APMA 1740: Recent Applications of Probability and Statistics

Milan Capoor

Spring 2025

0.1 Jan 22

0.1.1 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, $S_1, S_2, \ldots, S_n$ with positions $x_1, x_2, \ldots, x_n$.

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 $X_1, \ldots, X_n$ 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 $\widehat{n}_x = \widehat{p}_x \cdot n = \#\{i : X_i = x\}$,

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

0.2 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 $\sum p_x \mathcal{E}(x)\theta$.

0.3 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,

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

We can solve for λ with any numeric method.

0.3.1 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

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

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

0.4 Jan 29

0.4.1 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.

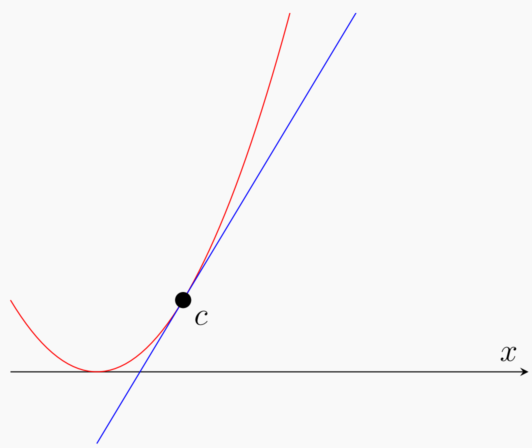
Jensen's Inequality: For every $g : \mathbb{R} \to \mathbb{R}$ convex,

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

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

$$q = \operatorname*{arg\,max}_{\sum q_x \mathcal{E}(x) = \theta} (-D(q \parallel p))$$

using Lagrange multipliers

0.5 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.

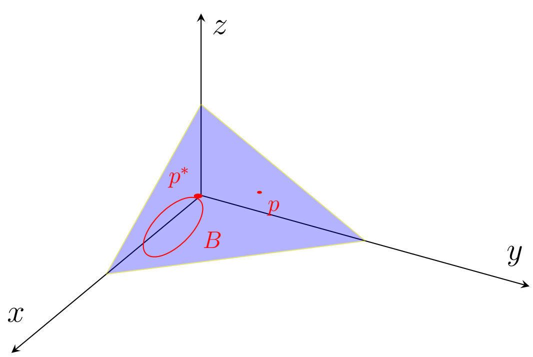
0.5.1 Sanov's Theorem

Motivation: As usual, let p be a PMF on $\{1, \ldots, s\}$ and $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} p$. 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 $\{1, \ldots, s\}$. 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 $||\widehat{p} - p^*|| = \max_{x \in \{1, \ldots, s\}} |\widehat{p}_x - p_x|$

Proof:

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

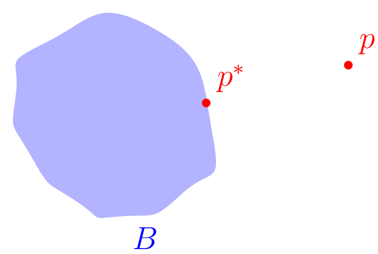
$$\frac{1}{n} \log \underbrace{e^{-o(n)}} \mathbb{P}(\widehat{p} \in B) = 0$$

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



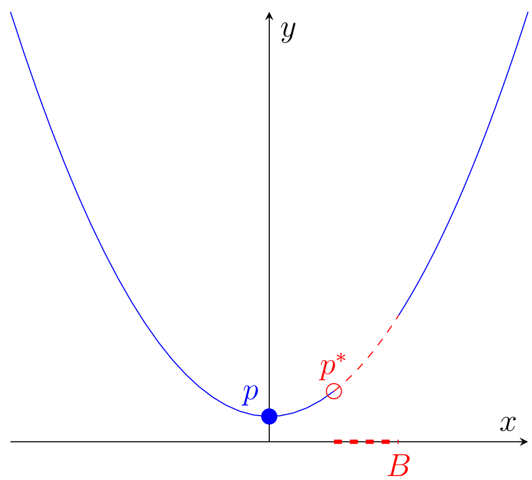
This leads to some interesting questions:

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

Now let's prove the claim:

Proof:

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

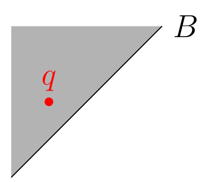
But $\forall v \in \mathbb{R}^s$, $v^T H v = \sum v_i^2 \frac{1}{q_i} \geq 0 \implies H$ is positive semi-definite. Hence F is convex.

0.6.1 Back to Gibbs' Heat Bath

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

Previously, 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 occurring?

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

0.7 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

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

TODO

3.

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

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

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

But

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

since

$$
\sum v_k v_l \operatorname{Cov}\left(Y_k, T_l\right) = \operatorname{Var}\left(\sum v_k y_k \right)
$$

0.8 Feb 10

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

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

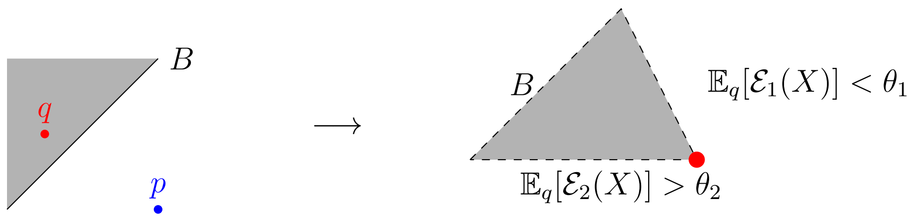
and we know

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

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:

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

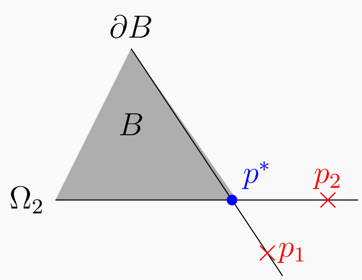
Further, we know if $p_2^* \in \overline{B}$, then $p^* = p_2^*$ and we are done.

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.

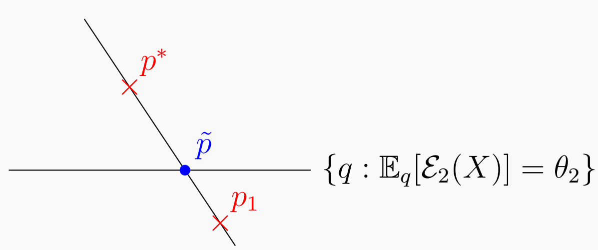
What if we get unlucky both times and $p_1^*, p_2^* \notin \overline{B}$?

