with some confidence the theoretical or underlying mean of our observed variables based on a large sampling of those variables. The difference in how we can understand this mean differs though between the Bayesian and frequentist approaches.

(b) Gibbs conditioning in a way is way of conditioning on the empirical statistics of observations from a system, an element of the frequentist school, while generating a posterior distribution on the true statistics of the larger population or system being observed. In this way, I agree that Gibbs conditioning combines elements present in both frequentist and Bayesian schools. I additionally think this is common practice in most interpretations of Bayesian statistics, so I'm unsure if the element of using empirical statistics can be uniquely attributed to the frequentist school.

Exercise 2. Consider a reference measure $\mathbb{P}_1$ and a sequence of measures $\mathbb{P}_2(\theta)$ which contains a continuous parameter θ. The RND

$$Z(\omega;\theta) = \frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)$$

is a random variable that contains the continuous parameter θ. We shall assume that the function $Z(;\theta)$ is differentiable as many times as you like w.r.t. θ, and resulting $\partial^k Z(\omega;\theta)/\partial\theta^k$ are also random variables.
Introducing

$$\mathcal{I}_k(\theta) = -\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial^k}{\partial\theta^k}\log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\right],$$

where the expectation is w.r.t. $\mathbb{P}_2(\theta)$. We shall assume the order of integration involved in taking the expectations and differentiations w.r.t. θ are exchangable. Show that
(a)

$$\mathcal{I}_0(\theta) = -\int_\Omega \left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\mathbb{P}_1(\mathrm{d}\omega)$$

is the Shannon entropy of $\mathbb{P}_2(\theta)$ w.r.t. the measure $\mathbb{P}_1$.
(b) $\mathcal{I}_1(\theta) = 0$.
(c)

$$\mathcal{I}_2(\theta) = \mathbb{E}^{\mathbb{P}_2}\left[\left(\frac{\partial}{\partial\theta}\log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\right)^2\right] \geq 0.$$

This is known as the *Fisher information.*

Solution 2. (a) Starting from the definition of $\mathcal{I}_0(\theta)$, we have that

$$\begin{aligned}
\mathcal{I}_0(\theta) &= -\mathbb{E}^{\mathbb{P}_2}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\right] \\
&= -\int_\Omega \log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\mathbb{P}_2(\mathrm{d}\omega) \\
&= -\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\int_\Omega \log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\mathbb{P}_1(\mathrm{d}\omega),
\end{aligned}$$

where the last line follows from a change of measure using the fact $\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)$ is the Radon-Nikodym derivative of $\mathbb{P}_2$ with respect to $\mathbb{P}_1$.
(b) Starting from the definition of $\mathcal{I}_1(\theta)$, we have that

$$\begin{aligned}
\mathcal{I}_1(\theta) &= -\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial}{\partial\theta}\log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\right] \\
&= -\int_\Omega \frac{\partial}{\partial\theta}\log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\mathbb{P}_2(\mathrm{d}\omega) \\
&= \int_\Omega \frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)^{-1}\mathbb{P}_2(\mathrm{d}\omega),
\end{aligned}$$

where we've used the chain rule to find the partial derivative of $\log Z(\omega;\theta)$ with respect to θ. Next, changing the measure to measure of integration to $\mathbb{P}_1$, we see

$$\begin{aligned}
\mathcal{I}_1(\theta) &= \int_\Omega \frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\mathbb{P}_1(\mathrm{d}\omega)\\
&= \frac{\partial}{\partial\theta}\int_\Omega \left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\mathbb{P}_1(\mathrm{d}\omega)\\
&= \frac{\partial}{\partial\theta}(1)\\
&= 0,
\end{aligned}$$

where we've interchanging the integral and the partial derivative.
(c) We'll begin by doing a bit of calculus using that

$$\frac{\partial}{\partial\theta}\log Z(\omega;\theta) = \left(\frac{\partial}{\partial\theta}Z(\omega;\theta)\right)\cdot Z(\omega;\theta)^{-1},$$

we have that

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\log Z(\omega;\theta) &= \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}Z(\omega;\theta)\right)\cdot Z(\omega;\theta)^{-1}\right]\\
&= \left(\left(\frac{\partial^2}{\partial\theta^2}Z(\omega;\theta)\right)\cdot Z(\omega;\theta)^{-1}\right) - \left(\left(\frac{\partial}{\partial\theta}Z(\omega;\theta)\right)\cdot Z(\omega;\theta)^{-1}\right)^2\\
&= \left(\left(\frac{\partial^2}{\partial\theta^2}Z(\omega;\theta)\right)\cdot Z(\omega;\theta)^{-1}\right) - \left(\frac{\partial}{\partial\theta}\log Z(\omega;\theta)\right)^2.
\end{aligned}$$

