APMA 1740: Recent Applications of Probability and Statistics

Milan Capoor

Spring 2025

Chapter 1

Information Theory

1.1　Jan 22

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

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

1.2 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,

$$C(\widehat{p}) = \frac{n^n e^{-n}\sqrt{2\pi n}}{\prod_{i=1}^{s}(\widehat{p}_i n)^{\widehat{p}_i n} e^{-\widehat{p}_i n}\sqrt{2\pi \widehat{p}_i n}}$$

$$\log C(\widehat{p}) = n \log n - n + \log\sqrt{2\pi n} - \sum_{i=1}^{s}\left[\widehat{p}_i n \log(\widehat{p}_i n) - \widehat{p}_i n + \log\sqrt{2\pi n}\right]$$

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

$$= \log n - \frac{1}{n}\log\sqrt{2\pi n} - \sum_{i=1}^{s}\left[\widehat{p}_i \log(\widehat{p}_i) + \frac{1}{n}\log\sqrt{2\pi n}\right]$$

$$= -\sum_{i=1}^{s}\widehat{p}_i \log\widehat{p}_i - \frac{1}{n}\sum_{i=1}^{s}\log\sqrt{2\pi\widehat{p}_i n} + \frac{1}{n}\log\sqrt{2\pi n}$$

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,

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

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

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

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

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

1.3   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,

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

We can solve for λ with any numeric method.

1.3.1   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

$$H(p) = -\log p_x - 1 + \lambda_1 \mathcal{E}_1(x) + \cdots + \lambda_k \mathcal{E}_k(x) + \gamma = 0$$

$$\implies p_x = \frac{1}{Z} \exp\left[\lambda_1 \mathcal{E}_1(x) + \cdots + \lambda_k \mathcal{E}_k(x) \right]$$

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:

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

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

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

1.3.2   Large Deviation Principle

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

$$\mathbb{E}_{\widehat{p}} \mathcal{E}(X) = \sum_{x=1}^{s} \widehat{p}_x \mathcal{E}(x)$$

$$= \frac{1}{n} \sum_{x=1}^{s} \mathcal{E}(x) \sum_{i=1}^{n} \mathbb{1}_{X_i}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{x=1}^{s} \mathbb{1}_{X_i = x} \cdot \mathcal{E}(x)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathcal{E}(X_i) = \theta$$

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

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

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

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

1. How many ways can we sample 5 and 1 from X_1, X_2? Two ways: $(1, 5)$ or $(5, 1)$.

2. Now wat is the probability $X_1 = 1, X_2 = 5$? This is $p_1 p_5$. Similarly, $\mathbb{P}(X_1 = 5, X_2 = 1) = p_5 p_1$.

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

1.4   Jan 29

1.4.1   Relative Entropy Function

Motivation:

- p a PMF $\{1, \ldots, s\}$

- $\mathcal{E} : \{1, \ldots, s\} \to \mathbb{R}$ an energy function

- $X_1, X_2, \ldots, X_n \overset{iid}{\sim} p$

- $\widehat{p}$ 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.
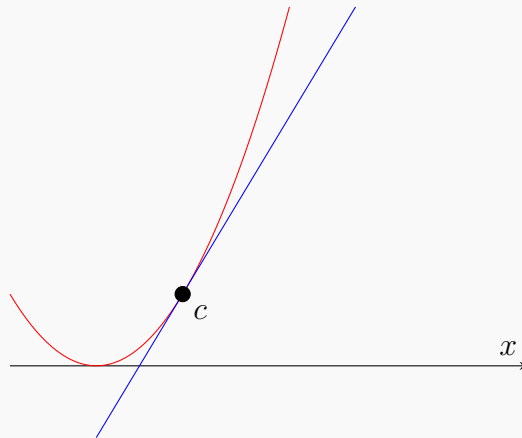
Hence,
$$\mathbb{E}g(X) \geq \mathbb{E}g'(c)(X - c) + \mathbb{E}g(c) = g'(c)(\mathbb{E}X - c) + g(c) = g(c) = g(\mathbb{E}X)$$

Properties of KL Entropy:

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

Proof:

1.

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

$$= \mathbb{E}_q \log \frac{q(X)}{p(X)}$$

$$= -\mathbb{E}_q \log \frac{p(X)}{q(X)}$$

$$= -\mathbb{E}_q \log Y$$

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

1.5   Jan 31

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

Proof:

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

$$X \sim q = \mathbb{E}[\log \frac{q_x}{p_x}] = -\mathbb{E}[\log \frac{p_x}{q_x}]$$

$$\overset{\text{Jensen}}{\geq} -\log[\mathbb{E}\frac{p_x}{q_x}]$$

$$= -\log[\sum q_x \frac{p_x}{q_x}] = 0$$

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.
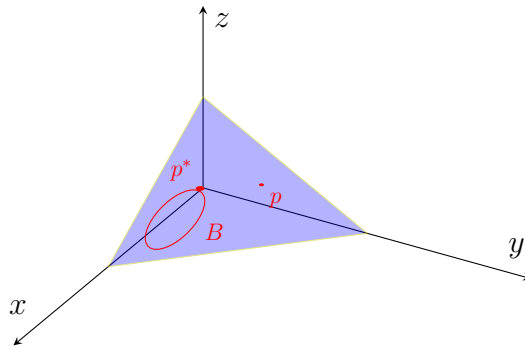
1.5.1   Sanov's Theorem

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

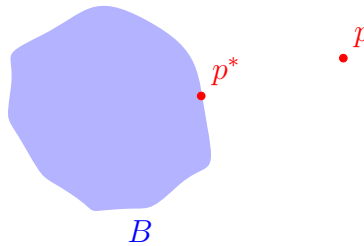$$\frac{1}{n} \log \underbrace{e^{-o(n)}}_{} \mathbb{P}(\widehat{p} \in B) = 0$$

1.6   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

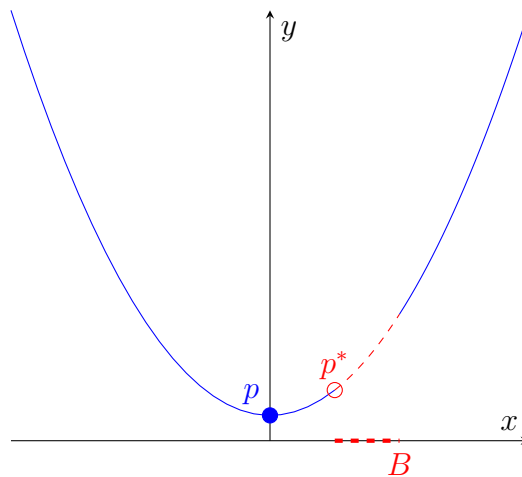$$\lim_{n\to\infty} \mathbb{P}(||\widehat{p} - p|| > \varepsilon \mid \widehat{p} \in B) = 0 \quad \forall \varepsilon > 0$$

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:

$$F(q) = D(q \parallel p) = \sum q_x \log \frac{p_x}{q_x}$$

$$= \sum q_x \log q_x - \sum q_x \log p_x$$

$$\frac{\partial F}{\partial q_x} = \log q_x + 1 - \log p_x$$

$$\frac{\partial^2 F}{\partial q_x \, \partial q_y} = \begin{cases} 1/q_x & x = y \\ 0 & x \neq y \end{cases}$$

$$H = \begin{pmatrix} \frac{1}{q_1} & & & \\ & \frac{1}{q_2} & & \\ & & \ddots & \\ & & & \frac{1}{q_s} \end{pmatrix}$$
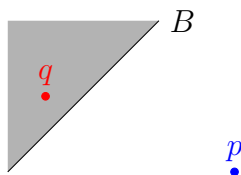
But $\forall v \in \mathbb{R}^s$, $v^T H v = \sum v_i^2 \frac{1}{q_i} \geq 0 \implies H$ is positive semi-definite. Hence F is convex.

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

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

1.7 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
$$\mathbb{E}_{p^*}[\mathcal{E}(X)] = \theta \iff \sum p_x^* \mathcal{E}(x) = 0 \iff \sum \frac{1}{Z_\lambda} p_x e^{\lambda \mathcal{E}(x)} \mathcal{E}(x) = 0$$

Proposition:

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

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

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.

$$\frac{\partial}{\partial \lambda_k} \log Z_\lambda = \frac{1}{Z_k} \cdot \frac{\partial}{\partial \lambda_k} Z_\lambda$$

$$= \frac{1}{Z_\lambda} \cdot \frac{\partial}{\partial \lambda_k} \left[\sum p_x e^{\lambda_1 \mathcal{E}_1(x) + \cdots + \lambda_c \mathcal{E}_c(x)} \right]$$

$$= \frac{1}{Z_\lambda} \cdot \sum_x p_x e^{\lambda_1 \mathcal{E}_1(x) + \cdots + \lambda_c \mathcal{E}_c(x)} \cdot \mathcal{E}_k(x)$$

$$= \frac{1}{Z_\lambda} \cdot \sum_x p_x \mathcal{E}_k(x) e^{\lambda \mathcal{E}(x)}$$

$$= \sum_x p_x^* \mathcal{E}_k(x)$$

$$= \mathbb{E}_{p^*}[\mathcal{E}_k(X)] = \mathbb{E}_\lambda[\mathcal{E}_k(X)]$$

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

<div style="border:2px solid red; padding:10px;">

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

</div>

2.

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 $\forall v \neq \vec{0}$,

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

But

$$\sum v_k v_l H_{kl} = \sum v_k v_l \mathrm{Cov}\left(\mathcal{E}_k(X), \mathcal{E}_l(X)\right)$$

$$= \mathrm{Var}\left(\sum v_k \mathcal{E}_k(X)\right) \geq 0$$

since

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

1.8   Feb 10

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

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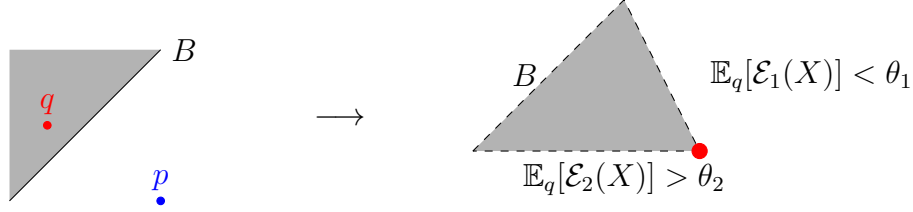- $\mathbb{E}_{\hat{p}}[\mathcal{E}_2(X)] = \theta_2$

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:

$$B_2 = \{q : \mathbb{E}_q[\mathcal{E}_2(X)] > \theta_2\}$$
$$\Omega_2 = \{q : \mathbb{E}_q[\mathcal{E}_2(X)] = \theta_2\} p_2^* \qquad\qquad = \underset{q \in \Omega_2}{\arg\min}\, D(q \parallel p)$$

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:
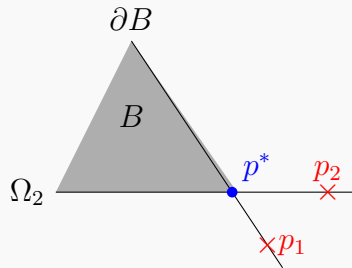
WLOG, $p^* \in \Omega_1$ so let $\tilde{p} = [p^*, p_1^*] \cap \Omega \implies \tilde{p} \in \Omega$.
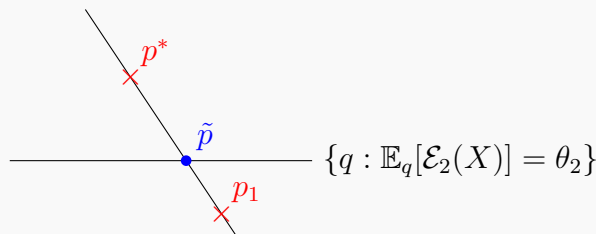
Then the $\tilde{p}$ should have been p^* (contradiction.)