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.

0.8.1 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:

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

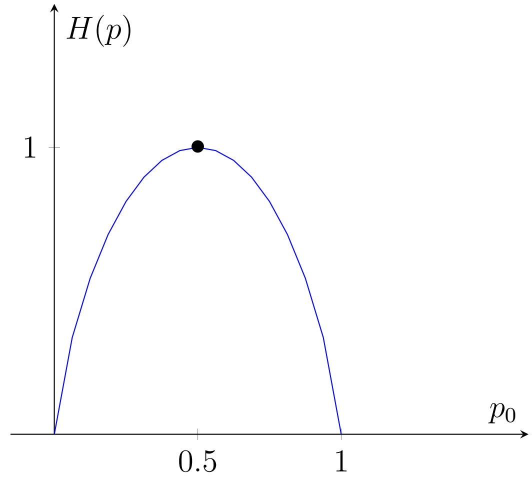
and this makes some sense: if we have no information, we would expect the probability of any sequence to be 2^{-n}.

0.9 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 justifies 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, $H(p) = 1 \implies 2^{n \cdot 1} = 2^n$.

0.9.1 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: What 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 $\ell_1, \ldots, \ell_t$ such that

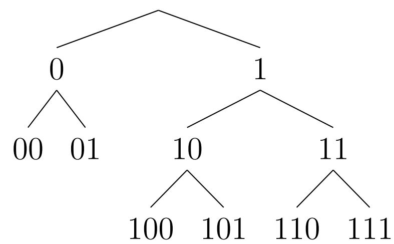
$$\sum_{x=1}^{t} 2^{-\ell_x} \leq 1$$

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

0.10 Feb 14

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

Sample $(X_1, \ldots, X_n)$ 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, $C(x) = x$, $\mathbb{E}_p[\left|C_{(X_1,\ldots,X_n)}\right|] = n$.

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, $\vec{X} = (X_1, \ldots, X_n), \quad X_i \overset{\text{iid}}{\sim} p$ 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 $\{x_1, \ldots, x_A\}$ and Y take values $\{y_1, \ldots, y_B\}$.

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,

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

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 $C^* = \arg\min_{C \text{ prefix}} \mathbb{E}_{\vec{p}}[|C(X)|]$,

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

Proof: Let $X \sim p$.

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

   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 $H(p) + 1$.

0.11 Feb 19

Example: $s = 3$ with $p = (1/2, 1/4, 1/4)$.

Then

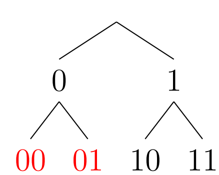
$$H(p) = \sum p_x \log \frac{1}{p_x} = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = \frac{3}{2}$$

If we want to encode $X_1 \cdots X_n$, we have 3^n possible sequences. We would naturally like to design a prefix code C with length $\left\lceil \log_2 \frac{1}{p_x} \right\rceil$.

One way is via block coding. We first choose the lengths:

X_1	p_x	$\ell_x = \left\lceil \log_2 \frac{1}{p_x} \right\rceil$
1	1/2	1
2	1/4	2
3	1/4	2

If we say $C(1) = 0$, then we can prune the resulting tree for all other encodings:



which naturally leads us to a full prefix code:

X_1	$C(x)$
1	0
2	10
3	11

Example: Now consider $s = 3$, $p = (1/3, 1/3, 1/3)$. Then $H(p) = \log 3 \approx 1.58$. So

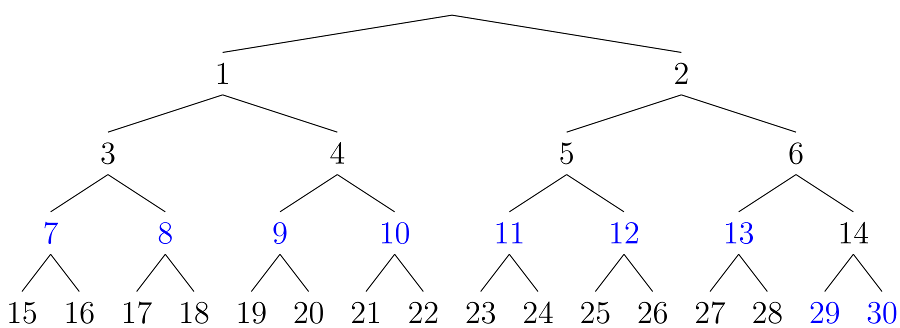
For $n = 1$,

x	$p(x)$	ℓ_x	$C(x)$
1	1/3	$\lceil \log_2(3) \rceil = 2$	0
2	1/3	2	10
3	1/3	2	11

with

$$\mathbb{E}\,|C_x| = \frac{2}{3}(2) + \frac{1}{3}(1) = \frac{5}{3}$$

But with $n = 2$, we have $3^2 = 9$ possible sequences. Looking at the tree, we can choose a reasonable minimal encoding:



which gives

x	$p(x)$	ℓ_x	$C(x)$
11	1/3	4	000
12			001
13			$\vdots$
21			
22			
23			
31			110
32			1110
33			1111

which has

$$\mathbb{E}\,|C_x| = \frac{7}{9}(3) + \frac{2}{9}(4) \approx 3.222 = 1.611 \cdot 2$$

which means we use 1.611 bits per signal.

If $n \to \infty$, then the best prefix code has an average $H(p)$ bits per symbol.

Chapter 1

Statistical Inference

1.1 Feb 19

1.1.1 Probability Estimation

Motivation: Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta$. We want to estimate θ.

Example: If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then $\theta = (\mu, \sigma)$.

Unbiased Estimation: Suppose $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$ is an estimation of θ. If $\mathbb{E}[\widehat{\theta}] = \theta$, we say $\widehat{\theta}$ is *unbiased.*

Example: Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

- $\widehat{\mu} = \frac{1}{n}(X_1 + \cdots + X_n)$ is unbiased since

$$\mathbb{E}[\widehat{\mu}] = \frac{1}{n}\sum \mathbb{E}[X_i] = \frac{1}{n}(n)(\mu) = \mu$$

- What is an unbiased estimator for σ^2? We know $\sigma^2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \mathbb{E}[(X-\mu)^2]$ so

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$$

- In fact, $\widehat{\widehat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$ is a biased estimator:

Proof: WLOG $\mu = 0$ (else $Y_i = X_i - \mu \sim \mathcal{N}(0, \sigma^2) \implies \widehat{\mu}_X = \widehat{\mu}_Y - \mu$).

