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 $X_1, \dots, X_n$ 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 $\{1, \dots, s\}$ that yield $\widehat{p}$:

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

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

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.

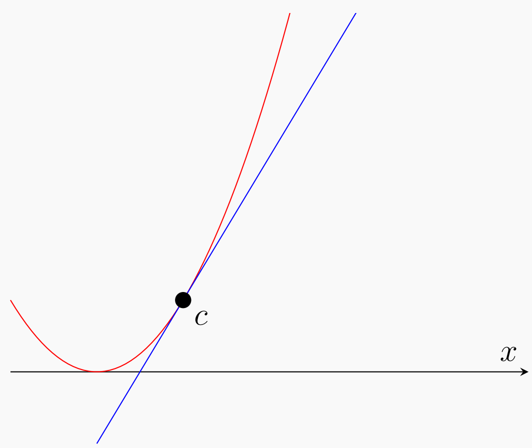
Jensen's Inequality: For every $g : \mathbb{R} \to \mathbb{R}$ convex,

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

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

5 Jan 31

Recall: $D(q \parallel p) = 0$ iff $p = q$.

Proof:

$$\begin{aligned} D(q \parallel p) &= \sum_{x=1}^{s} q_x \log \frac{p_x}{q_x} \\ X \sim q &= \mathbb{E}[\log \frac{q_x}{p_x}] = -\mathbb{E}[\log \frac{p_x}{q_x}] \\ &\overset{\text{Jensen}}{\geq} -\log[\mathbb{E}\frac{p_x}{q_x}] \\ &= -\log[\sum q_x \frac{p_x}{q_x}] = 0 \end{aligned}$$

Hence, we get the equality iff $\mathbb{E}g(Y) = g(\mathbb{E}Y)$ where $Y = \frac{p_x}{q_x}$ $(x \sim q)$ and $g(Y) = -\log Y$. (g is strictly convex, i.e. $\mathbb{E}g(Y) = g(\mathbb{E}Y)$, iff Y is a const a.s.)

But since $Y = \mathbb{E}Y = 1$, $\frac{p_x}{q_x} = 1 \implies p_x = q_x$ a.s.

Last time, we discussed the cases in which the approximation $\mathbb{P}(\widehat{p} = q) \approx D(q \parallel p)$ fails. But why does this happen?

Recall

$$\mathbb{P}(\widehat{p} = q) = \binom{n}{n_1 \cdots n_s} \prod_i p_i^{n_i}$$

where $n_i = q_i \cdot n$.

But this binomial coefficient is well defined only if $q_i n \in \mathbb{N}$ for all i. Hence, the approximation only holds for distributions q with $q_i \cdot n \in \mathbb{N}$ for all i.

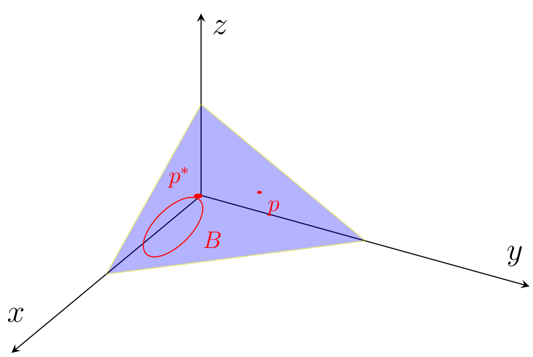
5.1 Sanov's Theorem

Motivation: As usual, let p be a PMF on $\{1, \dots, s\}$ and $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} p$. We know that for large n, $\widehat{p} \approx p$. But this relation is only probabilistic. How do we quantify the probability that $\widehat{p}$ is far from p?