1.8.1 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}.

1.9 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.
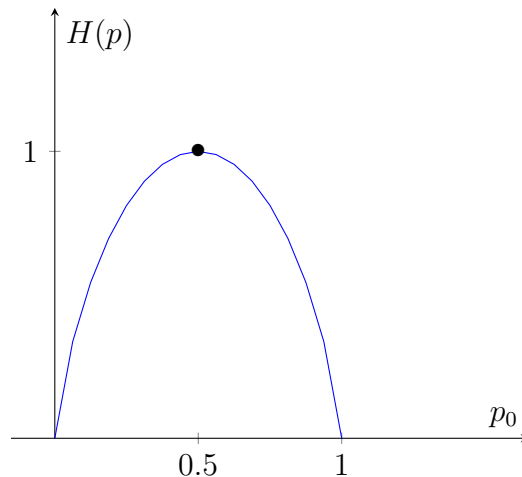
What is the relationship between the Shannon Entropy and the KL-Divergence?

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

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

1.9.1 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,

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

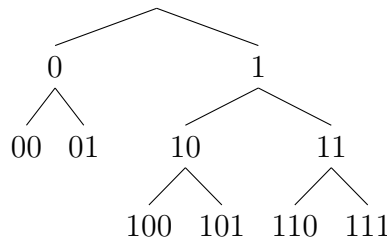and for any code lengths $\ell_1, \ldots, \ell_t$ such that

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

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

1.10   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 $n = 3$,

$X_1 X_2 X_3$	$C(X_1 X_2 X_3)$
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, $C(x) = x$, $\mathbb{E}_p\big[\big|C_{(X_1, \ldots, X_n)}\big|\big] = n$.

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,

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

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, $S \le 1$ so

$$= \sum_x p_x \log \frac{p_x}{q_x S} \tag{1.1}$$

$$= \sum_x p_x \log \frac{p_x}{q_x} - \sum_x p_x \log S \tag{1.2}$$

$$= D(p \parallel q) - \log S \ge 0 \tag{1.3}$$

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

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

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

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 $\sum_x 2^{-\ell_x} \le 1$.

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 $H(p) + 1$.

1.11   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:
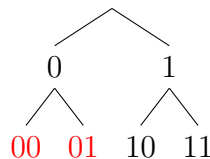
which naturally leads us to a full prefix code:

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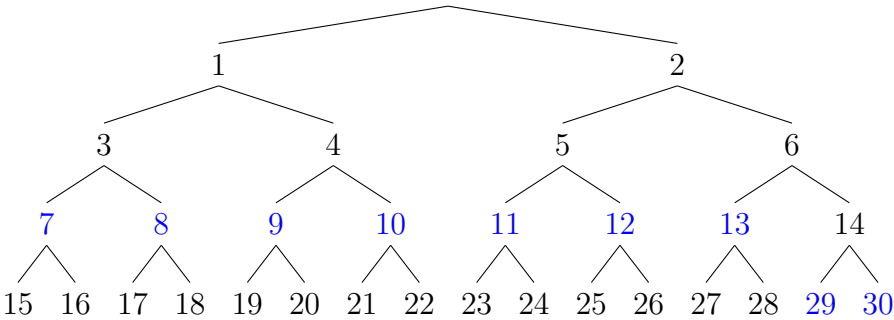
x	$p(x)$	ℓ_x	$C(x)$
11	1/3	4	000
12			001
13			$\vdots$
21			
22			
23			
31			110
32			1110
33			1111

which gives

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 2

Statistical Inference

2.1 Feb 19

2.1.1 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

$$\widehat{\mu} = \frac{1}{n} \sum X_i$$

$$\widehat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \widehat{\mu})^2 \mathbb{E}[\widehat{\sigma}^2] \qquad\qquad = \mathbb{E}\left[\frac{1}{n-1} \sum (X_i - \widehat{\mu})^2\right]$$

$$= \frac{1}{n-1} \sum \mathbb{E}[(X_i - \widehat{\mu})^2]$$

$$= \frac{n}{n-1} \mathbb{E}[(X_i - \widehat{\mu})^2]$$

$$= \frac{n}{n-1} \mathbb{E}\left[\left(X_i - \frac{X_1 + \cdots + X_n}{n}\right)^2\right]$$

$$= \mathbb{E}\left[\left(\frac{n-1}{n} X_1 - \frac{1}{n} X_2 \cdots - \frac{1}{n} X_n\right)^2\right]$$

$$= \mathbb{E}\left[\left(\frac{n-1}{n}\right)^2 X_1^2 + \sum_{i=2}^{n} \frac{1}{n^2} X_1^2 + 2 \sum_{i \neq j} X_i X_j\right]$$

$$= (\frac{n-1}{n})^2 \mathbb{E}[X_1^2] + \frac{n-1}{n^2} \mathbb{E}[X_1^2]$$

$$= \frac{(n-1)^2}{n^2} \sigma^2$$

$$= \frac{n-1}{n} \sigma^2$$

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,

- $\widehat{\theta}_n \xrightarrow{a.s.} \theta \implies \mathbb{P}(\lim_{n\to\infty} \widehat{\theta}_n = \theta) = 1$

- $\widehat{\theta}_n \xrightarrow{P} \theta \implies \forall \varepsilon > 0, \mathbb{P}(|\widehat{\theta}_n - \theta| > \varepsilon) \xrightarrow{n\to\infty} 0$

- $\widehat{\theta} \xrightarrow{\text{mean square}} \theta \implies \mathbb{E}[(\widehat{\theta}_n - \theta)^2] \to 0$.

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

2.2 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): $\text{MSE}(\widehat{\theta}_n) = \mathbb{E}\left|\widehat{\theta}_n - \theta\right|^2$.

Notice,

$$\text{MSE}\,(\hat\theta) = \mathbb{E}(\hat\theta_n - \theta)^2$$
$$= \mathbb{E}(\underbrace{\hat\theta_n - \mathbb{E}\hat\theta_n}_{a} + \underbrace{\mathbb{E}\hat\theta_n + \mathbb{E}\hat\theta_n - \theta}_{b})^2$$
$$= \mathbb{E}(a + b^2)$$
$$= \mathbb{E}a^2 + 2b\underbrace{\mathbb{E}a}_{0} + \underbrace{b^2}_{\text{bias}^2}$$
$$= \text{Var}\,(\hat\theta) + \text{bias}^2$$

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

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

We know

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

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 $X_i^4, X_i^3 X_j, X_i^2 X_j^2, X_i^2 X_j X_k, X_i X_j X_k X_l$ terms in the expansion.

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

After a little more work, we find

$$\text{MSE}\,(S_n^2) = \frac{2}{n-1}\sigma^4$$
$$\text{MSE}\,(\hat\sigma_n^2) = \frac{2n-1}{n^2}\sigma^4$$

but then MSE $(\hat\sigma_n^2)$ < MSE (S_n^2) so even though it is biased, it is a better estimator (in the sense of minimizing MSE).

2.2.1 Nonparametric Estimation

Example: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} p$. 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,

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

> **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 $p(B_n)$.

2.3 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?

2.3.1 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:

$$F_Y(x) = \mathbb{P}(Y \leq x) = int_{-\infty}^x k(t)\ dt$$

$$F_{\varepsilon Y}(x) = \mathbb{P}(\varepsilon Y \leq x) = \mathbb{P}\left(Y \leq \frac{x}{\varepsilon}\right) = \int_{-\infty}^{x/\varepsilon} k(s)\ ds$$

$$\overset{s=t/\varepsilon}{=\!=\!=} \int_{-\infty}^x k\left(\frac{t}{\varepsilon}\right) \frac{dt}{\varepsilon}$$

$$\implies k_\varepsilon(t) = \frac{1}{\varepsilon} k\left(\frac{t}{\varepsilon}\right)$$

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

$$k_w(x) = \frac{1}{w} k\left(\frac{x}{w}\right)$$
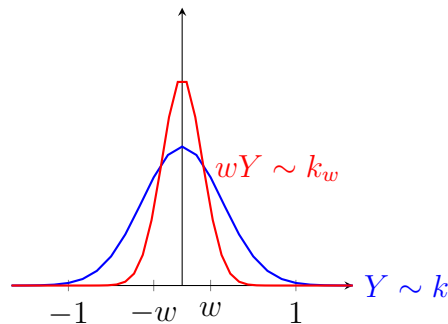
Now, our goal is to find the optimal w to approximate $Z(\sim \delta_0)$ by $Z + Y_w$.