Then $\sigma^2 = \mathbb{E}[X^2]$ so

$$\begin{aligned}
\widehat{\mu} &= \frac{1}{n}\sum X_i \\
\widehat{\sigma}^2 &= \frac{1}{n-1}\sum (X_i - \widehat{\mu})^2 \mathbb{E}[\widehat{\sigma}^2] & = \mathbb{E}\left[\frac{1}{n-1}\sum (X_i - \widehat{\mu})^2\right] \\
&= \frac{1}{n-1}\sum \mathbb{E}[(X_i - \widehat{\mu})^2] \\
&= \frac{n}{n-1}\mathbb{E}[(X_i - \widehat{\mu})^2] \\
&= \frac{n}{n-1}\mathbb{E}\left[\left(X_i - \frac{X_1 + \cdots + X_n}{n}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{n-1}{n}X_1 - \frac{1}{n}X_2 \cdots - \frac{1}{n}X_n\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{n-1}{n}\right)^2 X_1^2 + \sum_{i=2}^{n}\frac{1}{n^2}X_1^2 + 2\sum_{i\neq j} X_i X_j\right] \\
&= (\frac{n-1}{n})^2\mathbb{E}[X_1^2] + \frac{n-1}{n^2}\mathbb{E}[X_1^2] \\
&= \frac{(n-1)^2}{n^2}\sigma^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

since for $i \neq j$, $\mathbb{E}[X_i X_j] \overset{X_i \perp X_j}{=} (\mathbb{E}X_i)(\mathbb{E}X_j)$

Consistent: We say $\widehat{\theta}_n$ is *consistent* if $\widehat{\theta}_n \longrightarrow \theta$ in some sense as $n \to \infty$. For example,

- $\widehat{\theta}_n \xrightarrow{a.s.} \theta \implies \mathbb{P}(\lim_{n\to\infty} \widehat{\theta}_n = \theta) = 1$
- $\widehat{\theta}_n \xrightarrow{P} \theta \implies \forall \varepsilon > 0, \mathbb{P}(\left|\widehat{\theta}_n - \theta\right| > \varepsilon) \xrightarrow{n\to\infty} 0$
- $\widehat{\theta} \xrightarrow{\text{mean square}} \theta \implies \mathbb{E}[(\widehat{\theta}_n - \theta)^2] \to 0.$

Is $\widehat{\sigma}^2$ consistent in any sense? As we will see, yes. But not trivially so.

1.2 Feb 21

Recall: Let $\theta = \sigma^2$ and take $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} p$. Then

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator for σ^2.

Further,

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is a biased estimator for σ^2.

Mean Squared Error (MSE): $\operatorname{MSE}(\widehat{\theta}_n) = \mathbb{E}\left|\widehat{\theta}_n - \theta\right|^2$.

Notice,

$$
\begin{aligned}
\text{MSE}\,(\widehat{\theta}) &= \mathbb{E}(\widehat{\theta}_n - \theta)^2 \\
&= \mathbb{E}(\underbrace{\widehat{\theta}_n - \mathbb{E}\widehat{\theta}_n}_{a} + \underbrace{\mathbb{E}\widehat{\theta}_n + \mathbb{E}\widehat{\theta}_n - \theta}_{b})^2 \\
&= \mathbb{E}(a + b^2) \\
&= \mathbb{E}a^2 + 2b\underbrace{\mathbb{E}a}_{0} + \underbrace{b^2}_{\text{bias}^2} \\
&= \text{Var}\,(\widehat{\theta}) + \text{bias}^2
\end{aligned}
$$

Example: Calculate MSE (S_n^2) vs. MSE $(\widehat{\sigma}_n^2)$. For simplicity, assume $\mu = 0$,$\sigma^2 = 1$ and $\mathbb{E}_p X^4 = 3$.

$$
\begin{aligned}
\text{MSE}\,(S_n^2) &= \text{Var}\,(S_n^2) + \text{bias}^2 \\
&= \text{Var}\,(S_n^2) \qquad \text{since } S_n^2 \text{ is unbiased} \\
&= \mathbb{E}[(S_n^2 - \mathbb{E}S_n^2)^2] \\
&= \mathbb{E}[(S_n^2 - \sigma^2)^2] \\
&= \mathbb{E}[(S_n^2 - 1)^2] \\
&= \mathbb{E}[S_n^4] - 2\mathbb{E}[S_n^2] + 1 \\
&= \mathbb{E}[S_n^4] - 2 + 1 \\
&= \mathbb{E}[S_n^4] - 1
\end{aligned}
$$

We know

$$
\begin{aligned}
S_n^2 &= \frac{1}{n-1}\left(\sum (X_i - \frac{\sum X_j}{n})\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{n-1}{n}X_i - \frac{1}{n}\sum_{j\neq i} X_j\right)^2
\end{aligned}
$$

We want

$$
\mathbb{E}[S_n^4] = \frac{1}{(n-1)^2}\mathbb{E}\left[\sum_i \left(\frac{n-1}{n}X_i - \frac{1}{n}\sum_{j\neq i} X_j\right)^2\right]^2
$$

up to coefficients, we will only have $X_i^4, X_i^3X_j, X_i^2X_j^2, X_i^2X_jX_k, X_iX_jX_kX_l$ terms in the expansion.

Under expectation, however, only the X_i^4 and $X_i^2X_j^2$ terms will survive.

After a little more work, we find

$$
\begin{aligned}
\text{MSE}\,(S_n^2) &= \frac{2}{n-1}\sigma^4 \\
\text{MSE}\,(\widehat{\sigma}_n^2) &= \frac{2n-1}{n^2}\sigma^4
\end{aligned}
$$

but then MSE $(\widehat{\sigma}_n^2) <$ MSE (S_n^2) so even though it is biased, it is a better estimator (in the sense of minimizing MSE).

1.2.1 Nonparametric Estimation

Example: Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} p$. We want to estimate p.

Suppose we have one observation $\widehat{p}_x = \frac{1}{n}\#\{i : X_i = x\}$. How good an estimator is this?