Example: Let $s = 3$ and say $\widehat{p} = (\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (a, b, c)$. Then

$$\begin{cases} a, b, c \geq 0 \\ a + b + c = 1 \end{cases}$$

gives us a triangle in $\mathbb{R}^3$:



Sanov's Theorem: Let B be an open subset of the space of all PMF on $\{1, \dots, s\}$. Then

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(\widehat{p} \in B) = -\inf_{q \in B} D(q \parallel p)$$

Further, if $p^* = \arg\min_{q \in B} D(q \parallel p)$ is unique, then

$$\lim_{n \to \infty} \mathbb{P}(||\widehat{p} - p^*|| > \varepsilon \mid \widehat{p} \in B) = 0 \quad \forall \varepsilon > 0$$

where $||\widehat{p} - p^*||$ is any metric, say $||\widehat{p} - p^*|| = \max_{x \in \{1, \dots, s\}} |\widehat{p}_x - p_x|$

Proof:

Remark: What if $p \in B$? Then $\inf_{q \in B} D(q \parallel p) = 0$, so

$$\frac{1}{n} \log \underbrace{e^{-o(n)}} \mathbb{P}(\widehat{p} \in B) = 0$$

6 Feb 5

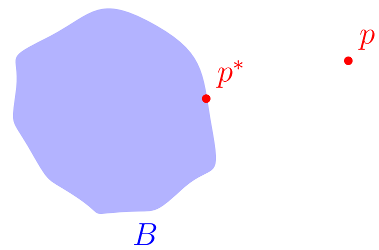
Recall (Sanov's Theorem): For B open,

1.

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p}_{x_1,\dots,x_n} \in B) = -\inf_{q\in B} D(q \parallel p)$$

2. If $\exists!\ p^* = \arg\min_{q\in\overline{B}} D(q \parallel p)$, then

$$\lim_{n\to\infty} \mathbb{P}(||\widehat{p} - p|| > \varepsilon \mid \widehat{p} \in B) = 0 \quad \forall \varepsilon > 0$$



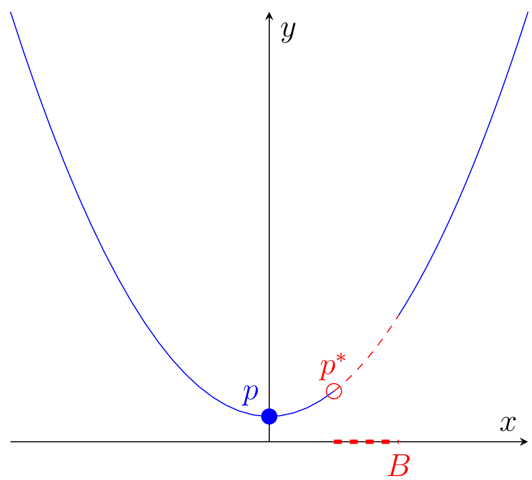
This leads to some interesting questions:

1. Why is p^* drawn on the boundary?
2. Is there a case when p^* lies in the interior?

For the second: yes, if $p \in B$ (in which case p is the global minimizer of $D(q \parallel p)$).

For the first, it suffices to show that since $D(q \parallel p)$ is a convex function, on any set B with $p \notin B$, the minimizer p^* must lie on the boundary.

Example:



Example: $B = \{q \mid \exists x : |q_x - p_x| > 0\}$

By Sanov,

$$\mathbb{P}(\widehat{p}_n \in B) \approx \exp(-n \inf_{q\in B} D(q \parallel p)) \le e^{-n/2} < 10\%$$

Now let's prove the claim:

Proof:

$$
\begin{aligned}
F(q) = D(q \parallel p) &= \sum q_x \log \frac{p_x}{q_x} \\
&= \sum q_x \log q_x - \sum q_x \log p_x \\
\frac{\partial F}{\partial q_x} &= \log q_x + 1 - \log p_x \\
\frac{\partial^2 F}{\partial q_x \, \partial q_y} &= \begin{cases} 1/q_x & x = y \\ 0 & x \neq y \end{cases} \\
H &= \begin{pmatrix} \frac{1}{q_1} & & & \\ & \frac{1}{q_2} & & \\ & & \ddots & \\ & & & \frac{1}{q_s} \end{pmatrix}
\end{aligned}
$$

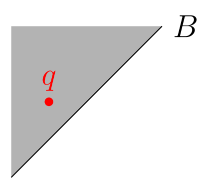
But $\forall v \in \mathbb{R}^s$, $v^T H v = \sum v_i^2 \frac{1}{q_i} \geq 0 \implies H$ is positive semi-definite. Hence F is convex.