The above follows from the product rule for differentiation. We'll now turn our attention to the quantity of interest

$$\begin{aligned}
\mathcal{I}_2(\theta) &= -\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial^2}{\partial\theta^2}\log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\right]\\
&= \mathbb{E}^{\mathbb{P}_2}\left[\left(\frac{\partial}{\partial\theta}\log Z(\omega;\theta)\right)^2\right] - \mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial^2}{\partial\theta^2}\left(Z(\omega;\theta)\right)\cdot Z(\omega;\theta)^{-1}\right].
\end{aligned}$$

Focusing on the last term, we change the measure to $\mathbb{P}_1$ and exchange the order of integration and differentiation

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial^2}{\partial\theta^2}\left(Z(\omega;\theta)\right)\cdot Z(\omega;\theta)^{-1}\right] &= \int_\Omega \frac{\partial^2}{\partial\theta^2}\left(Z(\omega;\theta)\right)\cdot Z(\omega;\theta)^{-1}\mathbb{P}_2(\mathrm{d}\omega)\\
&= \int_\Omega \frac{\partial^2}{\partial\theta^2}\left(Z(\omega;\theta)\right)\mathbb{P}_1(\mathrm{d}\omega)\\
&= \frac{\partial^2}{\partial\theta^2}\int_\Omega Z(\omega;\theta)\mathbb{P}_1(\mathrm{d}\omega)\\
&= 0.
\end{aligned}$$

This leaves us with

$$\mathcal{I}_2(\theta) = \mathbb{E}^{\mathbb{P}_2}\left[\left(\frac{\partial}{\partial\theta}\log\left(\frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\mathbb{P}_1}(\omega;\theta)\right)\right)^2\right] \geq 0,$$

as the integrand itself is non-negative.

Exercise 3. The Legendre-Fenchel transform is introduced in MLN §3.5:

$$\Lambda^*(x) = \sup_{t\in\mathbb{R}} \big\{xt - \Lambda(t)\big\},\ \ x \in \mathbb{R}.$$

Assuming that the $\Lambda(t)$ is strictly convex and twice differentiable, then the supremum in the equation can be obtained by differentiation, which yields

$$\Lambda^*(x) = \begin{cases} \Lambda^*(t) &= x(t)t - \Lambda(t) \\ x(t) &= \Lambda'(t) \end{cases} \tag{3}$$

This gives function $\Lambda^*(x)$ in a parametric form in terms of t. Show that Eq. (3) implies :
(a) $\Lambda^*(x)$ is also convex; and
(b) an inverse, dual relation

$$\Lambda(t) = \sup_{x\in\mathbb{R}} \big\{tx - \Lambda^*(x)\big\}.$$

Solution 3. (a) As $\Lambda(t)$ is strictly convex and twice differentiable, we have that $\Lambda''(t) > 0$ for all t. In what follows, for a fixed point $x \in \mathbb{R}$, we have there is a specific t^* (which depends on x) so that

$$\Lambda^*(x) = xt^* - \Lambda(t^*).$$

Taking the derivative of this with respect to t^*, we have that

$$\Lambda'(t^*(x)) = x.$$

Due to the fact that $\Lambda''(t) > 0$ for all t, $\Lambda'(t)$ is strictly increasing and therefore, there $\Lambda'(t)$ is invertible so that

$$g(x) = (\Lambda')^{-1}(x) = t^\star.$$

Taking the derivative of this with respect to x, we see that

$$\frac{\mathrm{d}g}{\mathrm{d}x}(x) = \frac{1}{\Lambda''(g(x))} > 0$$

and g is differentiable. As

$$\Lambda^*(x) = xg(x) - \Lambda(g(x)),$$

is a composition of differentiable functions, it is differentiable. We can then compute

$$\begin{aligned}\frac{\mathrm{d}\Lambda^*}{\mathrm{d}x} &= g(x) + g'(x)x - g'(x)\Lambda'(g(x)) \\ &= g(x) + (x - \Lambda'(g(x)))g'(x) \\ &= g(x),\end{aligned}$$

where we've used that $g(x) = (\Lambda')^{-1}(x)$. This shows that

$$\frac{\mathrm{d}^2\Lambda^*}{\mathrm{d}x^2} = \frac{\mathrm{d}g}{\mathrm{d}x}(x) > 0$$

which shows that Λ^* is convex.
(b) Using a similar formulation to above, we have that for a fixed t

$$\Lambda^{**}(t) = tx^* - \Lambda^*(x^*)$$

for some x^* which depends on t. Writing t as $(\Lambda')^{-1}(x^*)$, we see that

$$\begin{aligned}\Lambda^{**}(t) &= (\Lambda')^{-1}(x^*)x^* - \Lambda^*(x^*)\\ &= \Lambda((\Lambda')^{-1}(x^*))\\ &= \Lambda(t),\end{aligned}$$

where we've used the definition for $\Lambda(t)$ as well as $t = (\Lambda')^{-1}(x^*)$.