First, is it unbiased? We know that for a set B,

$$\widehat{p}(B) = \frac{1}{n} \cdot \#\{i : X_i \in B\} = \sum_{x \in B} \widehat{p}_x$$

and

$$\mathbb{E}[\widehat{p}(B)] = \frac{1}{n}\sum_i \mathbb{E}[\mathbb{1}_{X_i \in B}] = \frac{1}{n}\sum_i p(B) = p(B)$$

so $\widehat{p}_x$ is unbiased.

Next, is it consistent? That is, for B measurable, does $\widehat{p}_n(B) \to p(B)$ in some sense?

By LLN,

$$\widehat{p}_n(B) = \frac{1}{n}\sum_i \mathbb{1}_{X_i \in B} = \frac{1}{n}\sum_i Y_i \xrightarrow{a.s.} \mathbb{E}Y = \mathbb{E}\mathbb{1}_{X_i \in B} = \mathbb{P}(X_i \in B) = p(B)$$

Exercise: In the above proof, we depended on B being fixed. Here we show that this condition was necessary.

Let $p = \mathcal{N}(0,1)$. For all n, show that there exists a set $B_n(X_1, \ldots, X_n)$ such that $\widehat{p}_n(B_n)$ is far from $p(B_n)$.

1.3 Feb 24

Motivation: Let f be the density of p. We want to estimate f. We can approximate $\widehat{p}$ but this is discrete so we cannot have a continuous $\widehat{f}$.

Formally, how can we approximate the Dirac measure $\delta_a(A) = \mathbb{1}_{a \in A}$ by a continuous measure?

1.3.1 Kernel Density Estimation

Density Function: a function k satisfying

1. $k(x) \geq 0$
2. $\int xk(x)\,dx = 0$
3. $\int x^2 k(x)\,dx = 1$

i.e $Y \sim k \implies \mathbb{E}Y = 0 \wedge \operatorname{Var} Y = 1$.

Example: $k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Example: We want to approximate δ_0. For $Z = 0$, we know $\delta_0 = \operatorname{dist}(Z)$.

One approach is to approximate Z by $Z + Y$ where Y is continuous (hence $\mathbb{E}Y = 0$) and therefore $Z + Y$ is continuous.

A natural solution is $Y_\varepsilon \sim \mathcal{N}(0, \varepsilon)$ for $\varepsilon \ll 1$. Notice, $Y_0 \sim \mathcal{N}(0,1) \implies \varepsilon Y_0 \sim \mathcal{N}(0, \varepsilon^2)$.

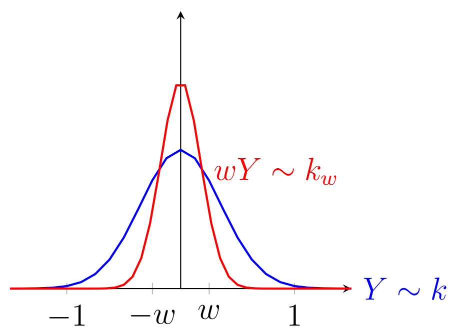
In general, if $Y \sim k$, what is the density of εY?

We can consider the CDF:

$$
\begin{aligned}
F_Y(x) &= \mathbb{P}(Y \leq x) = int^x_{-\infty} k(t)\, dt \\
F_{\varepsilon Y}(x) &= \mathbb{P}(\varepsilon Y \leq x) = \mathbb{P}\left(Y \leq \frac{x}{\varepsilon}\right) = \int_{-\infty}^{x/\varepsilon} k(s)\, ds \\
&\overset{s=t/\varepsilon}{=} \int_{-\infty}^{x} k\left(\frac{t}{\varepsilon}\right) \frac{dt}{\varepsilon} \\
&\implies k_\varepsilon(t) = \frac{1}{\varepsilon} k\left(\frac{t}{\varepsilon}\right)
\end{aligned}
$$

Definition: for each **smoothing parameter** w (aka bandwidth),

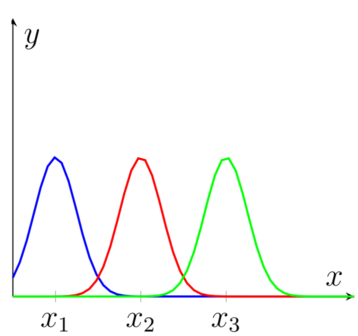
$$k_w(x) = \frac{1}{w} k\left(\frac{x}{w}\right)$$



Now, our goal is to find the optimal w to approximate $Z(\sim \delta_0)$ by $Z + Y_w$.

Correspondingly, we approximate $f(x)$ by

$$\widehat{f}(x) = \widehat{f}_{n,w}(x, X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} k_w(x - X_i)$$



Our plan is to use MSE $=$ bias^2 + variance as

$$
\begin{array}{c|cc}
w \searrow 0 & \text{bias} \searrow & \text{variance} \nearrow \\
w \nearrow \infty & \text{bias} \nearrow & \text{variance} \searrow
\end{array}
$$

Integrated Square Error (ISE):

$$\text{ISE } = \int_{\mathbb{R}} \left| \widehat{f}_n(x, X_1, \ldots, X_n) - f(x) \right|^2 \, dx$$

Since this is a random variable, we can also define *mean integrated square error*.

Mean Integrated Square ERROR (MISE):

$$\begin{aligned}
\text{MISE} = \mathbb{E}[\text{ISE}\,] &= \int_{\mathbb{R}} \mathbb{E}\left|\widehat{f}(x, X_1, \ldots, X_n) - f(x)\right|^2 \, dx \\
&= \int_{\mathbb{R}} \mathbb{E}\left|\widehat{f}(x, X_{1:n}) - \mathbb{E}[\widehat{f}_n(x, X_{1:n})] + \mathbb{E}[\widehat{f}_n(x, X_{1:n})] - f(x)\right|^2 \, dx \\
&= \int_{\mathbb{R}} \left|\mathbb{E}\widehat{f}_n(x, X_{1:n}) - f(x)\right|^2 \, dx + \int_{\mathbb{R}} \mathbb{E}\left|\widehat{f}_n(x, X_{1:n}) - \mathbb{E}[\widehat{f}_n(x, X_{1:n})]\right|^2 \, dx \\
&= \int_{\mathbb{R}} \underbrace{\left|\mathbb{E}\widehat{f}_n(x, X_{1:n}) - f(x)\right|^2}_{\text{bias}^2} \, dx + \int_{\mathbb{R}} \underbrace{\text{Var}\left[\widehat{f}_n(x, X_{1:n})\right]}_{\text{variation}} \, dx
\end{aligned}$$

We can apply this formula to the kernel density estimator so we have bias:

$$\begin{aligned}
B_{n,w}(x) &= \mathbb{E}[\widehat{f}_n(x, X_1, \ldots, X_n)] - f(x) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[k_w(x - X_i)] - f(x) \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} f(t)k_w(x-t)\,dt - f(x) \\
&= \int_{\mathbb{R}} f(t)k_w(x-t)\,dt - f(x)
\end{aligned}$$

1.4 Feb 26

Recall: For a continuous density f with $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$, we would like to estimate f but our normal method $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ is discrete, hence insufficient.