Correspondingly, we approximate $f(x)$ by

$$\widehat{f}(x) = \widehat{f}_{n,w}(x, X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^n k_w(x - X_i)$$
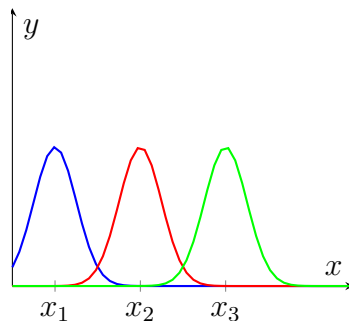
Our plan is to use MSE $=$ bias^2 + variance as

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

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):

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

$$= \int_{\mathbb{R}} \mathbb{E}\left|\widehat{f}(x, X_{1:n}) - \mathbb{E}[\widehat{f}_n(x, X_{1:n})] + \mathbb{E}[\widehat{f}_n(x, X_{1:n})] - f(x)\right|^2 \, dx$$

$$= \int_{\mathbb{R}} \left|\mathbb{E}\widehat{f}_n(x, X_{1:n}) - f(x)\right|^2 \, dx + \int_{\mathbb{R}} \mathbb{E}\left|\widehat{f}_n(x, X_{1:n}) - \mathbb{E}[\widehat{f}_n(x, X_{1:n})]\right|^2 \, dx$$

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

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

$$B_{n,w}(x) = \mathbb{E}[\widehat{f}_n(x, X_1, \ldots, X_n)] - f(x)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[k_w(x - X_i)] - f(x)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} f(t) k_w(x - t) \, dt - f(x)$$

$$= \int_{\mathbb{R}} f(t) k_w(x - t) \, dt - f(x)$$

2.4   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

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

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)

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

where

$$\text{bias(x)} = \mathbb{E}[\widehat{f}(x)] - f(x)$$

$$\mathbb{E}[\widehat{f}(x)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{X_i}[k_w(x - X_i)]$$

$$= \mathbb{E}[k_w(x - X_1)] \qquad (X_i \overset{\text{iid}}{\sim} f)$$

$$= \int_{\mathbb{R}} f(t) k_w(x - t) \, dt$$

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:

$$\text{Var}(x) = \text{Var}(\widehat{f}(x))$$

$$= \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} k_w(x - X_i)\right)$$

$$= \frac{1}{n^2} \sum \text{Var}(k_w(x - X_i)) \qquad \text{(independence)}$$

$$= \frac{1}{n} \text{Var}(k_w(x - X_i)) \qquad \text{(identically distributed)}$$

$$= \underbrace{\frac{1}{n} \mathbb{E}[(k_w(x - X_i))^2]}_{V^{(1)}} - \underbrace{\frac{1}{n} [\mathbb{E}[k_w(x - x_i)]]}_{V^{(2)}}^{2}$$

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

$$V^{(1)} = \frac{1}{n} \int f(g) k_w^2 (x - t) \, dt$$

$$= \frac{1}{n} \frac{1}{w} \int f(t) \frac{1}{w} k^2 \left(\frac{x - t}{w} \right) dt$$

$$= \frac{1}{n} \frac{1}{w} \int f(ws + t) k^2(s) \, ds \qquad (s = \frac{x - t}{w})$$

$$\to \begin{cases} \infty & w \to 0 \\ 0 & w \to \infty \end{cases}$$

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

2.5   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:

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

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:

$$\text{ISE} = \int_x \left| \widehat{f}_{n,w}(x) - f(x) \right|^2 dx$$

$$= \int \widehat{f}_{n,w}(x) - 2 \int \widehat{f} \cdot f + \int f^2(x) \, dx$$

Define

$$I = \int_x \widehat{f}_{n,w}(x) \cdot f(x) \, dx$$

$$= \mathbb{E}_{X_{n+1} \sim f}[\widehat{f}_{n,w}(X_{n+1})] \qquad (X_{1:n} \overset{\text{iid}}{\sim} f)$$

$$= \mathbb{E}[\widehat{f}_{n,w}(X_{n+1}; X_1, \ldots, X_n)]$$

$$\approx \mathbb{E}[\widehat{f}_{n-1,w}(X_n; X_1, \ldots, X_{n-1})]$$

$$\approx \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\widehat{f}_{n-1,w}(X_i; i^X)] \qquad (i^X = X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

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

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

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

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

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

(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

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

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

which is very fast growth

2.6 March 3

2.6.1 Maximum Likelihood Estimation

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

$$\widehat{\theta} = \arg\max_{\widetilde{\theta}} p_{\widetilde{\theta}}(X_1 = x_1, \ldots, X_n = x_n)$$

$$= \arg\max_{\widetilde{\theta}} p_{\widetilde{\theta}}(x_1) \cdots p_{\widetilde{\theta}}(x_n)$$

$$= \arg\max_{\widetilde{\theta}} \frac{1}{n} \sum_{i=1}^{n} \log p_{\widetilde{\theta}}(x_i)$$

$$= \arg\max_{\widetilde{\theta}} \sum_{x=1}^{s} (\log p_{\widetilde{\theta}}(x)) \widehat{p}(x)$$

$$= \arg\min_{\widetilde{\theta}} - \sum_{x=1}^{s} \widehat{p}(x) \log p_{\widetilde{\theta}}(x)$$

$$= \arg\min_{\widetilde{\theta}} - \sum_{x=1}^{s} \widehat{p}(x) \log \frac{\widehat{p}(x)}{p_{\widetilde{\theta}}(x)} - \sum \widehat{p}(x) \log \widehat{p}(x)$$

$$= \arg\min_{\widetilde{\theta}} D(\widehat{p} \parallel p_{\widetilde{\theta}}) + H(\widehat{p})$$

$$= \arg\min_{\widetilde{\theta}} D(\widehat{p} \parallel p_{\widetilde{\theta}})$$

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,

$$\widehat{\mu} = \arg\max_{\widetilde{\mu}} \prod_{i=1}^{n} f_\mu(x_i)$$

$$= \arg\max \exp\left(\sum -\frac{(x_i - \widetilde{\mu})^2}{2}\right)$$

$$= \arg\min \sum_{i=1}^{n} (x_i - \widetilde{\mu})^2$$

$$\overset{*}{=} \frac{1}{n} \sum_{i=1}^{n} x_i$$

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

$$\widehat{\lambda} = \arg\max \prod f_{\tilde{\lambda}}(x_i)$$

$$= \arg\min -\frac{1}{n}\sum_{i=1}^{n}\log(f_{\tilde{\lambda}}(x_i)) \qquad \text{(invariant under constant)}$$

$$= \arg\min -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{Z_\lambda}p(x_i)\exp\left(\sum_{j=1}^{c}\lambda_j\mathcal{E}_j(x)\right)\right)$$

$$= \arg\min -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{Z_\lambda}\exp\left(\sum_{j=1}^{c}\lambda_j\mathcal{E}_j(x)\right)\right) \qquad (p(x_i) \text{ known})$$

$$= \arg\min \log(Z_{\tilde{\lambda}}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\tilde{\lambda}_j\mathcal{E}_j(x_i)$$

$$= \arg\min \log(Z_{\tilde{\lambda}}) - \sum_{j=1}^{c}\lambda_j\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_j(x_i)\right)$$

$$= \arg\min \log(Z_{\tilde{\lambda}}) - \sum_{j=1}^{c}\lambda_j\theta_j$$

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

2.6.2 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
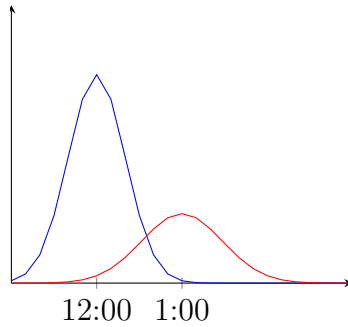
If we interviewed a student at 2:00, it is quite likely they will be from group two. Similarly, if we interviewed a student at 10:00, they are likely from group one. But what about at 12:30? This is the problem of classification.

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

2.7 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?

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

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 $\mathrm{FAR}_h \leq \mathrm{FAR}_{h_t}$, then $\mathrm{DR}_h \leq \mathrm{DR}_{h_t}$

Intuition:

We have

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

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

Under NP classification,
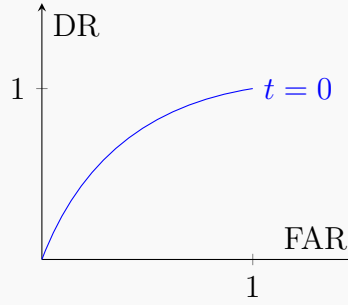
Proof:
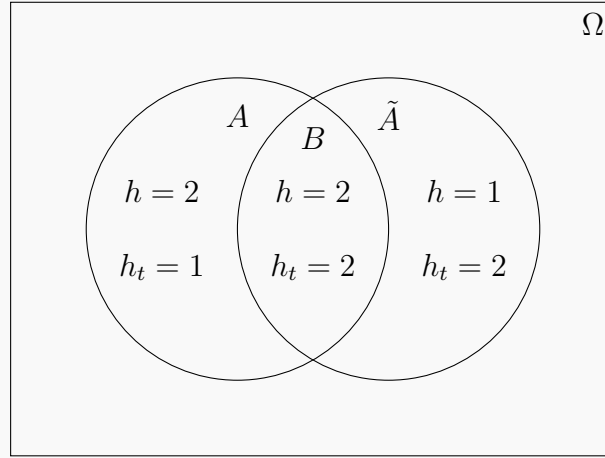
We have

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

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

$$A \subseteq \{h_t(x) = 1\} = \left\{ \frac{f_2(x)}{f_1(x)} < t \right\} = \{f_2(X) < t \cdot f_1(X)\}$$

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

We have

$$p_1(X \in A) \leq p_1(X \in \tilde{A})$$

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

2.8 March 7

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

There are multiple different approaces:

- Generative

- Discriminative

- Algorithmic

2.8.1 Generative Classifiers

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

A good estimator given data $(X_1, Y_1) \ldots (X_n, Y_n)$ 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 $X = (X^1, X^2, \ldots, X^d)$ and $f_c(x^1, \ldots, x^d) = f_c^1(x^1) \cdots f_c^d(x^d)$.

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 $\{1, \ldots, s\}$. In practice, this can lead to over fitting.

2.9 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.
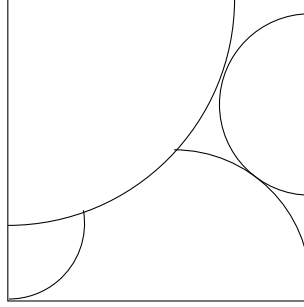
We can follow a similar but slightly less flexible approach.

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.

2.9.1 Discriminative Construction

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

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

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.

