APMA 1740: Recent Applications of Probability and Statistics

Milan Capoor

Spring 2025

1 Jan 22

1.1 Maximum Entropy Principle

A strange though experiment of Gibbs: Imagine a physical system S (say a gas) in an "infinite bath". Let x be the state of every particle (positions, velocities, ...) in S.

For simplicity, let S be be 3 particles in $\mathbb{Z}^2$ with $x \in \mathbb{Z}^6$ being the positions. Let s be the number of states of particles in S.

What is $p(x)$, the probability that S has state x?

In the simplest case (each particle is independent and the state distribution is uniform), we trivially have $P(x) = \frac{1}{s}$. But in general, these are incredibly strong assumptions.

We can create some constraints to do better.

1. Assume that the average kinetic energy $\mathcal{E}$ of the infinite heat bath is some constant θ.

   In this case, we expect the average kinetic energy of S is approximately θ:

   $$\sum_x p(x)\mathcal{E}(x) = \theta$$

2. Trivially, p is a probability distribution, so

   $$\sum_x p(x) = 1$$

But still this is far from enough: this gives us only 2 constraints for s many unknowns!

However, we can approximate with the LLN. Sample $n \gg s \gg 1$ iid copies of S, $S_1, S_2, \dots, S_n$ with positions $x_1, x_2, \dots, x_n$.

Define the **empirical distribution**

$$\widehat{p}_x = \frac{\#\{i : X_i = x\}}{n}$$

So with large n, $\widehat{p} = p$, and

$$\sum_x \widehat{p}(x)\mathcal{E}(x) \approx \theta$$

Claim: The vast majority of assignments of states to $X_1, \dots, X_n$ yield a single empirical distribution $\widehat{p}$.

Consider $C(\widehat{p})$, the number of ways to assign a state to each of n systems that would yield $\widehat{p}$. Then, with $\widehat{n}_x = \widehat{p}_x \cdot n = \#\{i : X_i = x\}$,

$$C(\widehat{p}) = \binom{n}{\prod_{i=1}^s n_i}$$

2 Jan 24

Recall: For a system S with s states, what is the probability $p(x)$ that S is in state x?

We know that $\sum_{x=1}^{s} p(x) = 1$ and $\sum_{x=1}^{s} p(x)\mathcal{E}(x) = \theta$ for some constant θ.

We sample $X_1, \ldots, X_n$ iid from S ($n \gg s \gg 1$) and define the empirical distribution $\widehat{p}_x = \frac{\#\{i : X_i = x\}}{n}$. By LLN, $\widehat{p} \approx p$.

Claim: $\widehat{p}$ should maximize $C(\widehat{p})$, the number of arrangements of n states $\{1, \ldots, s\}$ that yield $\widehat{p}$:

$$C(\widehat{p}) = \binom{n}{\widehat{p}_1 n \ldots \widehat{p}_s n} = \frac{n!}{(\widehat{p}_1 n)! \ldots (\widehat{p}_s n)!}$$

where $\widehat{p}_i n$ is the number of times we see state i in the sample.

Example: For $s = 2$, put n balls into 2 bins $\{1, 2\}$. Then $\widehat{p}_1 n = a$ balls in bin 1, $\widehat{p} + 2n = n - a$ balls in bin 2. We write this

$$C(\widehat{p}) = \binom{n}{a} = \binom{n}{a, n-a} = \frac{n!}{a!(n-a)!}$$

Stirling's Approximation:

$$k! \approx \frac{k^k}{e^k}\sqrt{2\pi k}$$

Hence,

$$\begin{aligned}
C(\widehat{p}) &= \frac{n^n e^{-n}\sqrt{2\pi n}}{\prod_{i=1}^{s} (\widehat{p}_i n)^{\widehat{p}_i n} e^{-\widehat{p}_i n}\sqrt{2\pi \widehat{p}_i n}} \\
\log C(\widehat{p}) &= n\log n - n + \log\sqrt{2\pi n} - \sum_{i=1}^{s}\left[\widehat{p}_i n \log(\widehat{p}_i n) - \widehat{p}_i n + \log\sqrt{2\pi n}\right] \\
\frac{1}{n}\log C(\widehat{p}) &= \log n - 1 + \frac{1}{n}\log\sqrt{2\pi n} - \sum_{i=1}^{s}\left[\widehat{p}_i \log(\widehat{p}_i n) - \widehat{p}_i + \frac{1}{n}\log\sqrt{2\pi n}\right] \\
&= \log n - \frac{1}{n}\log\sqrt{2\pi n} - \sum_{i=1}^{s}\left[\widehat{p}_i \log(\widehat{p}_i) + \frac{1}{n}\log\sqrt{2\pi n}\right] \\
&= -\sum_{i=1}^{s}\widehat{p}_i\log\widehat{p}_i - \frac{1}{n}\sum_{i=1}^{s}\log\sqrt{2\pi\widehat{p}_i n} + \frac{1}{n}\log\sqrt{2\pi n}
\end{aligned}$$