6.1 Back to Gibbs' Heat Bath

Recall the original motivating example where $X_1, \ldots, X_n \sim p$, and $\frac{1}{n} \sum_{i=1}^n \mathcal{E}(X_i) = \theta$.

Previosuly, we showed that $\theta = \frac{1}{n} \sum_{i=1}^n \mathcal{E}(X_i) = \mathbb{E}_{\widehat{p}}[\mathcal{E}(X)]$.

Now consider the set $B = \{q \mid \mathbb{E}_q[\mathcal{E}(X)] > \theta\}$ and define $\Omega = \{q : \mathbb{E}_q[\mathcal{E}(X)] = \theta\}$.



Imagine we observe some sample with energy higher than expected (i.e. $q \in B$). What is the probability of this occuring?

By Sanov, in order to find $\inf_{q \in B} D(q \parallel p)$, it suffices to find p^* such that $D(p^* \parallel p) = \inf_{q \in B} D(q \parallel p)$.

In the past, we used Lagrange multipliers to confirm our solution is in the **exponential family**

$$p_x^* = \frac{1}{Z_\lambda} p_x \exp(\lambda \mathcal{E}(x)) \quad \forall x$$

for some λ.

Example of Exponential Family: $\mathcal{N}(\mu, \sigma^2)$ has PDF $\frac{1}{Z} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

If instead we had many constraints $\mathbb{E}_{\widehat{p}}[\mathcal{E}_i(X)] = \theta_i$ for $i = 1, \ldots, k$, we found minimizer

$$p^* = \frac{1}{Z_{\lambda_1 \ldots \lambda_k}} p_x \exp(\lambda_1 \mathcal{E}_1(x) + \cdots + \lambda_k \mathcal{E}_k(x))$$

where we found $\lambda_1, \ldots, \lambda_k$ using Lagrange multipliers to satisfy the constraints and

$$Z_{\lambda_1 \ldots \lambda_k} = \sum_x p_x \exp(\lambda_1 \mathcal{E}_1(x) + \lambda_k \mathcal{E}_k(x))$$

These must also satisfy:

1. $\frac{\partial}{\partial \lambda_k} \log Z_k = \mathbb{E}_\lambda[\mathcal{E}_k(X)]$

2. $\frac{\partial^2}{\partial \lambda_k \lambda_l} \log Z_k = \mathrm{Cov}_\lambda(\mathcal{E}_k(X), \mathcal{E}_l(X)) \quad \forall k, l$

3. $\log Z_k$ is a convex function of λ and it is strictly convex unless $\exists \alpha = (\alpha_1, \ldots, \alpha_k)$ such that $\alpha \neq 0$ and $\sum_{k=1}^{c} \alpha_k \mathcal{E}_k(x) = \text{const} \quad \forall x$

4. $\log Z_\lambda - \sum \lambda_k \theta_k$ is convex in λ and minimized when $\mathbb{E}_\lambda[\mathcal{E}(X)] = \theta_k$

7 Feb 7

Last time, we defined the set

$$B = \{q : \mathbb{E}_q \mathcal{E}(X) < \theta\}$$

For $p \notin B$ known, we know that the minimizer $p^* = \arg\min_{q \in B} D(q \parallel p)$ lies on the boundary of B, $\Omega = \{q : \mathbb{E}_q[\mathcal{E}(X)] = \theta\}$.