Exercise 4. Let Ω be a simply connected compact domain in $\mathbb{R}^m$. In statistical mechanics, in addition to the concept of "mechanica energy function" $E(\mathbf{x})$ where $\mathbf{x} \in \Omega \subset \mathbb{R}^m$, there is a sequence of probability measures whose density functions w.r.t. the Lebesgue measure is

$$f_n(\mathbf{x}) = A_n e^{-nE(\mathbf{x})}, \tag{4a}$$

in which the A_n is a normalization factor

$$A_n^{-1} = \int_\Omega e^{-nE(\mathbf{x})} \mathrm{d}\mathbf{x}, \tag{4b}$$

and

$$\lim_{n\to\infty} \frac{\ln A_n}{n} = 0. \tag{4c}$$

Eq. (4) is called *Boltzmann-Gibbs distribution* if one replaces the n by the inverse temperature. Let $f_1(x; n)$ be a marginal distribution for x_1, the first component of $\mathbf{x}$:

$$f_1(x; n) = A_n \int_{\Omega\cap\mathbb{R}^{m-1}} e^{-nE(x,\mathbf{y})} \mathrm{d}\mathbf{y},$$

in which $\mathbf{y} = (x_2, \cdots, x_m)$. Assuming in the limit of $n \to \infty$,

$$-\lim_{n\to\infty} \frac{1}{n} \ln f_1(x; n) = \Lambda^*(x), \;\; x \in \mathbb{R}.$$

(a) Show that the n-scaled cumulant generating function, using e^{ntx} instead of e^{tx},

$$\log \int f_1(x; n) e^{ntx} \mathrm{d}x,$$

has the limit:

$$\lim_{n\to\infty} \frac{1}{n} \log \int f_1(x; n) e^{ntx} \mathrm{d}x = \max_{x\in\mathbb{R}} \big\{ tx - \Lambda^*(x) \big\},$$

which is the Legendre-Fenchel transform of $\Lambda^*(x)$. All functions are sufficiently smooth and you are allowed to freely exchange

$$\lim_{n\to\infty} \quad \text{and} \quad \int_{\Omega\cap\mathbb{R}} \mathrm{d}x.$$

(b) Denoting

$$\Lambda(t) = \max_{x\in\mathbb{R}} \big\{ tx - \Lambda^*(x) \big\},$$

show that one can obtain $\Lambda^*(x)$ parametrically as:

$$\Lambda^*(x) = \left\{ \begin{array}{rcl} \Lambda^*(t) & = & -\frac{\mathrm{d}}{\mathrm{d}(1/t)} \left(\frac{\Lambda(t)}{t} \right) \\ x(t) & = & \Lambda'(t) \end{array} \right.$$

Solution 4. (a)
Noting that by assumption,

$$\ln f_1(x;n) \approx -n\Lambda^*(x) \text{ for large } n.$$

In the case of large n, we can write

$$\begin{aligned}\Lambda(nt) &= \log \int f_1(x;n)e^{ntx}\mathrm{d}x \\ &\simeq \log \int e^{ntx} \cdot e^{-n\Lambda^*(x)} \\ &= \log \int e^{n(tx-\Lambda^*(x))}.\end{aligned}$$

Using the Lapalace's method for integral (setting $h(x) = tx - \Lambda^*(x)$) and the assumed smoothness of the functions involved, we see that

$$\lim_{n\to\infty} \frac{1}{n} \log \int f_1(x;n)e^{ntx}\mathrm{d}x = \max_{x\in\mathbb{R}} \big\{tx - \Lambda^*(x)\big\}.$$

(b) We'll repeat a bit from the proof of exercise 3. For any fixed $x \in \mathbb{R}$, there is a specific t which depends on x so that

$$\Lambda^*(x) = x \cdot t(x) - \Lambda(t)$$

Taking the derivative with respect to t, it follows

$$0 = x - \Lambda'(t)$$

Therefore, we know that $x = \Lambda'(t)$. Replacing this shows

$$\Lambda^*(x) = \Lambda'(t) \cdot t - \Lambda(t)$$

Writing this out as

$$\Lambda^*(x) = \Lambda'(u^{-1}) \cdot u^{-2} \cdot u - \Lambda(u^{-1}),\ t = u^{-1},$$

we can see that

$$\begin{aligned}\Lambda^*(x) &= -\frac{\mathrm{d}}{\mathrm{d}u}\big(\Lambda(u^{-1}) \cdot u\big) \\ &= -\frac{\mathrm{d}}{\mathrm{d}(1/t)}\left(\frac{\Lambda(t)}{t}\right).\end{aligned}$$

This suggests that for a given x and $\Lambda(t)$, we can obtain the value of $\Lambda^*(x)$ parametrically.