2.10 March 14

2.10.1 Discriminative Classification Continued

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.
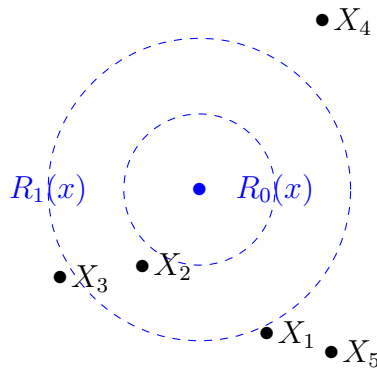
If we fix k, as $n \to \infty$, $R_k(x) \to 0$ which gives us the estimator

$$\widehat{r}_c(x) = \frac{\#\{i : X_i \in B(x, R_k(x)), \ Y_i = c\}}{k}$$

Claim: $\widehat{r}_c(x) \to r_c(x)$, i.e. the K-NN estimator is consistent

Proof:

$$
\begin{aligned}
\widehat{r}_c(x) &= \frac{\#\{i : X_i \in B(x, R_k(x)), Y_i = c\}}{\#\{i : X_i \in B(c, R_k(x))\}} \\
&= \frac{\#\{i : X_i \in B(x, \varepsilon), Y_i = c\}/n}{\#\{i : X_i \in B(c, \varepsilon)\}/n} \qquad (\varepsilon = R_k(x)) \\
&\xrightarrow{n \to \infty} \frac{\mathbb{P}(X \in B(x, \varepsilon),\ Y = c)}{\mathbb{P}(X \in B(x, \varepsilon))} \\
&\overset{*}{=} \frac{\mathbb{P}(Y = c)\mathbb{P}(X \in B(x, \varepsilon) \mid Y = c)}{\mathbb{P}(X \in B(x, \varepsilon))} \\
&= \mathbb{P}(Y = c)\frac{\int_{B(x, \varepsilon)} f_c(t)\, dt}{\int_{B(x, \varepsilon)} f(t)\, dt} \\
&\approx \mathbb{P}(Y = c)\frac{f_c(x)\,|B(x, \varepsilon)|}{f(x)\,|B(x, \varepsilon)|} \\
&= \mathbb{P}(Y = c)\frac{f_c(x)}{f(x)} \\
&= \frac{\mathbb{P}(Y = c)\mathbb{P}(X = x \mid Y = c)}{P(X = x)} \\
&= \frac{\mathbb{P}(X = x,\ Y = c)}{\mathbb{P}(X = x)} \\
&= \mathbb{P}(Y = c \mid X = x) \\
&= r_c(x)
\end{aligned}
$$

Remark: This is not entirely rigorous since we are taking both $n \to \infty$ and $\varepsilon \to 0$. It would not be too difficult to make this rigorous by estimating the error of the limit $n \to \infty$. Leaving the way open to rigor is also why we cannot say

$$
\frac{\mathbb{P}(Y = c)\mathbb{P}(X \in B(x, \varepsilon) \mid Y = c)}{\mathbb{P}(X \in B(x, \varepsilon))} = \mathbb{P}(Y = c \mid X \in B(x, \varepsilon)) \xrightarrow{\varepsilon \to 0} \mathbb{P}(Y = c \mid X = x) = r_c(x)
$$

in the starred equality.
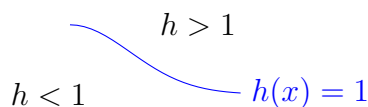
Support Vector Machine:

$$
h^*(x) = \arg\max_c r_c(x)
$$

Example: In the case $s = 2$, we want r_1, r_2 which gives

$$
h(x) = \frac{r_2(x)}{r_1(x)}
$$

so it suffices to find the precision boundary

The trick here is to find a transformation ϕ so that the precision boundary is linear

$$\phi$$
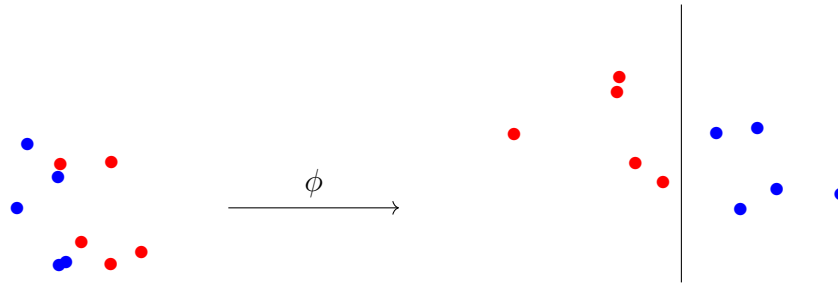
> **Bonus 3/14:** What is the setup for the classification problem? What does it mean to build a classifier?
>
> ---
>
> **Answer:** Given a set of data $\{(X_i, Y_i)\}_{i=1}^n$ on classes $Y_i \in \{1, \ldots, s\}$ following an unknown distribution f_c for $c \in \{1, \ldots, s\}$, we want to find a function $h : \mathbb{R}^d \to \{1, \ldots, s\}$ that best approximates the true distribution.
>
> Bayes' Rule suggests that the optimal behavior is
>
> $$h^*(x) = \arg\max_c \mathbb{P}(Y = c \mid Y = x)$$
>
> Everything else is finding different ways to estimate this function.

2.11 March 17

2.11.1 Support Vector Machine

Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we want to find a function $h : \mathbb{R}^d \to \{1, \ldots, s\}$ that best approximates the true distribution.

In $\mathbb{R}^d$, there exists a curve that partitions the data into regions where $h(x) = c$ for some $c \in \{1, \ldots, s\}$. We want to find this curve. The idea behind support vector machines is to apply a transformation ϕ such that it suffices to choose a hyperplane (rather than a curve) in $\mathbb{R}^{d'}$ with $d' \gg d$.

Example $(d = 1)$:
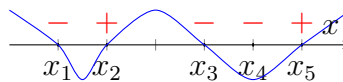
with classifier

$$
\begin{aligned}
h(x) &= \mathrm{sign}((x - X_1)(x - X_2)(x - X_3)(x - X_5)) \\
&= \mathrm{sign}(\alpha_4 x^4 + \alpha_3 x^3 + \alpha_2 x^2 + \alpha_1 x + \alpha_0)
\end{aligned}
$$

so for $\phi : \mathbb{R} \to \mathbb{R}^5$ by $x \mapsto (1, x, x^2, x^3, x^4)$, we have

$$h(x) = \mathrm{sign}(\vec{a} \cdot \phi(x))$$

However this presents a few problems: it depends on raising the dimension (something we are always looking to avoid), can lead to overfitting, and presents ambiguities – if there are multiple such hyperplanes, which do we choose?

We will spend some time on this last question.

Consider the case
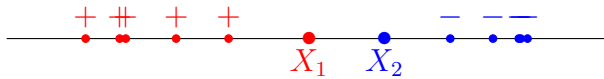
Intuitively, we would like our classifier not to be too close to the data points nor to favor one class over the other.

In this case, $\frac{X_1+X_2}{2}$ is a good choice which would give us classifier $h(x) = \text{sign}(\frac{X_1+X_2}{2} - x)$.

In the general case, we would like to find the hyperplane that *maximizes the minimal distance to the precision boundary.* We call the points closest to the hyperplane the *support vectors* (in our example, X_1, X_2)

Formally, we want to find the hyperplane $0 = \alpha + \vec{\beta} \cdot x$ for $\alpha \in \mathbb{R}$, $\vec{\beta}, x \in \mathbb{R}^d$ which maximizes the minimal distance to the support vectors, yielding classifier

$$h(x) = \begin{cases} + & \text{if } \alpha + \beta x > 0 \\ - & \text{if } \alpha + \beta x < 0 \end{cases} = \text{sign}(\alpha + \beta x)$$

following the convention that the normal vector β is always chosen in the direction of the positive class.

What is the distance between a point $x \in \mathbb{R}^d$ to the hyperplane $\alpha + \beta x = 0$? Precisely $||\alpha + \beta x||$.

Hence, the task moving forward is to find

$$(\alpha, \beta) = \arg\max \left(\min_{i=1:n} |\alpha + \beta x_i| \right)$$

2.12 March 19

Recall: We were looking for *maximum margin classifiers* for Support Vector Machines: On data $(X_1, Y_1), \ldots, (X_n, Y_n)$ find

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\max_{\alpha, \beta} \min_{i=1:n} \text{dist}(X_i, \{\alpha + \beta x = 0\})$$

for all $i : (\alpha + \beta X_i)Y_i \geq 0$.

Notice, the condition $(\alpha + \beta X_i)Y_i \geq 0$ is equivalent to the condition that the data is separable by the hyperplane $\alpha + \beta x = 0$.

Equivalently,

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\max_{\substack{\alpha, \beta \\ ||\beta||=1 \\ \forall i: (\alpha+\beta X_i)Y_i \geq 0}} \min_{i=1:n} (\alpha + \beta X_i)Y_i = \arg\max_{\alpha, \beta, ||\beta||=1} M$$

subject to $Y_i(\alpha + \beta X_i) \geq M$ for all M since

$$\min\{a_1, \ldots, a_n\} = \max M \quad a_i \geq M \; \forall M$$

Is this a convex optimazation problem, i.e are we optimizing a convex function on a convex set?

Already we run into difficulty with the condition $||B|| = 1$ since
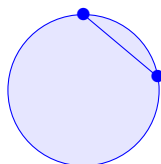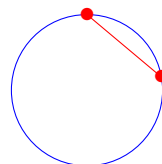
$||B|| \leq 1$ convex $||B|| = 1$ not convex

Hence, we would like to reformulate the problem to remove the constraint $||B|| = 1$.

Consider the adjusted constraint $Y_i(\tilde{\alpha}, \tilde{\beta}) = Y_i(\frac{\alpha}{M} + \frac{\beta}{M}X_i) \geq 1$ for all i. T

Now

$$\left\|\tilde{B}\right\| = \sum_{i=1}^{d} \tilde{\beta}_i^2 = \frac{1}{M^2}$$

so we can write

$$(\tilde{\alpha}, \tilde{\beta}) = \arg\max_{\tilde{\alpha},\tilde{\beta}} \frac{1}{\sum_{i=1}^{d} \tilde{\beta}_i^2} = \arg\min_{\tilde{\alpha},\tilde{\beta}} \sum_{i=1}^{d} \tilde{\beta}_i^2$$

subject to $Y_i(\tilde{\alpha} + \tilde{\beta}X_i) \geq 1 \quad \forall i$

And indeed this is a convex optimization problem since $(\alpha, \beta) \mapsto \sum_{i=1}^{d} \beta_i^2$ is convex and the domain $B = \{(a, b) : Y_i(a + bX_i) \geq 1\}$ is convex

> *Proof:* Pick $(\alpha, \beta) \in B$ and $(\alpha', \beta') \in B$. Then for all $\lambda \in (0, 1)$,
>
> $$(\lambda a + (1 - \lambda)a', \lambda\beta + (1 - \lambda)\beta') \in B$$
>
> since by assumption,
>
> $$(\alpha, \beta) \in B \implies Y_i(\alpha + \beta X_i) \geq 1 \implies \lambda Y_i(\alpha + \beta X_i) \geq \lambda$$
> $$(\alpha', \beta') \in B \implies Y_i(\alpha' + \beta'X_i) \geq 1 \implies (1 - \lambda)Y_i(\alpha' + \beta'X_i) \geq 1 - \lambda$$
>
> $$\lambda Y_i(\alpha + \beta X_i) \geq \lambda$$
> $$\frac{(1 - \lambda)Y_i(\alpha' + \beta'X_i) \geq 1 - \lambda}{\lambda Y_i(\alpha + \beta X_i) + (1 - \lambda)Y_i(\alpha' + \beta'X_i) \geq 1}$$
>
> so indeed,
>
> $$Y_i(\alpha'' + \beta''X_i) \geq 1$$

Recall in the past to solve $\hat{\theta} = \arg\min f(\theta)$ subject to $g(\theta) = 0$, we have used Lagrange multipliers:

$$\arg\min f(\theta) + \lambda g(\theta)$$

For convex optimization problems, we introduce the **Lagrange Dual**:

1. Find $\lambda = (\lambda_1, \ldots, \lambda_n)$ subject to $\lambda_i \geq 0$
2. Find $\hat{\alpha} = \hat{\alpha}(\lambda)$, $\hat{\beta} = \hat{\beta}(\lambda)$ that minimize

$$\frac{1}{2}\sum_{j=1}^{d} \hat{\beta}_j^2 + \sum_{i=1}^{n} \lambda(1 - Y_i(\alpha + \beta X_i)) \qquad (\star)$$

3. Maximize $(\star)$ over λ to get $\hat{\alpha}, \beta$

Chapter 3

Graphical Models

3.1 March 31

3.1.1 Gibbs Random Fields

Definition: Let $G = (V, E)$ be a graph. Then we say a subset $C \subseteq V$ is a *clique* if $\forall i \neq j \in C$, $(i, j) \in E$.