Using Lagrange Multipliers, we found

$$p^*_x = \frac{1}{Z_\lambda} p_x e^{\lambda \mathcal{E}(x)} \quad \forall x$$

with

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

Now, we want to find $\lambda = (\lambda_1, \ldots, \lambda_s)$ that satisfies

$$\mathbb{E}_{p^*}[\mathcal{E}(X)] = \theta \iff \sum p^*_x \mathcal{E}(x) = \theta \iff \sum \frac{1}{Z_\lambda} p_x e^{\lambda \mathcal{E}(x)} \mathcal{E}(x) = \theta$$

Proposition:

1. $\frac{\partial}{\partial \lambda_k} \log Z_\lambda = \mathbb{E}_\lambda[\mathcal{E}_k(X)] \quad \forall k = 1, \ldots, c$

2. $\frac{\partial^2}{\partial \lambda_k \, \partial \lambda_l} \log Z_\lambda = \mathrm{Cov}_\lambda(\mathcal{E}_k(X), \mathcal{E}_l(X)) \quad \forall k, l$

3. $\log Z_\lambda$ is convex in λ and, in general, strictly convex (unless the equations $\{\mathbb{E}_{p^*} \mathcal{E}_k(X) = \theta_k\}_{k=1}^{c}$ are redundant, i.e. $\nexists b_1, \ldots b_c \neq (0, \ldots, 0)$)

4. Assuming (3), the function

$$\log Z_\lambda - \sum_{k=1}^{c} \lambda_k \theta_k$$

is in general strictly convex and is minimized when

$$\mathbb{E}_\lambda[\mathcal{E}_k(X)] = \theta_k \quad \forall k$$

(i.e. at exactly the λ that we need to find)

Proof:

1.

$$\begin{aligned}\frac{\partial}{\partial\lambda_k}\log Z_\lambda &= \frac{1}{Z_k}\cdot\frac{\partial}{\partial\lambda_k}Z_\lambda \\ &= \frac{1}{Z_\lambda}\cdot\frac{\partial}{\partial\lambda_k}\left[\sum p_x e^{\lambda_1\mathcal{E}_1(x)+\cdots+\lambda_c\mathcal{E}_c(x)}\right] \\ &= \frac{1}{Z_\lambda}\cdot\sum_x p_x e^{\lambda_1\mathcal{E}_1(x)+\cdots+\lambda_c\mathcal{E}_c(x)}\cdot\mathcal{E}_k(x) \\ &= \frac{1}{Z_\lambda}\cdot\sum_x p_x\mathcal{E}_k(x)e^{\lambda\mathcal{E}(x)} \\ &= \sum_x p_x^*\mathcal{E}_k(x) \\ &= \mathbb{E}_{p^*}[\mathcal{E}_k(X)] = \mathbb{E}_\lambda[\mathcal{E}_k(X)]\end{aligned}$$

Remark: We write $\mathbb{E}_\lambda$ instead of $\mathbb{E}_{p^*}$ just to emphasize that this is a function of λ

2.

Exercise: Email the proof to oanh_nguyen1@brown.edu for bonus points.

Proof: In part 1, we showed that $\frac{\partial}{\partial\lambda_k}\log Z_\lambda = \mathbb{E}_\lambda[\mathcal{E}_k(X)]$. Hence, it suffices now to show

$$\frac{\partial}{\partial\lambda_l}\mathbb{E}_\lambda[\mathcal{E}_k(X)] = \mathrm{Cov}_\lambda(\mathcal{E}_k(X),\mathcal{E}_l(X))$$

TODO

3.

$$H(\lambda_1,\ldots,\lambda_c) = \left(\frac{\partial^2}{\partial\lambda_k\,\partial\lambda_l}\log Z_\lambda\right)_{c\times c}$$

We need to show $\forall v \neq \vec{0}$,

$$v^T Hv = \sum_{k,l} v_k v_l H_{kl} \geq 0 \implies \log_Z \text{ convex}$$

But

$$\begin{aligned}\sum v_k v_l H_{kl} &= \sum v_k v_l \mathrm{Cov}\left(\mathcal{E}_k(X),\mathcal{E}_l(X)\right) \\ &= \mathbb{V}\left(\sum v_k\mathcal{E}_k(X)\right) \geq 0\end{aligned}$$

since

$$\sum v_k v_l \mathrm{Cov}\left(Y_k, T_l\right) = \mathbb{V}\left(\sum v_k y_k\right)$$