Since, $\widehat{p}_i \leq 1$, $\frac{1}{n}\log\sqrt{2\pi\widehat{p}_i n} \leq \log n$. Further, $\frac{\log n}{n} \to 0$ so

$$\frac{1}{n}\log C(\widehat{p}) \approx -\sum \widehat{p}_i \log\widehat{p}_i$$

Definition: If p is a probability distribution, its **Shannon Entropy** is

$$H(p) = \sum p(x)\log\frac{1}{p(x)} = -\sum p(x)\log p(x)$$

Note: $H(p) \geq 0$ since $p(x) \leq 1$ for all p.

Back to our original problem, we seek $\widehat{p}$ that satisfies

- $\sum_{x=1}^{s} \widehat{p}_x = 1$
- $\sum_{x=1}^{s} \widehat{p}_x \mathcal{E}(x) \approx \theta$

- $\widehat{p}$ maximizes $C(\widehat{p})$, i.e. maximizes Shannon Entropy $H(\widehat{p})$

We turn to our trusty friend, Lagrange multipliers. We seek to chose p to maximize

$$H(p) + \gamma \sum_{x=1}^{s} p_x + \lambda \sum_{x=1}^{s} p_x \mathcal{E}(x)$$

Taking derivatives WRT p_x,

$$\begin{aligned}\frac{\partial}{\partial p_x}\left[H(p) + \gamma \sum_{x=1}^{s} p_x + \lambda \sum_{x=1}^{s} p_x \mathcal{E}(x)\right] &= \frac{\partial}{\partial p_x}\left[-\sum_x p_x \log p_x\right] + \gamma + \lambda\mathcal{E}(x) \\ &= -\log p_x - 1 + \gamma + \lambda\mathcal{E}(x) = 0\end{aligned}$$

So $\gamma + \lambda\mathcal{E}(x) - 1 = \log p(x)$ and

$$\begin{aligned}p(x) &= e^{-1}e^{\lambda\mathcal{E}(x)}e^{\gamma+\lambda\mathcal{E}(x)} \\ &= \frac{1}{z_\lambda}e^{\lambda\mathcal{E}(x)}\end{aligned}$$

where $Z_\lambda = \sum_{x=1}^{s} e^{\lambda\mathcal{E}(x)}$.

To find λ, we use the constraint $\sum p_x \mathcal{E}(x)\theta$.

3 Jan 27

Example: Find the maximum entropy distribution p on $\{1, 2, 3\}$ (i.e. $s = 3$) satisfying $\mathbb{E}_p X^2 = 2$, i.e. $\sum_{x=1}^{s} p_x x^2 = 2$.

Since $\mathbb{E}_p X^2 = \sum_{x=1}^{s} p(x)x^2 = 2$, $\mathcal{E}(x) = x^2$,

$$p(x) = \frac{1}{Z}e^{\lambda\mathcal{E}(x)} = \frac{1}{Z}e^{\lambda x^2}, \quad x = 1, 2, 3$$

We need to find Z, λ satisfying

- $\mathbb{E}_p X^2 = 2$
- $\sum p_x = 1$

Hence,

$$\begin{aligned}\begin{cases}\frac{1}{Z}[e^\lambda + 4e^{4\lambda} + 9e^{9\lambda}] = 2 \\ \frac{1}{Z}[e^\lambda + e^{4\lambda} + e^{9\lambda}] = 1\end{cases} &\implies Z = e^\lambda + e^{4\lambda} + e^{9\lambda} \\ &\implies e^\lambda + 4e^{4\lambda} + 9e^{9\lambda} = 2(e^\lambda + e^{4\lambda} + e^{9\lambda}) \\ &\implies e^\lambda - 2e^{4\lambda} - 7e^{9\lambda} = 0\end{aligned}$$

We can solve for λ with any numeric method.

3.1 Maximum Entropy Principle in the Continuum

Definition: Let p be a PDF. Its **entropy** is defined as

$$H(p) = -\int_{-\infty}^{\infty} p(x)\log p(x)\, dx$$

Example (MEP with multiple constraints): Find p that maximizes $H(p)$ subject to

$$\begin{cases} \sum p_x \mathcal{E}_1(x) = \theta_1 \\ \vdots \\ \sum p_x \mathcal{E}_k(x) = \theta_k \\ \sum p_x = 1 \end{cases}$$

Our Lagrange multipliers are given by

$$\max \left[H(p) + \lambda_1 \sum p_x \mathcal{E}_1(x) + \lambda_2 \sum p_x \mathcal{E}_2(x) + \cdots + \lambda_k \sum p_x \mathcal{E}_k(x) + \gamma \sum p_x\right]$$

Taking derivatives WRT p_x, we get

$$\begin{aligned} H(p) &= -\log p_x - 1 + \lambda_1 \mathcal{E}_1(x) + \cdots + \lambda_k \mathcal{E}_k(x) + \gamma = 0 \\ \implies p_x &= \frac{1}{Z} \exp\left[\lambda_1 \mathcal{E}_1(x) + \cdots + \lambda_k \mathcal{E}_k(x)\right] \end{aligned}$$

The rest follows as before.