Hence, we introduce the *Kernel Density Estimator*:

$$\widehat{f}_{n,w}(x, X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} k_w(x - X_i)$$

where

$$k_w(t) = \frac{1}{w}k(\frac{t}{w}), \quad k \text{ some density}$$

is parameterized by the bandwidth w.

Remark: Above, we are using the Dirac Measure $\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$ instead of the indicator function ($\mathbb{1} : \mathbb{R} \to \mathbb{R}$) because we need a measure and not a function.

Goal: Find the "optimal" w.

We introduced the *Integrated Square Error* (ISE), $\int_x \left|\widehat{f}(x) - f(x)\right|^2 \, dx$ and the *Mean Integrated Square Error* (MISE)

$$\text{MISE} = \mathbb{E}[\text{ISE}] = \int_x [(\text{bias}(x))^2 + \text{Var}\,(x)] \, dx$$

where

$$\begin{aligned}
\text{bias(x)} &= \mathbb{E}[\widehat{f}(x)] - f(x) \\
\mathbb{E}[\widehat{f}(x)] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X_i}[k_w(x - X_i)] \\
&= \mathbb{E}[k_w(x - X_1)] \qquad (X_i \overset{\text{iid}}{\sim} f) \\
&= \int_{\mathbb{R}} f(t) k_w(x-t)\; dt
\end{aligned}$$

Convolution: Let $Z \sim f$ and $Y \sim g$ be independent. Then

$$Z + Y \sim (f \star g)(x) = \int_{\mathbb{R}} f(t)\; g(x-t)\; dt$$

Hence,

$$\mathbb{E}[\widehat{f}(x)] = (f \star k_w)(x)$$

which means that $\mathbb{E}[\widehat{f}]$ is the density of $Z + Y_w$ where $Z \perp Y_w$ and $Z \sim f$ and $Y_w \sim k_w$.

What does this tell us about the behavior?

- For $Y \sim k$, $Y_w \sim wY$ so $\mathbb{E}[\widehat{f}] \to f$ as $w \to 0$.
- As $w \to \infty$, our support becomes infinitely large so $\mathbb{E}[\widehat{f}] \to 0$.

Hence,

$$(\text{bias}(x))^2 = (\mathbb{E}[\widehat{f}(x)] - f(x))^2 = \begin{cases} 0 & w \to 0 \\ f^2(x) & w \to \infty \end{cases}$$

Now, let's calculate the variance term:

$$\begin{aligned}
\text{Var}\,(x) &= \text{Var}\,(\widehat{f}(x)) \\
&= \text{Var}\,\left(\frac{1}{n}\sum_{i=1}^{n} k_w(x - X_i)\right) \\
&= \frac{1}{n^2}\sum \text{Var}\,(k_w(x - X_i)) \qquad \text{(independence)} \\
&= \frac{1}{n}\text{Var}\,(k_w(x - X_i)) \qquad \text{(identically distributed)} \\
&= \underbrace{\frac{1}{n}\mathbb{E}[(k_w(x - X_i))^2]}_{V^{(1)}} - \underbrace{\frac{1}{n}[\mathbb{E}[k_w(x - x_i)]]^2}_{V^{(2)}}
\end{aligned}$$

From our previous work,

$$V^{(2)} = \frac{1}{n}\mathcal{I}^2 \to \begin{cases} \frac{1}{n} f^2(x) & w \to 0 \\ 0 & w \to \infty \end{cases}$$

and

$$\begin{aligned}
V^{(1)} &= \frac{1}{n}\int f(g)k_w^2(x-t)\,dt \\
&= \frac{1}{n}\frac{1}{w}\int f(t)\frac{1}{w}k^2\left(\frac{x-t}{w}\right)\,dt \\
&= \frac{1}{n}\frac{1}{w}\int f(ws+t)k^2(s)\,ds \qquad (s=\frac{x-t}{w}) \\
&\to \begin{cases} \infty & w\to 0 \\ 0 & w\to\infty \end{cases}
\end{aligned}$$

since the constant $\frac{1}{w}$ term dominates the bounded f, k.

1.5 Feb 28

Theorem: Assume f and k smooth. Then as $w \to 0$,

$$\text{MISE}_{n,w} = \underbrace{\alpha w^4}_{\text{bias}} + \underbrace{\frac{\beta}{nw}}_{\text{variance}} + \text{error}$$

How do we choose w? Ignoring α, β, it makes sense we want to minimize MISE:

$$(w^4 + \frac{1}{nw})' = 4w^3 - \frac{1}{nw^2} = 0 \implies w^5 \propto \frac{1}{n} \implies w \propto n^{-1/5}$$

This is **Sylverman's Rule of Thumb:** up to unknown bias and variance, choose $w = n^{-1/5}$.

However, assuming we do not know α, β, this is not a very good estimate – it does not even depend on the density f! Can we do better?