8 Feb 10

Let $B = \{q : \mathbb{E}_q[\mathcal{E}(X)] < \theta\}$. Suppose we have two contraints

- $\mathbb{E}_{\widehat{p}}[\mathcal{E}_1(X)] = \theta_1$
- $\mathbb{E}_{\widehat{p}}[\mathcal{E}_2(X)] = \theta_2$

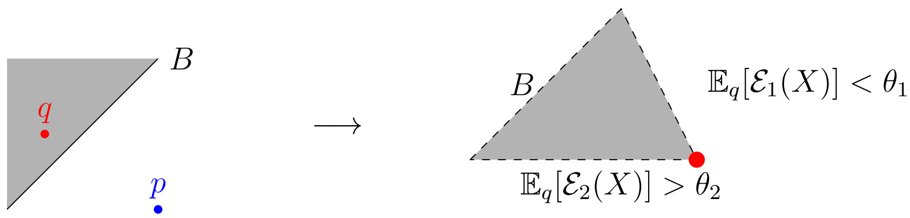
and we know

- $\mathbb{E}_p[\mathcal{E}_1(X)] > \theta_1$
- $\mathbb{E}_p[\mathcal{E}_2(X)] > \theta_2$

Then we can tighten

$$B = \{q : \mathbb{E}_q[\mathcal{E}_1(X)] < \theta_1,\ \mathbb{E}_q[\mathcal{E}_2(X)] > \theta_2\}$$

which updates our partition of the space from:



which tells us

$$\Omega = \{q : \mathbb{E}_q[\mathcal{E}_1(X)] = \theta_1, \quad \mathbb{E}_q[\mathcal{E}_2(X)] = \theta_2\}$$

We already know what to do if $p^* \in \Omega$, so consider just one constraint:

$$\mathbb{E}_q[\mathcal{E}_2(X)] = \theta_2$$

We can easily find p_2^* WRT this constraint:

$$\begin{aligned} B_2 &= \{q : \mathbb{E}_q[\mathcal{E}_2(X)] > \theta_2\} \\ \Omega_2 &= \{q : \mathbb{E}_q[\mathcal{E}_2(X)] = \theta_2\} p_2^* &= \arg\min_{q \in \Omega_2} D(q \parallel p) \end{aligned}$$

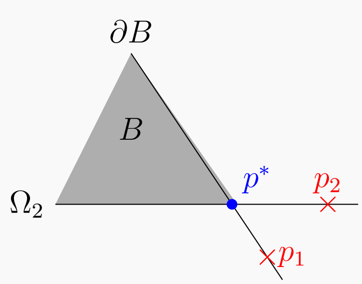
Further, we know if $p_2^* \in \overline{B}$, then $p^* = p_2^*$ and we are done.

Otherwise, we can just try again using the first constraint to find p_1^*. If $p_1^* \in \overline{B}$, then $p^* = p_1^*$ and we are done.

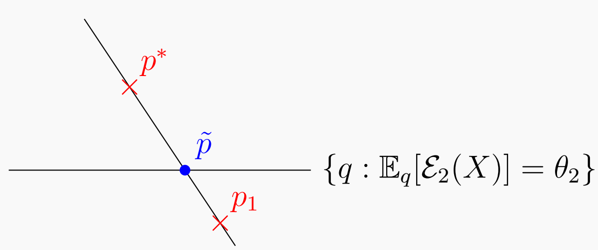
What if we get unlucky both times and $p_1^*, p_2^* \notin \overline{B}$?

Claim: Because of convexity, if $p_1^*, p_2^* \notin \overline{B}$, then $p^* \in \Omega$

Proof:



WLOG, $p^* \in \Omega_1$ so let $\tilde{p} = [p^*, p_1^*] \cap \Omega \implies \tilde{p} \in \Omega$.