Example: Let G_1 be given by

$$1 \overline{\hspace{1.5cm}} 2$$
$$|$$
$$3$$
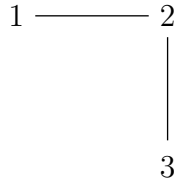
Then the cliques are $C_i = \{i\}$, $C_{12} = \{1, 2\}$, $C_{23} = \{2, 3\}$ but $C_{123} = \{1, 2, 3\}$ is not a clique.

Definition: A *Gibbs random field (GRF)* on a graph G is a set of random variables $\{X_v\}_{v \in V}$ with the distribution

$$p(x) = \frac{1}{Z} \prod_{C \text{ sums over cliques in } G} \phi_c(x_c)$$

for some function $\phi_c : \Omega_c \to [0, \infty)$ clique functions and Z a partition function.

Example:

$$3 \overline{\hspace{1.5cm}} 4$$
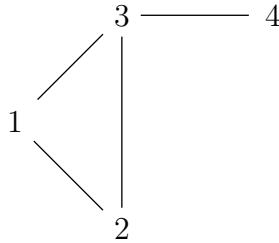
So for $C = \{1, 2, 3\}$ we have $X_C = (X_1, X_2, X_3)$ while for $C' = \{1, 2\}$ we have $X_{C'} = (X_1, X_2)$.

Hence $(X_1, \ldots, X_4)$ is a GRF on G if

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \phi_1(x_1) \phi_2(x_2) \phi_3(x_3) \phi_{12}(x_1, x_2) \phi_{13}(x_1, x_3) \phi_{23}(x_2, x_3) \phi_{34}(x_3, x_4) \phi_{123}(x_1, x_2, x_3)$$

Example: Let

- $X_1 \sim \text{Bernoulli}(1/2)$

- $X_2 \sim \text{Bernoulli}(1/2), \ X_2 \perp X_1$

- $X_3 = X_1 \oplus X_2 = \begin{cases} 0 & X_1 = X_2 \\ 1 & X_1 \neq X_2 \end{cases}$

- $X_4 = 1 - X_3$

$$p(1,0,1,0) = p(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0) = p(X_1 = 1, X_2 = 0) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

and

$$p(1,0,0,0) = 0$$

What graph does this respect?

$$p(x_1, x_2, x_3, x_4) = \frac{1}{2} \cdot \frac{1}{2} \cdot \mathbb{1}_{x_3 = x_1 \oplus x_2} \cdot \mathbb{1}_{x_4 = 1 - x_3} = \frac{1}{Z} \phi_{123}(x_1, x_2, x_3) \phi_{34}(x_3, x_4)$$

Hence this is a GRF with respect to any graph that contains
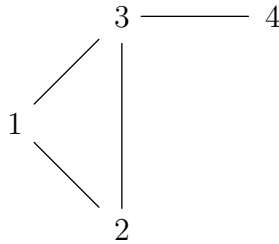
For example:
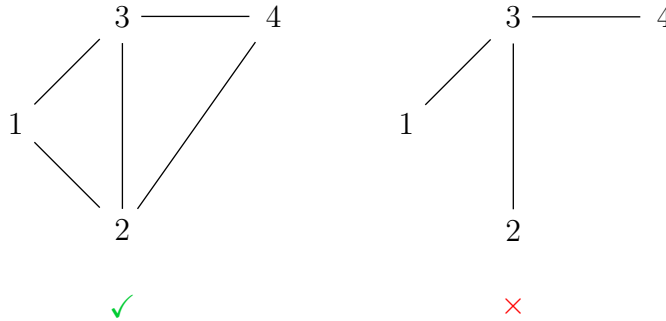
What would a GRF on the second graph look like?

We know we must have a density of the form

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \phi_1(x_1) \cdots \phi_4(x_4) \phi_{13}(x_1, x_3) \phi_{23}(x_2, x_3) \phi_{34}(x_3, x_4)$$

but we cannot find functions $\phi_{13}(x_1, x_3)\phi_{23}(x_2, x_3)$.

Remark: Neither the representation $p(x) = \frac{1}{Z} \prod_{C \text{ sums over cliques}} \phi_c(x_c)$ nor the graph representing a particular distribution must not be unique.

Example:

1. Instead of

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \frac{1}{4} \mathbb{1}_{x_3 = x_1 \oplus x_3} \cdot \mathbb{1}_{x_4 = 1 - x_3}$$

   we could write

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \frac{1}{4} \mathbb{1}_{x_3 = x_1 \oplus x_2} \cdot \mathbb{1}_{x_4 = 1 - x_1 \oplus x_2}$$

   which would instead yield the graph
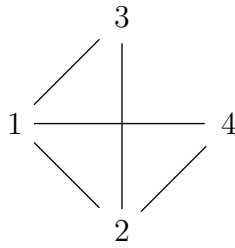
2. Working with our same example
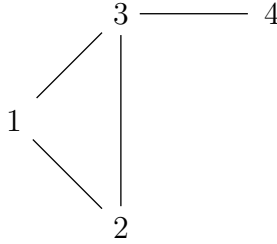
suppose

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \left| \cos(x_1 + x_2 + x_2) \right| x_2^{x_3}$$

but this could be parameterized $\frac{1}{Z} \phi_{123}(x_1, x_2, x_3) \phi_{23}(x_2, x_3)$ or $\frac{1}{Z} \phi_{123}(x_1, x_2, x_3)$.

Definition: If $\phi_c > 0$ for all cliques c, we say the GRF is *strictly positive*. Equivalently, $\forall x_1, \ldots, x_M, p(x_1, \ldots, x_M) > 0$

As we will see, this is exactly the condition for a GRF to be a Markov random field.

Note that GRFs are incredibly general. For any distribution $p(x)$, $x \in \mathbb{R}^n$, $p(x)$ will be a GRF on the complete graph K_n with n vertices.

Example: What is the minimal graph for a GRF on $X_1, \ldots, X_n$ independent? Trivially,

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

so the graph is $G = (V, E) = (\{1, \ldots, n\}, \emptyset)$.

Example (Markov Chain): A Markov chain satisfies

$$p(x_1, \ldots, x_n) = p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_2) \cdots p(x_n \mid x_{n-1})$$

Its minimal graph is the chain

$$1 \overline{\qquad} 2 \overline{\qquad} 3 \overline{\qquad} \cdots \overline{\qquad} n$$

3.2   April 2

Recall: $(X_1, \ldots, X_n)$ is a *Gibbs Random Field* wrt a graph F if

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{\text{cliques } c \subseteq G} \phi_c(x_c)$$

for clique functions ϕ_c and Z a partition function, i.e. if $p(x_1, \ldots, x_n)$ factors on G.

Example (Hidden Markov Model): Suppose we know our friend's behavior very well. Based on *observations* (e.g. they choose to go for a walk, shop, or clean) we can guess a hidden state unknown to us (e.g. it is sunny or it is rainy).

Suppose the transition probability is given by

$$
\begin{array}{cc}
 & \begin{array}{cc}\text{Sunny} & \text{Rainy}\end{array}\\
\begin{array}{c}\text{Sunny}\\ \text{Rainy}\end{array} & \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}
\end{array}
$$

Then $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is a Markov Chain with with emission probabilities

$\mathbb{P}(Y_i = y_i \mid X_i = x_i)$	Walk	Shop	Clean
Sunny	0.6	0.3	0.1
Rainy	0.1	0.4	0.5

Further, $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ is a GRF wrt G.
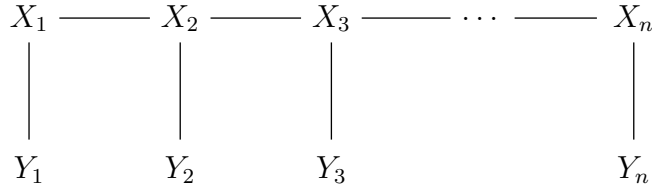
If we assume $\mathbb{P}(X_1) = (0.2, 0.8)$,

$$
\begin{aligned}
\mathbb{P}(\text{Sunny}, \text{Rainy}, \text{Walk}, \text{Shop}) &= \mathbb{P}(X_1 = \text{Sunny}, X_2 = \text{Rainy}, Y_1 = \text{Walk}, Y_2 = \text{Shop})\\
&= \mathbb{P}(X_1 = \text{Sunny}) \cdot \mathbb{P}(X_2 = \text{Rain} \mid X_1 = \text{Sunny}) \cdot \mathbb{P}(Y_1 = \text{Walk} \mid X_1 = \text{Sunny}) \cdot \mathbb{P}(Y_2 = \text{S}\\
&= 0.2 \cdot 0.3 \cdot 0.6 \cdot 0.4\\
&= 0.0144
\end{aligned}
$$

From which we deduce

$$
p(x_1, \ldots, x_n, y_1, \ldots, y_n) = p(x_1) \cdot p(x_2 \mid x_1) \cdots p(x_{n-1} \mid x_n) \cdot p(y_1 \mid x_1) \cdots p(y_n \mid x_n)
$$

Example: Let $X_1, X_2 \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$ and $X_3 = X_1 \oplus X_2 = \begin{cases} 0 & X_1 = X_2 \\ 1 & X_1 \neq X_2 \end{cases}$.