Example: Find the max entropy density subject to $\mathbb{E}_p X^2 = 1$ and $\mathbb{E}_p X = 0$.

In this case,

$$p_x = \frac{1}{Z} \exp\left[\lambda_1 \mathcal{E}_1(x) + \lambda_2 \mathcal{E}_2(x)\right]$$

where

$$\mathcal{E}_1(x) = x^2, \quad \mathcal{E}_2(x) = x$$

Hence, we have constraints

$$\begin{cases} \frac{1}{Z}\left[\int_{-\infty}^{\infty} e^{\lambda_1 x^2 + \lambda_2 x} x^2 \, dx\right] = 1 \\ \frac{1}{Z}\left[\int_{-\infty}^{\infty} e^{\lambda_1 x^2 + \lambda_2 x} x \, dx\right] = 0 \\ \frac{1}{Z}\left[\int_{-\infty}^{\infty} e^{\lambda_1 x^2 + \lambda_2 x} \, dx\right] = 1 \end{cases}$$

We can complete the square to get the integrals in the forms of a Gaussian:

$$\frac{1}{Z} e^{\lambda_1 x^2 + \lambda_2 x} = \frac{1}{Z} \exp\left[\lambda_1 \left(x - \frac{\lambda_2}{2\lambda_2}\right)^2\right] \sim N(\frac{\lambda_2}{2\lambda_1}, \frac{-1}{2\lambda_1})$$

But we have mean 0 and variance 1 so

$$\frac{\lambda_2}{2\lambda_1} = 0 \implies \lambda_2 = 0, \quad -\frac{1}{2\lambda_1} = 1 \implies \lambda_1 = -\frac{1}{2}$$

Z follows from simply computing

$$Z = \int_{-\infty}^{\infty} \exp(\lambda_1 x^2 + \lambda_2 x) \, dx$$

3.2 Large Deviation Principle

Large Deviation Principle: Take p on $\{1, 2, \ldots, s\}$, $\mathcal{E} : \{1, \ldots, s\} \to \mathbb{R}$. Observe $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} p$. Define

$$\frac{1}{n} \sum_{x=1}^{n} \mathcal{E}(X_k) = \theta$$

. Define the empirical distribution $\widehat{p}_x = \frac{1}{n} \cdot \#\{i : X_i = x\}$. Then $\mathbb{E}_{\widehat{p}} \mathcal{E}(X) = \theta$

Proof:

$$\begin{aligned}
\mathbb{E}_{\widehat{p}} \mathcal{E}(X) &= \sum_{x=1}^{s} \widehat{p}_x \mathcal{E}(x) \\
&= \frac{1}{n} \sum_{x=1}^{s} \mathcal{E}(x) \sum_{i=1}^{n} \mathbb{1}_{X_i} \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{x=1}^{s} \mathbb{1}_{X_i = x} \cdot \mathcal{E}(x) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathcal{E}(X_i) = \theta
\end{aligned}$$

Let q be some probability distribution on $\{1, \ldots, s\}$. What is $\mathbb{P}(\widehat{p} = q)$?

Recall that the $C(\widehat{p})$ function gave the number of ways to assign a state to each of n systems that would yield $\widehat{p}$. Similarly, here we have

$$\mathbb{P}(\widehat{p} = q) = \binom{n}{n_1 \cdots n_s} \prod_{x=1}^{s} p_x^{q_x \cdot n}$$

Example: Take $X_1, X_2 \sim p$. Let $q = \frac{1}{2}\delta_\{1\} + \frac{1}{2}\delta_\{2\}$. What is $\mathbb{P}(\widehat{p} = q)$?

1. How many ways can we sample 5 and 1 from X_1, X_2? Two ways: $(1, 5)$ or $(5, 1)$.
2. Now wat is the probability $X_1 = 1, X_2 = 5$? This is $p_1 p_5$. Similarly, $\mathbb{P}(X_1 = 5, X_2 = 1) = p_5 p_1$.

Hence, $\mathbb{P}(\widehat{p} = q) = 2 p_1 p_5$.