Then the $\tilde{p}$ should have been p^* (contradiction.)

Or

$$\tilde{p} = \lambda p^* + (1-\lambda)p^*_\perp \quad \lambda(0,1)$$

so

$$D(\tilde{p} \parallel p) \leq \lambda D(p^* \parallel p) + (1-\lambda)D(p^*_\perp \parallel p)$$

but $D(p^* \parallel p)$ and $D(p^*_\perp \parallel p)$ are the smallest among the points while $D(\tilde{p} \parallel p)$ should be the largest. Contradiction.

8.1 Information Point of View for Shannon Entropy

In the following section, let $\log = \log_2$

Here, **Shannon Entropy** "measures the minimal number of bits needed to encode a message optimally".

For example, let $X_1, \ldots, X_n \sim \{1,2\}$ with $p = (p_1, p_2)$ and $p_2 = 1 - p_1$.

As before, let $\widehat{p}_1 = \frac{\#\{i : X_i = 1\}}{n}$ and $\widehat{p}_2 = 1 - \widehat{p}_1$.

Question: What is the probability of any particular sequence? (say $\widehat{p}_1 \approx p_1, \widehat{p}_2 \approx p_2$)

Answer:

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) &= p_1^{\widehat{p}_1 n} p_2^{\widehat{p}_2 n} \\
&\approx p_1^{p_1 n} p_2^{p_2 n} \\
&= 2^{n(\log p_1)p_1} \cdot 2^{n(\log p_2)p_2} \\
&= 2^{-nH(p)}
\end{aligned}$$

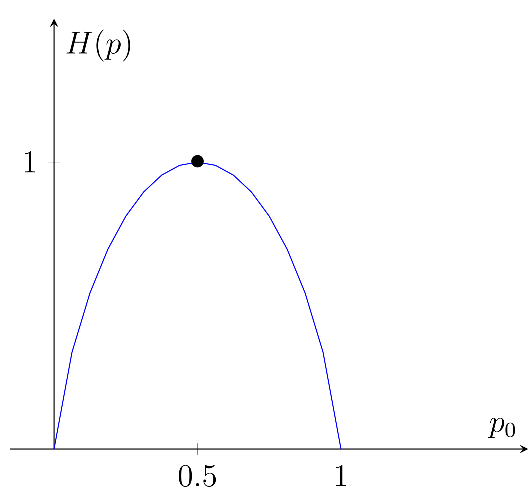
and this makes some sense: if we have no information, we would expect the probability of any sequence to be 2^{-n}.

9 Feb 12

Let $\{X_i\}_{i=1}^n \sim \{0,1\}$ with $p = (p_0, p_1) = (p_0, 1-p_0)$. The Shannon Entropy is

$$\begin{aligned}
H(p) &= -\sum p_x \log p_x \\
&= -p_0 \log p_0 - p_1 \log p_1 \\
&= -p_0 \log p_0 - (1-p_0)\log(1-p_0) = F(p_0)
\end{aligned}$$

for some function F.



What is the relationship between the Shannon Entropy and the KL-Divergence?

$$\begin{aligned}
D(p \parallel h) &= \sum p_x \log \frac{p_x}{h_x} \\
&= \sum p_x \log p_x - \sum p_x \log h_x \\
&= -H(p) - \log \frac{1}{s}
\end{aligned}$$

for $h \sim \text{Unif}(1, s)$. Hence, up to a constant, $H(p) \approx D(p \parallel \text{Unif}\{1, \ldots, s\})$.

And indeed this jusitifies that $H(p)$ has its max at $1/2$ when $p = (1/2, 1/2)$.

This also explains what we found last class: we only need $2^{nH(p)}$ bits rather than 2^n because in the worst case, $H(p) = 1 \implies 2^{n \cdot 1} = 2^n$.