Is (X_1, X_2, X_3) a GRF wrt the following?
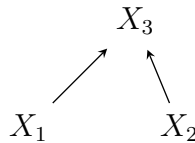
We know it respects G iff

$$
p(x_1, x_2, x_3) = \frac{1}{Z} f(x_1, x_3) \cdot g(x_2, x_3)
$$

We can calculate:

$$
\begin{aligned}
p(x_1, x_2, x_3) &= \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)\\
&= \frac{1}{2} \cdot \frac{1}{2} \mathbb{P}(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2)\\
&= \frac{1}{4} \mathbb{1}_{x_3 = x_1 \oplus x_2}
\end{aligned}
$$

Assume $\mathbb{1}_{x_3 = x_1 \oplus x_2} = f(x_1, x_2) \cdot g(x_2, x_3)$.

But if $g(x_2, x_3) = 0$, then $\mathbb{1}_{x_3 = x_1 \oplus x_2} = 0$ regardless of x_1. But by definition, for any value of $x_3 = x_1 \oplus x_2$, there exists a value of x_1 such that $\mathbb{1}_{x_3 = x_1 \oplus x_2} = 1$. Contradiction. In particular, this means that the smallest possible graph for (X_1, X_2, X_3) is the complete graph K_3.

Example (Multinomial): Roll a die 5 times giving $X_1, \ldots, X_5$, the number of times we get each face. We know that $X_1 + \cdots + X_6 = 5$

Which graph does this respect?

$$p(x_1, \ldots, x_6) = \frac{5!}{x_1! \cdots x_6!} \cdot \frac{1}{6^5} \cdot \mathbb{1}_{x_1 + \cdots + x_6 = 5}$$

As before, we cannot factor this indicator, so the minimal graph is K_6.

Example (Naive Bayes): Run d tests to classify $Y \in \{1, \ldots, s\}$ giving data $(X_1, \ldots, X_d, Y)$.

$$f(x_1, \ldots, x_d, y) = p(y)p(x_1, \ldots, x_d \mid y) \overset{NB}{=} p(y) \prod_{i=1}^{d} f_y(x_i)$$

Exercise: What is the minimal graph for a Naive Bayes model?

Solution: By the Naive Bayes' assumption, each test is independent and only depends on the class.
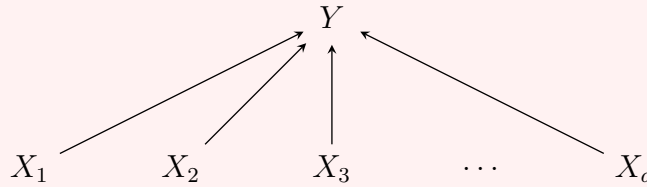
3.3 April 4

3.3.1 Independence

Suppose $(X_v)_{v \in V(G)}$ is GRF wrt G where G is partitioned into two disjoint sets A and B ($V(G) = A \cup B$) with no edges between them
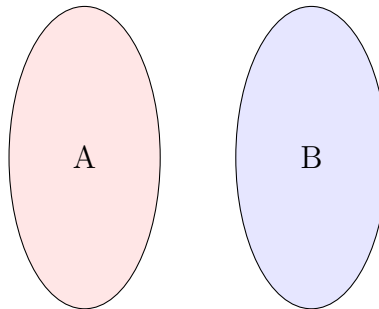
then $(X_v)_{v \in A} \perp (X_u)_{u \in B}$ so we say $(X_v)_{v \in A}$ and $(X_u)_{u \in B}$ are *independent*.

Example: Let $X_1, \ldots, X_5$ respect
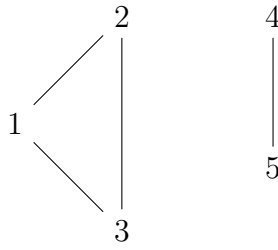
Claim: $(X_1, X_2, X_3) \perp (X_4, X_5)$

Proof: $x_A \perp x_B$ iff $p(x_A, x_B) = p(x_A)p(x_B)$.

We have

$$p(x_A, x_B) = \frac{1}{Z} \prod_{c \text{ cliques}} f_c(x_c)$$

by definition of GRF.

Notice that any clique in A must be entirely contained in A (and similarly for B). Hence,

$$p(x_A, x_B) = \frac{1}{Z} \prod_{\substack{c \text{ cliques} \\ c \subseteq A}} f_c(x_c) \cdot \prod_{\substack{c \text{ cliques} \\ c \subseteq B}} f_c(x_c z)$$

$$= \frac{1}{Z} f_A(x_A) \cdot f_B(x_B)$$

But since Z is a normalization factor,

$$Z = \sum_{x_A} \sum_{x_B} f_A(x_A) f_B(x_B) = \sum_{x_A} f_A(x_A) \left[\sum_{x_B} f_B(x_B) \right] = \left[\sum_{x_A} f_A(x_A) \right] \left[\sum_{x_B} f_B(x_B) \right] = Z_A \cdot Z_B$$

So

$$p(x_A, x_B) = \left[\frac{1}{Z} f_A(x_A) \right] \left[\frac{1}{Z_B} f_B(x_B) \right] = p_A(x_A) \cdot p_B(x_B)$$

for pmfs p_A, p_B.

Example: Consider a case with three sets A, B, D with no edges between A, B but edges between A, D and B, D. Now is it true that $A \perp B$?
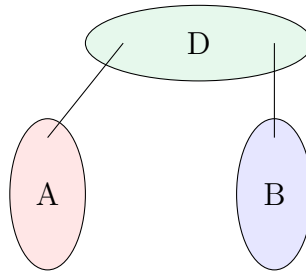
Consider a simple case:

$$1 \;\;\rule{1cm}{0.4pt}\;\; 2 \;\;\rule{1cm}{0.4pt}\;\; 3$$
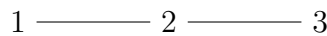
If this is a Markov chain, then there are no edges between X_1 and X_2 but there is certainly still a dependence between them.

Is there a sufficient condition to conclude $A \perp B$? Clearly, we just need there to not exist a path between A and B.

Proof: Suppose there is no path between A and B.

Let

$$\tilde{A} = \{v : \exists \text{ path from } v \text{ to } A\}$$
$$\tilde{B} = V \setminus \tilde{A}$$

Clearly, $\tilde{A} \cap \tilde{B} = \emptyset$ and $A \subseteq \tilde{A}$, $B \subseteq \tilde{B}$ since they share no edges.

But $X_{\tilde{A}} \perp X_{\tilde{B}} \implies X_A \perp X_B$.

To summarize:

- No edges $\;\not\!\!\!\implies\;$ independence
- Independence $\;\not\!\!\!\implies\;$ no edges (even on a minimal graph!)
- No path $\implies$ independence

3.3.2 Conditioning

Suppose $(X_v)_{v \in V}$ respects G conditioned on $(X_v)_{v \notin A}$.
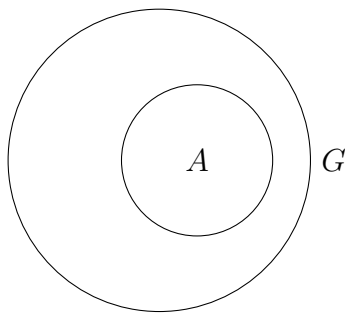
Conditioning: What graph does (X_v) respect?

Marginalizing: What graph does (X_A) respect?

Example: Let $X \sim \text{Bernoulli}(1/2)$ and $Y = 1 - X$.

1. Conditioned on X, what is the distribution of Y? Answer: fixed point, since there is no randomness

2. What is the marginal distribution of Y? Answer: Bernoulli(1/2)

Example: Let G be
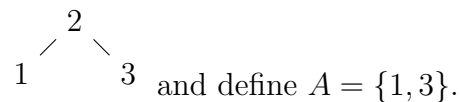
 and define $A = \{1, 3\}$.

Conditioned on X_2, what graph does (X_n) represent? What graph respects the marginal distribution of X_A?

Answer:

- The conditioned graph is

- The marginal graph is

$$\bullet \underset{1}{\rule{0pt}{0pt}} \!\!\!\!\!\!\!\!\! \underline{\hspace{2cm}} \!\!\!\!\!\!\!\!\! \underset{3}{\bullet}$$
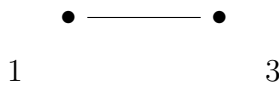
Remark: While this marginal graph works, it is the worst case graph. i.e., in most cases there is a smaller graph possible.

> **Theorem:**
>
> - Conditioning: G_1 is the reduced subgraph of G on A: $G_1 = (A, E_A)$ where $E_A = E(G) \cap (A \times A)$. (That is, edges in G_1 are simply edges in G restricting on A.
>
> - Marginalizing: $G_2 = (A, E_2(A))$ defined by $(u, v) \in E_2(A)$ if $u \sim v$ or u, v connected by a path outside A.
>
>)

Example: Let G be
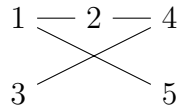
Then for $A = \{1, 2, 3\}$,

G_1	G_2
$1 - 2$	$1 - 2$
3	3
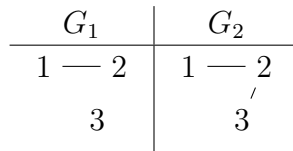
3.4 April 7

> **Theorem:** Let $(X_v)_{v \in G}$ be aGRF wrt a graph G if $A \subseteq V(G)$. Conditioned on X_{A^c}, the rv X_A is a GRF wrt the induced graph on A, i.e. $v \overset{G'}{\sim} u$ in this graph iff $v, u \in A$ and $v \overset{G}{\sim} u$.
>
> *Example:* Let G be
>
> $$3$$
> $$1$$
> $$2$$
>
> and $A = \{1, 2\}$ so $G = (\{1, 2\}, \emptyset)$.
>
> *Proof for the example:* Take X_1, X_2, X_3 from the joint distribution. Then
>
> $$p(x_1, x_2, x_3) = \frac{1}{Z}\phi_1(x_1)\phi_2(x_2)\phi_{12}(x_1, x_2)\phi_{13}(x_1, x_3)\phi_{23}(x_2, x_3)$$
> $$= \frac{1}{Z}f_{13}(x_1, x_3) \cdot f_{23}(x_2, x_3)$$
>
> ---
>
> *Proof for the General Case:* Suppose (X_v) is a GRF wrt G so
>
> $$p(x_1, \ldots, x_n) = \frac{1}{Z}\prod_{c \text{ cliques}} \phi_c(x_c) = \frac{1}{Z}\prod_{c \text{ cliques}} \phi_c(x_{A \cap C}, x_{A^c \cap C})$$

for any C satisfying $C = (A \cap C) \cup (A^c \cap C)$ (i.e. the measurable sets C)

Fix x_{A^c} so

$$\mathbb{P}_{x_A \mid X_{A^c} = x_{A^c}}(x_A) = \frac{\mathbb{P}(X_A = x_A, X_{A^c} = x_{A^c})}{\mathbb{P}(X_{A^c} = x_{A^c})}$$

but notice the denominator does not depend on x_A so we can call it a constant Z'.