Recall the setup: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f$ with estimator

$$\widehat{f}_{n,w}(x) = \frac{1}{n}\sum_{i=1}^{n} k_w(x - X_i)$$

We want to find w. Last time, we looked at the MISE. This time, consider only the ISE. Our goal is to minimize:

$$\begin{aligned}
\text{ISE} &= \int_x \left|\widehat{f}_{n,w}(x) - f(x)\right|^2\,dx \\
&= \int \widehat{f}_{n,w}(x) - 2\int \widehat{f}\cdot f + \int f^2(x)\,dx
\end{aligned}$$

Define

$$\begin{aligned}
I &= \int_x \widehat{f}_{n,w}(x) \cdot f(x)\, dx \\
&= \mathbb{E}_{X_{n+1}\sim f}[\widehat{f}_{n,w}(X_{n+1})] \qquad (X_{1:n} \overset{\text{iid}}{\sim} f) \\
&= \mathbb{E}[\widehat{f}_{n,w}(X_{n+1}; X_1, \dots, X_n)] \\
&\approx \mathbb{E}[\widehat{f}_{n-1,w}(X_n; X_1, \dots, X_{n-1})] \\
&\approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\widehat{f}_{n-1,w}(X_i; {}_i X)] \qquad ({}_i X = X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} \widehat{f}^{(i)}_{n-1,w}(X_i)
\end{aligned}$$

We call this the **cross-validation** (leave-one-out) estimator.

Since the last term does not depend on w, it suffices to find

$$\arg\min_w \widehat{J}(w) = \int \widehat{f}_{n,w}(x)^2 - 2\frac{1}{n}\sum_{i=1}^{n} \widehat{f}^{(i)}_{n-1,w}(X_i)$$

And this is exactly what we want since this minimization problem depends only on the kernel and not on the distribution f.

Theorem (Stone 1984):

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{\text{ISE}\left(\widehat{f}_{\widehat{w}_n}, f\right)}{\inf_w \text{ISE}\left(\widehat{f}_{w,n}, f\right)} = 1\right) = 1$$

(i.e. almost surely)

However, this convergence could be very slow (especially for $X_i \sim f \in \mathbb{R}^d,\ d \gg 1$)

Example: For f Gaussian in $\mathbb{R}^d$ with $f(0) = \left(\frac{1}{\sqrt{2\pi}}\right)^d$, to have

$$\left|\widehat{f}_{\widehat{w}_n} - f(0)\right| \le \frac{1}{10} f(0)$$

d	n
1	4
2	19
5	768
10	842000
$\vdots$	

which is very fast growth

1.6 March 3

1.6.1 Maximum Likelihood Estimation

Setup: Sample $X_1, \dots, X_n \sim p_\theta$ with θ unknown. We want to find θ that makes $X_1 = x_1, \dots, X_n = x_n$ most likely (i.e. the parameter that defines the distribution that best fits the observation)

$$
\begin{aligned}
\widehat{\theta} &= \arg\max_{\tilde{\theta}} p_{\tilde{\theta}}(X_1 = x_1, \ldots, X_n = x_n) \\
&= \arg\max_{\tilde{\theta}} p_{\tilde{\theta}}(x_1) \cdots p_{\tilde{\theta}}(x_n) \\
&= \arg\max_{\tilde{\theta}} \frac{1}{n} \sum_{i=1}^{n} \log p_{\tilde{\theta}}(x_i) \\
&= \arg\max_{\tilde{\theta}} \sum_{x=1}^{s} (\log p_{\tilde{\theta}}(x)) \widehat{p}(x) \\
&= \arg\min_{\tilde{\theta}} - \sum_{x=1}^{s} \widehat{p}(x) \log p_{\tilde{\theta}}(x) \\
&= \arg\min_{\tilde{\theta}} - \sum_{x=1}^{s} \widehat{p}(x) \log \frac{\widehat{p}(x)}{p_{\tilde{\theta}}(x)} - \sum \widehat{p}(x) \log \widehat{p}(x) \\
&= \arg\min_{\tilde{\theta}} D(\widehat{p} \,\|\, p_{\tilde{\theta}}) + H(\widehat{p}) \\
&= \arg\min_{\tilde{\theta}} D(\widehat{p} \,\|\, p_{\tilde{\theta}})
\end{aligned}
$$

since the entropy term does not depend on θ.

Conclusion: MLE is equivalent to finding the distribution that is closest in KL-divergence to the empirical distribution.

Example: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$. Equivalently, $f_\mu = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$.

Hence,

$$
\begin{aligned}
\widehat{\mu} &= \arg\max_{\tilde{\mu}} \prod_{i=1}^{n} f_\mu(x_i) \\
&= \arg\max \exp\left(\sum - \frac{(x_i - \tilde{\mu})^2}{2} \right) \\
&= \arg\min \sum_{i=1}^{n} (x_i - \tilde{\mu})^2 \\
&\overset{*}{=} \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}
$$

Exercise: Prove the starred equality above.

Example: $X_1, \ldots, X_n \sim f_\lambda = \frac{1}{Z_\lambda} p(x) \exp\left(\sum_{j=1}^{c} \lambda_j \mathcal{E}_j(x) \right)$

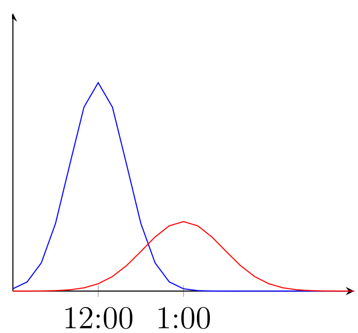
Then

$$
\begin{aligned}
\widehat{\lambda} &= \arg\max \prod f_{\tilde{\lambda}}(x_i) \\
&= \arg\min -\frac{1}{n}\sum_{i=1}^{n} \log(f_{\tilde{\lambda}}(x_i)) \qquad \text{(invariant under constant)} \\
&= \arg\min -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{Z_\lambda} p(x_i)\exp\left(\sum_{j=1}^{c}\lambda_j\mathcal{E}_j(x)\right)\right) \\
&= \arg\min -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{Z_\lambda}\exp\left(\sum_{j=1}^{c}\lambda_j\mathcal{E}_j(x)\right)\right) \qquad (p(x_i)\text{ known}) \\
&= \arg\min \log(Z_{\tilde{\lambda}}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\tilde{\lambda}_j\mathcal{E}_j(x_i) \\
&= \arg\min \log(Z_{\tilde{\lambda}}) - \sum_{j=1}^{c}\lambda_j\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_j(x_i)\right) \\
&= \arg\min \log(Z_{\tilde{\lambda}}) - \sum_{j=1}^{c}\lambda_j\theta_j
\end{aligned}
$$

where $\theta_j = \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_j(x_i)$ are the observed statistics.

1.6.2 Classification

Motivation: Suppose we set up outside the dining hall and observe the patterns of the rush. There is a large group that comes in at noon and another group that comes in later



If we interviewed a student at 2:00, it is quite likely they will be from group two. Similarly, if we interviewed a student at 10:00, they are likely from group one. But what about at 12:30? This is the problem of classification.