9.1 Source Coding

More generally, we can take $X = (X_1, \ldots, X_n) \sim p$ on states $\{1, \ldots, t\}$ for $t = 2^n$.

Let $C : \{1, \ldots, t\} \to \{0, 1\}^*$ be a **source code** where $\{0, 1\}^*$ is the set of finite non-empty strings of 0s and 1s.

We let $|C(x)|$ denote the length of the code. In general, we want $|C(x)|$ to be small across different x.

Example: A trivial code is the identity: $C(x) = x$ for all x. For $p = 1/2$, this is the best we can do.

If, however, $p = (0.99, 0.01)$ we can do better in expectation.

Prefix: A *prefix code* is a code C for which $C(x)$ is not a prefix for $C(\tilde{x})$ for any $x \neq \tilde{x}$.

Example:

x	$C(x)$	$C'(x)$
1	0	0
2	1	10
3	00	11

Here, C is not a prefix because under C, if we are trying to encode 0100, we do not know if it should be 120 or 1211. However, C' is a prefix because there is no ambiguity.

Remark: Being a prefix is not necessary for unique decoding. For example,

x	$C(x)$
1	0
2	01
3	011

is not a prefix but any string can be uniquely decoded by looking back.

Question: Whaat is the minimal $(|C(x)|)_x$ (i.e. $C = \arg\min \mathbb{E}_p |C(x)| = \sum p_x |C_x|$) where C is a prefix code?

If we simply return the message, every encoded message is of equal length so C is a prefix code of expected length n. Can we do better?

Proposition (Kraft-McMillan Inequality): For all prefix codes C,

$$\sum_{x=1}^{t} 2^{-|C(x)|} \leq 1$$

and for any code lengths $\ell_1, \dots, \ell_t$ such that

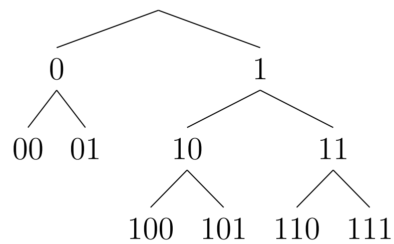
$$\sum_{x=1}^{t} 2^{-\ell_x} \leq 1$$

there exists a a prefix code C with $|C_x| = \ell_x$ (letting $C_x = C(x)$).

Example: In the non-prefix example, we say $\ell_1 = 1, \ell_2 = 2, \ell_3 = 3$ so

$$\sum_{x=1}^{t} 2^{-\ell_x} = 2^{-1} + 2^{-2} + 2^{-3} \leq 1 \quad \checkmark$$

We can visualize this as a tree:



We will see next time that the optimal code C^* satisfies $H(p) \leq \mathbb{E}\,|C^*(X)| \leq H(p)$

10 Feb 14

Motivation: Let $p = (p_1, p_2)$ be a distribution on $\{0, 1\}$ ($s = 2$).

Sample $(X_1, \dots, X_n)$ corresponding to n bits. Hence, there are 2^n possible sequences.

We can design a prefix code $C : \{0,1\}^n \to \{0,1\}^*$.

Example: For $n = 3$,

$X_1X_2X_3$	$C(X_1X_2X_3)$
000	00
001	01
$\vdots$	
111	

with $\mathbb{E}_p[|C_x|] \approx H(p)n$. And indeed this is a prefix since every image is the same length.

We know that for the identity code, $C(x) = x$, $\mathbb{E}_p[\left|C_{(X_1,\dots,X_n)}\right|] = n$.

Theorem: Let $\vec{X} \sim \vec{p}$. For the optimal code $C^* = \arg\min_{C \text{ prefix}} \mathbb{E}_{\vec{p}}[|C(X)|]$,

$$H(\vec{p}) \leq |\mathbb{E}_{\vec{p}}|\, C^*(X) \leq H(\vec{p}) + 1$$

Remark: In our example, $\vec{X} = (X_1, \dots, X_n), \quad X_i \overset{\text{iid}}{\sim} p$ so

$$H(\vec{p}) \leq \mathbb{E}_{\vec{p}}\,|C(X)| \leq H(\vec{p}) + 1$$

where $\vec{p} = p \otimes \cdots \otimes p$.