4 Jan 29

4.1 Relative Entropy Function

Motivation:

- p a PMF $\{1, \ldots, s\}$
- $\mathcal{E} : \{1, \ldots, s\} \to \mathbb{R}$ an energy function
- $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} p$
- $\widehat{p}$ the empirical distribution, $\widehat{p}_x = \frac{1}{n} \cdot \#\{i : X_i = x\}$

Question: what does $\widehat{p}$ look like?

Let q be a given PMF on $\{1, \ldots, s\}$.

Heuristic: $\frac{1}{n} \log \mathbb{P}(\widehat{p} = q) \approx -D(q \parallel p)$

Remark: We have to be careful about this approximation. Indeed, it holds under LLN for $q = p$ and since we can approximate p via an arbitrary distribution, it holds in general under certain conditions. However, we could easily construct a pathological example:

- $p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
- $q = (\frac{1}{3} + \frac{\sqrt{2}}{K}, \frac{1}{3} + \frac{\sqrt{2}}{K}, \frac{1}{3} + \frac{\sqrt{2}}{K})$ for very large K

Now since p is rational, $\mathbb{P}(\widehat{p}q) = 0$ so $\frac{1}{n} \log \mathbb{P}(\widehat{p} = q) = -\infty$.

KL Entropy:

$$D(q \parallel p) = \sum_{x=1}^{s} q_x \log \frac{q_x}{p_x}$$

measures how close q is to p.

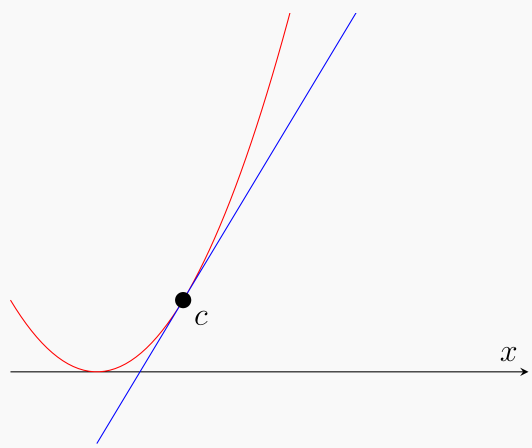
Jensen's Inequality: For every $g : \mathbb{R} \to \mathbb{R}$ convex,

$$\mathbb{E}g(X) \geq g(\mathbb{E}X)$$

Special Case: $\mathbb{E}(X^2) \geq (\mathbb{E}X)^2$

Proof: Consider the tangent line to g at $c = \mathbb{E}X$: $y = g'(c)(x - c) + g(c)$.

By convexity, $g(x) \geq g(c) + g'(c)(x - c)$ for all x.



Hence,

$$\mathbb{E}g(X) \geq \mathbb{E}g'(c)(X - c) + \mathbb{E}g(c) = g'(c)(\mathbb{E}X - c) + g(c) = g(c) = g(\mathbb{E}X)$$

Properties of KL Entropy:

1. $D(q \parallel p) \geq 0$
2. $D(q \parallel p) = 0 \iff q = p$

Proof:

1.

$$
\begin{aligned}
D(q \parallel p) &= \sum_{x=1}^{s} q_x \log \frac{q_x}{p_x} \\
&= \mathbb{E}_q \log \frac{q(X)}{p(X)} \\
&= -\mathbb{E}_q \log \frac{p(X)}{q(X)} \\
&= -\mathbb{E}_q \log Y
\end{aligned}
$$

where $Y = \frac{p_x}{q_x}$. Define $g(y) = -\log y$.

Note g is convex: $g''(y) = \frac{1}{y^2} > 0$. Hence, by Jensen's inequality,

$$
\mathbb{E}g(Y) \geq g(\mathbb{E}Y) = -\log(\mathbb{E}Y) = -\log\left(\mathbb{E}_q \frac{p_x}{q_x}\right) = -\log \underbrace{\left(\sum_{x=1}^{s} q_x \frac{p_x}{q_x}\right)}_{\sum p_x \leq 1} \geq 0
$$

2. For $Y = \frac{p_x}{q_x}$,

$$
\mathbb{E}Y = \sum q_x \frac{p_x}{q_x} = 1 \implies Y = \mathbb{E}Y \text{ a.s.} \implies \frac{p_x}{q_x} = 1 \text{ a.s.} \implies p_x = q_x \quad \forall x \text{ a.s.}
$$

Another Heuristic:

$$
\frac{1}{n} \log \mathbb{P}(\widehat{q} = q) \approx -D(q \parallel p) = -\sum q_x \log \frac{q_x}{p_x}
$$

Find

$$
q = \underset{\sum q_x \mathcal{E}(x) = \theta}{\arg\max} \left(-D(q \parallel p)\right)
$$

using Lagrange multipliers