Formally, let $X \in \mathbb{R}^d$ be some random variable. Let $Y = \{1, \ldots, c\}$ be classes with $\pi_i = \mathbb{P}(Y = i)$ and $f_i(x)$ the class conditioned density (in the example above, f_2 would be the red curve).

Then for any set A,

$$P_X(A) = \sum_{i=1}^{c}\pi_i\mathbb{P}(A)$$

and

$$f_X(A) = \sum_{i=1}^{c}\pi_i f_i(A)$$

We define a **classification** $h : \mathbb{R}^d \to \{1, \ldots, c\}$.

1.7 March 5

Bayes' Classification Rule:

$$h^*(x) = \arg\max_{i=1:c} \mathbb{P}(Y = i \mid X = x)$$

Example: $Y \in \{1, 2\}$.

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = x) &= \frac{\mathbb{P}(Y = 1, X = x)}{\mathbb{P}(X = x)} \\
&= \frac{\mathbb{P}(Y = 1)\mathbb{P}(X = x \mid Y = 1)}{\mathbb{P}(X = x)} \\
&= \frac{\pi_1 \cdot f_1(x)}{\mathbb{P}(X = x)} \\
\mathbb{P}(Y = 2 \mid X = x) &= \frac{\pi_2 \cdot f_2(x)}{\mathbb{P}(X = x)}
\end{aligned}$$

Hence,

$$h^*(x) = \begin{cases} 2 & \pi_1 f_1(x) < \pi_2 f_2(x) \\ 1 & \text{otherwise} \end{cases}$$

In what sense can we say h^* is the "best" classifier?

$$\mathbb{P}(h^*(X) \neq Y) \leq \mathbb{P}(h(x) \neq Y) \qquad \forall h : \mathbb{R}^d \to \{1, \ldots, c\}$$

Exercise: Prove the optimality of Bayes' classification rule. Hint:

$$\mathbb{P}(h^*(X) = Y) = \int_x \mathbb{P}(Y = h^*(x) \mid X = x) f_X(x)\, dx$$

Proof:

$$\begin{aligned}
\mathbb{P}(h^*(x) \neq Y) &= 1 - \mathbb{P}(h^*(x) = Y) \\
&= 1 - \int_x \mathbb{P}(Y = h^*(x) \mid X = x) f_X(x)\, dx \\
&= 1 - \int_x \mathbb{P}(Y = \arg\max_i [\mathbb{P}(Y = i \mid X = x)]; \mid X = x) f_X(x)\, dx \\
&\leq 1 - \int_X \mathbb{P}(Y = h(x) \mid X = x) f_X(x)\, dx \\
&= 1 - \mathbb{P}(h(X) = Y) \\
&= \mathbb{P}(h(X) \neq Y)
\end{aligned}$$

In applications, however, we may be able to approximate the f_i's by sampling but not necessarily the π_i's.

Neyman-Pearson (NP) Classification: Fix $t \in (0, \infty)$. Then

$$h_t(x) = \begin{cases} 1 & \text{if } \frac{f_2(x)}{f_1(x)} < t \\ 2 & \text{if } \frac{f_2(x)}{f_1(x)} > t \end{cases}$$

Remark: In the case $t = \frac{\pi_1}{\pi_2}$, then NP is equivalent to Bayes.

If $Y = 1$ represents a negative test and $Y = 2$ represents a positive test, we have

- The detection rate $\mathbb{P}(h(X) = 2 \mid Y = 2)$
- The false alarm rate $\mathbb{P}(h(X) = 2 \mid Y = 1)$

Intuitively, we would like to maximize the detection rate while minimizing the false alarm rate.

Theorem: Fix $t \in (0, \infty)$. Let h be any other classifier. If $\text{FAR}_h \leq \text{FAR}_{h_t}$, then $\text{DR}_h \leq \text{DR}_{h_t}$

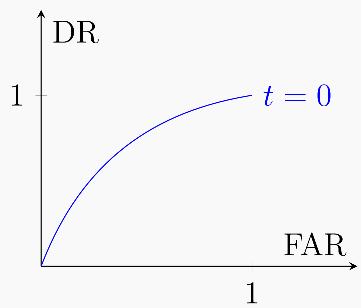
Intuition:

We have

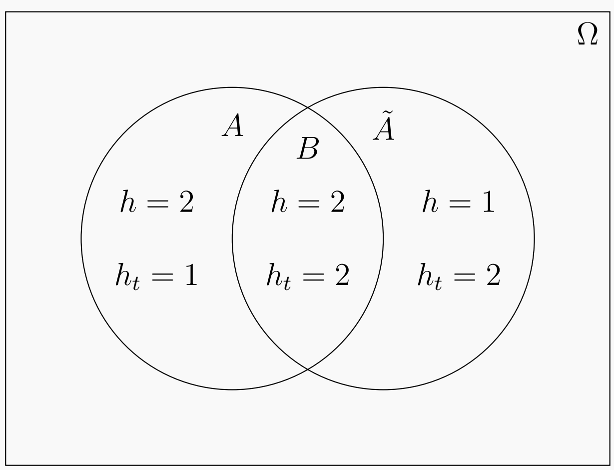
$$\text{FAR}_{h_t} = \mathbb{P}(h(X) = 2 \mid Y = 1) = \mathbb{P}(\frac{f_2(x)}{f_1(x)} > t \mid Y = 1)$$

so as $t \to \infty$, $\text{FAR}_{h_t} \searrow$ and $\text{DR}_t \searrow$

Under NP classification,



Proof:



We have

$$\begin{aligned}\text{FAR}_h &= \mathbb{P}(h(X) = 2 \mid Y = 1) \\ &= \mathbb{P}(A \cup B \mid Y = 1) \\ &= p_1(A \cup B)\end{aligned}$$

where p_1 is the marginal conditioned on the class being 1.

Then:

1. $p_1(A \cup B) \leq p_1(\tilde{A} \cup B) \implies p_1(A) \leq p_1(\tilde{A})$ (since A, B disjoint).