Then,

$$\frac{\mathbb{P}(X_A = x_A, X_{A^c} = x_{A^c})}{Z'} = \frac{\frac{1}{Z}\prod_{c \text{ cliques}} \phi_c(x_{A \cap C}, x_{A^c \cap C})}{Z'}$$

$$= \frac{1}{Z \cdot Z'} \overbrace{\left[\prod_{\text{cliques } C \subseteq A^c} \phi_c(c_{A^c \cap C})\right]}^{Z''} \left[\prod_{\text{cliques } C \not\subseteq A^c} \phi_c(x_{A \cap C}, x_{A^c \cap C})\right]$$

$$= \frac{Z''}{ZZ'} \prod_{\text{cliques } C \cap A \neq \emptyset} \phi_c(x_{A \cap C}, x_{A^c \cap C})$$
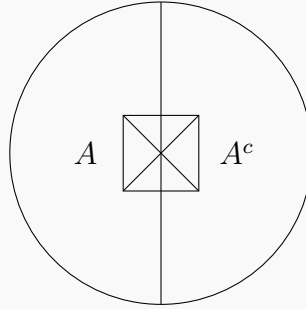
Since $\phi_c(x_{A \cap C}, x_{A^c \cap C})$ is a function on the clique $C \cap A$, $p(x_A)$ can be factored into cliques.

Definition 1: (X_v) is a GRF wrt G if

$$p(x_v) = \frac{1}{Z} \prod_{c \text{ cliques}} \phi_c(x_c)$$

Definition 2: (X_v) is a GRF wrt G if

$$p(x_v) = \frac{1}{Z} \prod_{c \text{ maximum cliques}} \phi_c(x_c)$$

Hence,

$$p(x_1, x_2 \mid x_3) = \frac{1}{Z'_{x_3}} f_{13,x_3}(x_1) \cdot f_{23,x_3}(x_2)$$

$$= \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{\mathbb{P}(X_3 = x_3)}$$

Example:

$$X_1 \sim \text{Bernoulli}(1/2)$$
$$X_3 = X_1 + \text{Bernoulli}(0.01)$$
$$X_2 = X_3 + \text{Bernoulli}(0.9)$$

gives

$$X_1 \sim \begin{cases} 0 & \text{with prob } 0.5 \\ 1 & \text{with prob } 0.5 \end{cases}$$

$$X_3 = \begin{cases} 0 & \frac{1}{2} \cdot \frac{99}{100} \\ 1 & \frac{1}{2} \\ 2 & \frac{1}{2} \cdot \frac{1}{100} \end{cases}$$

$$X_2 = \begin{cases} 0 \\ 1 \\ 2 \\ 3 \end{cases}$$

$$p(x_1, x_2, x_3) = \frac{1}{Z} f_{13}(x_3, x_1) \cdot f_{23}(x_2, x_3)$$
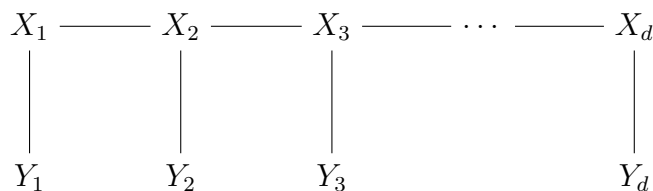
Conditioned on $X_3 = 0$,

$$p(x_1, x_2) = p_{x_3=0}(0, 1) = \frac{\mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 0)}{\mathbb{P}(X_3 = 0)} = \frac{\mathbb{P}(0, 1, 0)}{\sum_{x,y} \mathbb{P}(x, y, 0)} = \frac{1}{Z} f_{13}(0, 0) = \frac{f_{23}(1, 0)}{Z'}$$

Example (Naive Bayes):
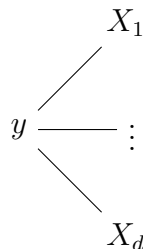
Conditioned on Y, the distribution of $X_1, \ldots, X_d$ is $(X_1, \ldots, X_d) \sim \text{Multinomial}(p_1, \ldots, p_d)$.

Example (Hidden Markov Model):

Conditioned on $Y_1, \ldots, Y_d$, the distribution of $X_1, \ldots, X_d$ is

$$X_1 \text{ ——— } X_2 \text{ ——— } X_3 \text{ ——— } \cdots \text{ ——— } X_d$$
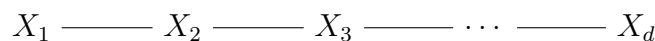
Conditioned on $X_1, \ldots, X_d$ the distribution of $Y_1, \ldots, Y_d$ is

$$Y_1 \qquad Y_2 \qquad Y_3 \qquad \cdots \qquad Y_d$$
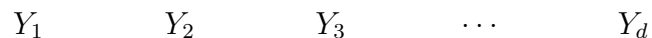
that is, they are independent.

Theorem (Marginalizing): Let the marginal distribution of (X_A) be a GRF wrt $G'' = (A, E'')$. Then for $u, v \in A$,

$$u \overset{E''}{\sim} v \impliedby \begin{cases} u \overset{G}{\sim} v \\ \text{there exists a path between } u \text{ and } v \text{ entirely in } A^c \end{cases}$$

Example:
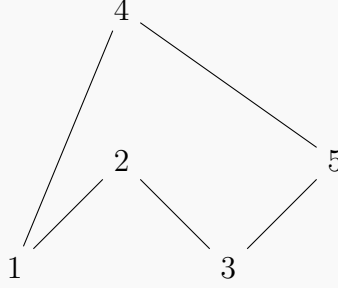
with $A = \{1, 2, 3, 5\}$ we have the subgraph
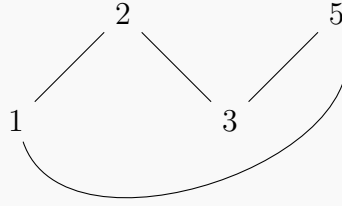
Proof of theorem for the example:

$$p(x_1, x_2, x_3, x_5) = \sum_{x_4} p(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_4} \frac{1}{Z} \phi_{12}(x_1, x_2) \phi_{23}(x_2, x_3) \phi_{35}(x_3, x_5) p_{14}(x_1, x_4) \phi_{45}(x_4, x_5)$$

$$\sum_{x_4} p_{14} p_{45} = f(x_1, x_5)$$

Proof Sketch:

$$p_{X_A}(x_A) = \mathbb{P}(X_A = x_A) = \sum_{x_{A^c}} \mathbb{P}(x_A, x_{A^c})$$

but we know

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{\text{cliques } c \subseteq G} \phi_c(x_c)$$

$$= \frac{1}{Z} \sum_{x_{A^c}} \prod_{\text{cliques } c \subseteq G} \phi_c(x_{c \cap A}, x_{C \cap A^c})$$

$$\overset{?}{=} \frac{1}{Z'} \prod_{\text{cliques } c \subseteq G''} \phi_c(x_{C \cap A})$$

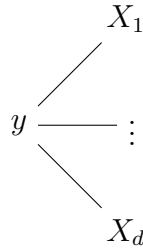**Exercise:** Prove the theorem for the general case.

Example: For
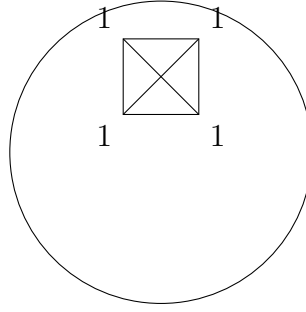
and fixed c,

$$\sum_{x_{34}} \phi_c(x_{12}, x_{34}) = \phi_{12}(x_{12})$$

Example (Naive Bayes): The marginal distribution of (X_i) is the complete graph

Example (Hidden Markov Model): If we were to condition on the observations (Y_i), then we would have a Markov chain. However, the marginal distribution on $Y_1, \ldots, Y_d$ is the complete graph
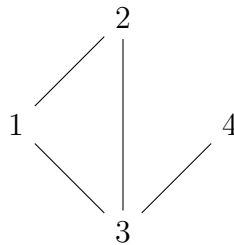
3.4.1 Markov Random Fields

Markov Random Field: $(X_v)_{v \in G}$ is a *Markov Random Field* (MRF) if

$$\mathbb{P}(X_i = x_i \mid X_{i^c} = x_{i^c}) = \mathbb{P}(X_i = x_i \mid x_{N_i} = x_{N_i})$$

where $N(i) = \{j : (j, i) \in E\}$ is the neighborhood of i.

Example:

$$\mathbb{P}(X_1 = x_1 \mid X_2 = x_2, X_3 = x_3, X_4 = x_4) = \mathbb{P}(X_1 = x_1 \mid X_2 = x_2, X_3 = x_3)$$

Theorem (Hammersley-Clifford): Assume (X_v) is positive (i.e. $p(x_1, \ldots, x_n) > 0$ for all $x_1, \ldots, x_n \in \Omega_1 \times \cdots \times \Omega_n$). Then X is a Gibbs Random Field iff it is a Markov Random Field.

Proof: (GRF $\implies$ MRF) Conditioned on $X_{N(i)}$, the distribution of is $A = \{i\} \cup \{j : j \notin N(i)\}$ from which it immediately follows that $i \perp \{j : j \notin N(i)\}$ hence

$$\mathbb{P}(X_i = x_i \mid X_{N_i}, \ldots) = \mathbb{P}(X_i = x_i \mid X_{N_i})$$

3.4.2 Computing Grpahical Models
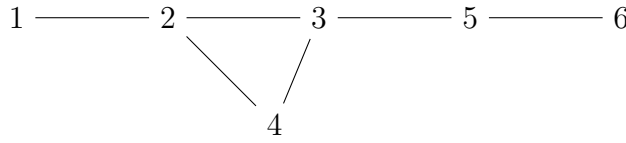
To use any graphical model

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_c \phi_c(x_c)$$

we need to know

$$Z = \sum_{x_1, \ldots, x_n} \prod_c \phi_c(x_c)$$

which is very difficult in practice.