Claim:

1. $H(\vec{p}) = nH(p)$.
2. $H(X,Y) = H(X) + H(Y)$ if X, Y independent

Proof: 1. Follows as a corollary from (2).

---

2. Let X take values $\{x_1, \dots, x_A\}$ and Y take values $\{y_1, \dots, y_B\}$.

Then

$$\begin{aligned}
H(X,Y) &= -\sum_{i=1}^{AB} p_i \log p_i \\
&= -\sum_{x=1}^{A}\sum_{y=1}^{B} p_{xy} \log p_{xy} \\
&= -\sum_{x}\sum_{y} p_x q_y \log p_x q_y \qquad (X, Y \text{ independent}) \\
&= -\sum_{x}\sum_{y} p_x q_y \log p_x + p_x q_y \log q_y \\
&= -\sum_{y} p_y \sum_{x} p_x \log p_x - \sum_{x} p_x \sum_{y} q_y \log q_y \qquad (\text{Tonelli}) \\
&= \sum_{y} q_y H(x) + \sum_{x} p_x H(y) \\
&= H(X) + H(Y) \quad \blacksquare
\end{aligned}$$

Hence,

$$nH(p) \le \mathbb{E}|C(X)| \le nH(p) + 1$$

In particular, our propositions from earlier in the week follow immediately. Most importantly, we have confirmed that we indeed only need $2^{nH(p)}$ bits to encode a message.

At last, we are ready to actually prove the theorem:

Theorem: Let $\vec{X} \sim \vec{p}$. For the optimal code $C^* = \arg\min_{C \text{ prefix}} \mathbb{E}_{\vec{p}}[|C(X)|]$,

$$H(\vec{p}) \le |\mathbb{E}_{\vec{p}}|\, C^*(X) \le H(\vec{p}) + 1$$

Proof: Let $X \sim p$.

1. $H(p) \le \mathbb{E}_p\,|C(X)|$

   Let $\ell_x = |C_x|$. Then

$$\begin{aligned}
\mathbb{E}\,|C(X)| - H(p) &= \sum p_x \ell_x + \sum p_x \log p_x \\
&= \sum p_x \log(2^{\ell_x} p_x) \\
&= \sum p_x \log \frac{p_x}{2^{-\ell_x}} \\
&= \sum p_x \log \frac{p_x}{2^{-\ell_x} \cdot \frac{\sum_y 2^{-\ell_y}}{\sum_y 2^{-\ell_y}}}
\end{aligned}$$

Let $S = \sum_x 2^{-\ell_x}$. By Kraft-McMillan, $S \leq 1$ so

$$= \sum_x p_x \log \frac{p_x}{q_x S} \tag{1}$$

$$= \sum_x p_x \log \frac{p_x}{q_x} - \sum_x p_x \log S \tag{2}$$

$$= D(p \parallel q) - \log S \geq 0 \tag{3}$$

2. $\mathbb{E}\,|C^*(X)| \leq H(p) + 1$.

   It suffices to show $\exists C$ prefix such that

$$\mathbb{E}_p\,|C(X)| \leq H(p) + 1$$

   In fact, our Part I gives us a place to start: We would like to find ℓ_x such that $q_x \propto 2^{-\ell_x} \approx p_x$. Hence, let $\ell_x = \left\lceil \log_2 \frac{1}{p_x} \right\rceil$.

   Now, we just need to show $\exists C$ prefix such that $\ell_x = |C_x|$. But by Kraft-Mcmillan, it suffices to show $\sum_x 2^{-\ell_x} \leq 1$.

   With a little more work, we can show this exactly. Heuristically, if we did not need to round to get an integer ℓ_x, we would have $H(p)$ exactly. Rounding, we get $H(p) + 1$.