We want to show $p_\alpha(A) \leq p_2(\tilde{A})$

Notice

$$\begin{aligned} A \subseteq \{h_t(x) = 1\} = \left\{\frac{f_2(x)}{f_1(x)} < t\right\} = \{f_2(X) < t \cdot f_1(X)\} \\ \tilde{A} \subseteq \{h_t(x) = 2\} = \left\{\frac{f_2(x)}{f_1(x)} > t\right\} = \{f_2(X) > t \cdot f_1(X)\} \end{aligned}$$

We have

$$p_1(X \in A) \leq p_1(X \in \tilde{A})$$
$$p_1(X \in A) = t \int_A f_1(X)\, d\mathbb{P} \leq t \int_{\tilde{A}} f_1(X)\, d\mathbb{P}$$

1.8 March 7

Motivation: for training data $(X_1, Y_1) \dots (X_n, Y_n)$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \{1, \dots, s\}$, we would like to build $h : \mathbb{R}^d \to \{1, \dots, s\}$.

There are multiple different approaces:

- Generative
- Discriminative
- Algorithmic

1.8.1 Generative Classifiers

$$h^*(x) = \underset{x \in \{1,\dots,s\}}{\arg\max}\, \mathbb{P}(Y = c \mid X = x) = \arg\max_c \frac{\pi_c f_c(x)}{\mathbb{P}(X = x)}$$

A good estimator given data $(X_1, Y_1) \dots (X_n, Y_n)$ is clearly

$$\widehat{\pi}_c = \frac{\#\{i : Y_i = c\}}{n}$$

But what if we do not know f_c? This gets especially difficult when d is large.

Naive Bayes: Assume $X = (X^1, X^2, \dots, X^d)$ and $f_c(x^1, \dots, x^d) = f_c^1(x^1) \cdots f_c^d(x^d)$.

Then instead of needing to find $(f_c)^s$ with $f_c : \mathbb{R}^d \to \mathbb{R}$, it suffices to find $(f_c^k)_{k=1:d}^{c=1:s}$ for $f_c^k : \mathbb{R} \to \mathbb{R}$

Quadratic Discriminant Analysis (QDA): Assume

$$f_c(x) \sim \mathcal{N}(\mu_c, \Sigma_c)$$

Then

$$f_c(x, \mu_c, \Sigma_c) = \frac{1}{\sqrt{(2\pi)^d\, |\Sigma|}} \exp(-\frac{1}{2}(x - \mu_c)^T \Sigma_c^{-1}(x - \mu))$$

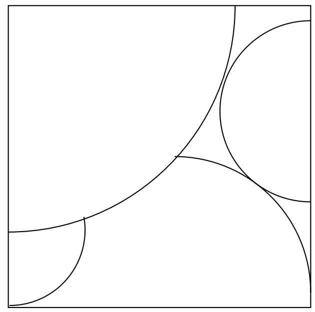
for $x \in \mathbb{R}^{d \times 1}$, $\mu_c \in \mathbb{R}^{d \times 1}$, $\Sigma_c \in \mathbb{R}^{d \times d}$

However, if we are to attempt MLE on μ_c, Σ_c, we need to ensure that we have enough data n to estimate the $d^2 s$ parameters across classes $\{1, \dots, s\}$. In practice, this can lead to over fitting.

1.9 March 10

Definition: If we partition a space $\Omega = \bigcup_{i=1}^{s} A_i$ into disjoint sets, then *precision boundary* of A_i is $\partial A_i = A_i \setminus \mathring{A}_i$.

Last time, we saw a classification method that let us use the MLE on high-dimensional spaces but which required a lot of data in practice. This was the Quadratic Discriminant Analysis (QDA), which had a quadratic precision boundary.



We can follow a similar but slightly less flexible approach.

Linear Discriminant Analysis:

Assume $f_c \sim \mathcal{N}(\mu_c, \Sigma)$.

Then the precision boundary is given by

$$\left\{ x \in \mathbb{R}^d : \frac{f_2(x)}{f_1(t)} = t \right\}$$

from NP. We claim this is a linear set.

Proof: By assumption, $f_1 \sim \mathcal{N}(\mu_1, \Sigma)$ where $\mu_1 \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$.

Then

$$f_1(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{d/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1) \right)$$

notice

$$(x-\mu_1)^T \Sigma^{-1} (x-\mu) \in \mathbb{R}^{(1\times d)(d \times d)(d \times 1)} = \mathbb{R}^1$$

Similarly, we can write

$$f_2(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{d/2}} \exp\left(-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1} (x-\mu_2) \right)$$

so

$$\log t = \log \frac{f_1}{f_2} = \frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1) - (x-\mu_2)^T \Sigma^{-1} (x-\mu_2)$$

In the case $d = 1$, we have $\Sigma = \sigma^2$ so

$$\log t = \frac{(x-\mu_1)^2}{\sigma^2} - \frac{(x-\mu_2)^2}{\sigma^2} = -\frac{2x(\mu_1-\mu_2) + \mu_1^2 + \mu_2^2}{\sigma^2}$$

but in the QDA case, we would have $\Sigma = (\sigma_1^2, \sigma_2^2)$ so the terms would not cancel.

1.9.1 Discriminative Construction

Recall that in the Bayes' classification rule (the optimal case),

$$h^*(x) = \underset{c \in \{1,\ldots,s\}}{\arg\max}\, \mathbb{P}(Y = c \mid X = x)$$

Earlier, we wrote $\mathbb{P}(Y = c \mid X = x) = \pi_c f_c(x)$ and tried to estimate π_c and f_c. But what if we tried to estimate $r_c(x) = \mathbb{P}(Y = c \mid X = x)$ directly?

Linear Regression: For $s = 2$, we want $r_1(x)$ and $r_2(x)$ satisfying $r_1(x) + r_2(x) = 1$. It seems reasonable to try linear regression.

We can model

$$\log \frac{r_2(x)}{r_1(x)} = \alpha + \beta x$$
$$\log \frac{r_2}{1 - r_2} = \alpha + \beta x$$
$$r_2 = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

Softmax: Softmax is a generalization of logistic regression for $s > 2$.

As before,

$$\log \frac{r_k(x)}{r_1(x)} = \alpha_k + \beta_k x \implies r_k(x) = \frac{e^{\alpha_k+\beta_k x}}{1 + \sum_{k=n}^{s} e^{\alpha_k+\beta_k x}}$$

Then we can use MLE to estimate α_k, β_k.

k-Nearest Neighbor Classification:

Let $D_k(x)$ be the closed ball centered at x with radius $R_k(x)$, the smallest radius that contains k data points.

Then