Example:

$$1 \quad\rule{1cm}{0.4pt}\quad 2 \quad\rule{1cm}{0.4pt}\quad 3 \quad\rule{1cm}{0.4pt}\quad 5 \quad\rule{1cm}{0.4pt}\quad 6$$
$$4$$

Recall that we can specify p by the product of all the cliques or just the maximum cliques:

$$p(x_1, \ldots, x_6) = \frac{1}{Z} \phi_{12}(x_{12}) \phi_{234}(x_{234}) \phi_{25}(x_{35}) \phi_{56}(x_{56})$$

Since $\sum_{x_1, \ldots, x_6} p = 1$,

$$
\begin{aligned}
Z &= \sum_{x_1, \ldots, x_6} \phi_{12}(x_{12}) \phi_{234}(x_{234}) \phi_{25}(x_{35}) \phi_{56}(x_{56}) \\
&= \sum_{x_6} \sum_{x_5} \sum_{x_4} \sum_{x_3} \sum_{x_2} \sum_{x_1} \phi_{12}(x_{12}) \phi_{234}(x_{234}) \phi_{25}(x_{35}) \phi_{56}(x_{56}) \\
&= \sum_{x_6} \left(\sum_{x_5} \phi_{56}(x_{56}) \left(\sum_{x_3} \phi_{35}(x_{35}) \left(\sum_{x_4, x_2} \phi_{234}(x_{234}) \left(\sum_{x_1} \phi_{12}(x_{12}) \right) \right) \right) \right) \\
&= \sum_{x_6} \left(\sum_{x_5} \phi_{56}(x_{56}) \left(\sum_{x_3} \phi_{35}(x_{35}) \left(\sum_{x_4, x_2} \phi_{234}(x_{234}) T_1(x_2) \right) \right) \right) \\
&= \sum_{x_6} \left(\sum_{x_5} \phi_{56}(x_{56}) \left(\sum_{x_3} \phi_{35}(x_{35}) T_2(x_3) \right) \right) \\
&= \sum_{x_6} \left(\sum_{x_5} \phi_{56}(x_{56}) T_3(x_5) \right) \\
&= \sum_{x_6} T_5(x_6) \\
&= T_6
\end{aligned}
$$

We can calculate

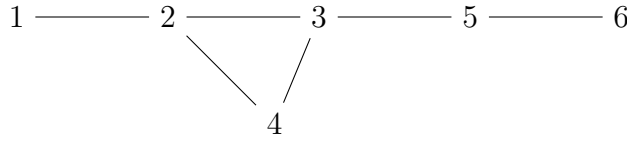|	cost
$T_1(x_2)$	$|\Omega|^2$
$T_2(x_3)$	$|\Omega|^3$
$T_3(x_5)$	$|\Omega|^2$
$T_4(x_5)$	$|\Omega|^2$
$T_5(x_6)$	$|\Omega|^2$
T_6	$|\Omega|^1$

giving total cost $|\Omega| + 4|\Omega|^2 + |\Omega|^3$ where the cost is calculated by $|\Omega|^{\#\text{new neighbors}+1}$ in the graph.
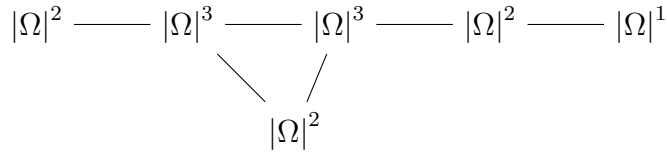
3.5   April 11

Let
$$p(x) = \frac{1}{Z}\phi_1(x_1)\phi_{12}(x_{12})\phi_{23}(x_{23})\phi_5(x_5)\phi_{56}(x_{56})$$

on

If we visit the vertices in the order $1, 2, 3, 4, 5, 6$, we can calculate costs

giving total $|\Omega| + 3|\Omega|^2 + 2|\Omega|^3$

by letting the currectly visited vertex v_k having visited $v_1, \ldots, v_{k-1}$ with cost $|\Omega|^{D_A+1}$ where

$$A = [n] \setminus \{v_1, \ldots, v_k\}$$
$$D_A = \partial A = \{u \in A : \exists v \notin A,\ u \sim v\}$$

Notice, though, that if we visit the vertices instead in the order $1, 2, 4, 3, 5, 6$, we have cost

$$|\Omega|^2 + |\Omega|^3 + |\Omega|^2 + |\Omega|^2 + |\Omega|^1 = |\Omega| + 4|\Omega|^2 + |\Omega|^3$$

which is much better. We call this the **boundary trick**.

> **Marginalization Theorem:** Let the marginal distribution of (X_A) be a GRF wrt $G'' = (A, E'')$. Then for $u, v \in A$,
> $$u \overset{E''}{\sim} v \impliedby \begin{cases} u \overset{G}{\sim} v \\ \text{there exists a path between } u \text{ and } v \text{ entirely in } A^c \end{cases}$$

Proof: Suppose A^c has connected components $D_1, \ldots, D_k$. Denote $M_1 = D_1 \cup \underbrace{N(D_1)}_{\subseteq A}$.

What can we say about $N(D_1)$? Since D_1 is connected, we know that there must be a path between points u, v in D_1. By definition, $u \overset{N(D_1)}{\sim} v$ so $N(D_1)$ must be a complete graph.

Observation 1: $\forall i$, $N(D_i)$ is complete in G''

Example:

With $A = \{12, 3, 5\}$, $D_1 = \{4\}$ and $D_2 = \{6\}$ so

$$\frac{1}{Z} \sum_{x_4} \sum_{x_6} \prod_c \phi(x_{A \cap c}, x_{A^c \cap c}) = \frac{1}{Z} \sum_{x_4} \sum_{x_6} \left(\prod_{c \in \{12,23,25\}} \phi(x_{1235}) \right) \left(\prod_{c \in \{34,54\}} \phi(x_{12356}) \right)$$

and

$$\sum_{x_4} \prod_{c \in \{34,45\}} \phi_c(x_{A \cap c}, x_{A^c \cap c}) = \phi_{35}$$

3.6   April 14

Suppose we know

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_c \phi_c(x_c)$$

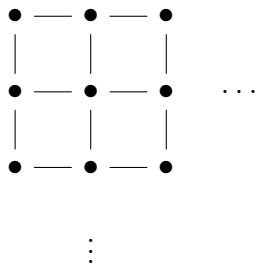where we know ϕ_c but do not know Z. How do we sample from p?

3.6.1   Dynamic Programming

One method is **Dynamic Programming**:

$$Z = \sum_{x_v} \prod_c \phi_c(x_c)$$

This is advantageous because it allows exact calculation of Z but is very slow in practice. (Recall that on the graphs last week this method had $O(|\Omega|^3)$ time.)

Consider if we were on an $n \times n$ lattice

Letting $G = (V, E)$ and $V = \{(i, j) : 1 \le i, j \le n\}$ we have n^2 vertices and $2n^2$ edges.
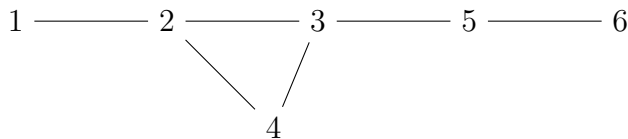
In this case, calculating Z directly on only $\Omega = \{\pm 1\}$ would require a sum over 2^{n^2} distributions and exponential time.

In any case, how do we use dynamic programming to sample from p? One easy way is to

1. Sample X_1

2. Sample $X_2 \mid X_1$

3. Sample $X_n \mid X_1, \ldots, X_{n-1}$.

Using this method, we only need to use the arrays T_i that we got from calculating Z.

Example: With graph

and GRF

$$p(x_1, \ldots, x_6) = \frac{1}{Z} \phi_{12} \phi_{234} \phi_{35} \phi_{56}$$

we have

$$Z = \sum_{x_6} \sum_{x_5} \sum_{x_3} \sum_{x_4} \sum_{x_2} \sum_{x_1} \phi_{12} \phi_{234} \phi_{35} \phi_{56}$$

$$= \sum_{x_6} \sum_{x_5} \phi_{56} \sum_{x_3} \phi_{35} \sum_{x_4} \sum_{x_2} \phi_{234} \underbrace{\sum_{x_1} \phi_{12}}_{T_1(x_2)}$$

$$\underbrace{\phantom{\sum_{x_4} \sum_{x_2} \phi_{234} \sum_{x_1} \phi_{12}}}_{T_2(x_{34})}$$

$$\underbrace{\phantom{\sum_{x_3} \phi_{35} \sum_{x_4} \sum_{x_2} \phi_{234} \sum_{x_1} \phi_{12}}}_{T_3(x_3)}$$

$$\underbrace{\phantom{\sum_{x_5} \phi_{56} \sum_{x_3}}}_{T_4(x_5)}$$

$$\underbrace{\phantom{\sum_{x_6} \sum_{x_5} \phi_{56}}}_{T_5(x_6)}$$

$$\underbrace{\phantom{\sum_{x_6}}}_{T_6}$$

Before writing down the conditional distribution, let's write down the joint distribution

$$p(x_1, \ldots, x_6) = \frac{1}{Z}\phi_{12}\phi_{234}\phi_{35}\phi_{56}$$

$$p(x_2, \ldots, x_6) = \sum_{x_1} p(x_1, \ldots, x_6)$$

$$= \frac{1}{Z}\phi_{234}\phi_{35}\phi_{56}T_1(x_2)$$

$$p(x_1 \mid x_2, \ldots, x_6) = \frac{p(x_1, \ldots, x_6)}{p(x_2, \ldots, x_6)}$$

$$= \frac{\phi_{12}}{T_1(x_2)}$$

Similarly over the other vertices,

$$p(x_3, \ldots, x_6) = \sum_{x_2} p(x_2, \ldots, x_6)$$

$$= \frac{1}{Z}\phi_{35}\phi_{56}T_2(x_{34})$$

$$p(x_2 \mid x_3, \ldots, x_6) = \frac{\phi_{234}T_1(x_2)}{T_2(x_{34})}$$

$$p(x_4 \mid x_3, \ldots, x_6) = \frac{T_2(x_{34})}{T_3(x_3)}$$

$$p(x_3 \mid x_5, x_6) = \frac{\phi_{35}T_3(x_3)}{T_4(x_5)}$$

$$p(x_5 \mid x_6) = \frac{T_4(x_5)}{T_5(x_6)}$$

$$p(x_6) = \frac{\phi_{56}T_5(x_6)}{Z}$$

3.6.2   Gibbs Samplers

Gibbs Sampling provides a cost effective alternative to dynamic program at the cost of asymptotic (rather than exact) convergence.

Example: Let G be

$$1 — 2$$

so the joint distribution is

$$p(x_1, x_2) = p(X_1 = x_1, X_2 = x_2)$$

Using the *Gibbs Sampler*, let π be any distribution on $V(G)$.

1. Initialize $X_1^{(0)}, \ldots, X_n^{(0)}$ to any values.

2. Sample a vertex i from π

3. Sample $X_i \mid X_{i^c} \sim p(x_i \mid x_{i^c})$

4. Iterate

Now we have RVs $(X_i^{(t)})_{i \in V, t=0,1,2,\ldots}$. As we will see, $\text{dist}(X_1^{(t)}, \ldots, X_n^{(t)}) \overset{t \to \infty}{\longrightarrow} p$.