What is this?

- Classical AI and ML research ignored this phenomena
- The Problem (an example):
  - you want to catch a flight at 10:00am from Pitt to SF, can I make it if I leave at 7am and take a 28X at CMU?
  - partial observability (road state, other drivers' plans, etc.)
  - noisy sensors (radio traffic reports)
  - uncertainty in action outcomes (flat tire, etc.)
  - immense complexity of modeling and predicting traffic
- Reasoning under uncertainty!
Basic Probability Concepts

- A sample space $S$ is the set of all possible outcomes of a conceptual or physical, repeatable experiment. ($S$ can be finite or infinite.)
  - E.g., $S$ may be the set of all possible outcomes of a dice roll: $S = \{1, 2, 3, 4, 5, 6\}$
  - E.g., $S$ may be the set of all possible nucleotides of a DNA site: $S = \{A, T, C, G\}$
  - E.g., $S$ may be the set of all possible positions time-space positions of an aircraft on a radar screen: $S = \{0, R_{\text{max}}\} \times \{0, 360^\circ\} \times \{0, +\infty\}$

- An event $A$ is the any subset $S$:
  - Seeing "1" or "6" in a roll; observing a "G" at a site; UA007 in space-time interval $X$

- An event space $E$ is the possible worlds the outcomes can happen
  - All dice-rolls, reading a genome, monitoring the radar signal

Visualizing Probability Space

- A probability space is a sample space of which, for every subset $s \in S$, there is an assignment $P(s) \in S$ such that:
  - $0 \leq P(s) \leq 1$
  - $\sum_{s \in S} P(s) = 1$

- $P(s)$ is called the probability (or probability mass) of $s$

$P(a)$ is the area of the oval

Event space of all possible worlds.
Its area is 1

Worlds in which $A$ is true

Worlds in which $A$ is false
Kolmogorov Axioms

- All probabilities are between 0 and 1
  - \( 0 \leq P(X) \leq 1 \)
- \( P(\text{true}) = 1 \)
  - regardless of the event, my outcome is true
- \( P(\text{false}) = 0 \)
  - no event makes my outcome true
- The probability of a disjunction is given by
  - \( P(A \lor B) = P(A) + P(B) - P(A \land B) \)

Why use probability?

- There have been attempts to develop different methodologies for uncertainty:
  - Fuzzy logic
  - Qualitative reasoning (Qualitative physics)
  - ...
- “Probability theory is nothing but common sense reduced to calculation”
  - — Pierre Laplace, 1812.
- In 1931, de Finetti proved that it is irrational to have beliefs that violate these axioms, in the following sense:
  - If you bet in accordance with your beliefs, but your beliefs violate the axioms, then you can be guaranteed to lose money to an opponent whose beliefs more accurately reflect the true state of the world. (Here, “betting” and “money” are proxies for “decision making” and “utilities”.)
- What if you refuse to bet? This is like refusing to allow time to pass: every action (including inaction) is a bet
Random Variable

- A random variable is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)
  - Discrete r.v.:
    - The outcome of a dice-roll
    - The outcome of reading a nt at site $i$: $X_i$
  - Binary event and indicator variable:
    - Seeing an "A" at a site $\Rightarrow X=1$, o/w $X=0$.
    - This describes the true or false outcome a random event.
    - Can we describe richer outcomes in the same way? (i.e., $X=1, 2, 3, 4$, for being A, C, G, T) — think about what would happen if we take expectation of $X$.
  - Unit-Base Random vector
    $X_\omega=[X_a, X_g, X_c, X_t]$, $X_\omega=[0,0,1,0]$ $\Rightarrow$ seeing a "G" at site $i$
  - Continuous r.v.:
    - The outcome of recording the true location of an aircraft: $X_{true}$
    - The outcome of observing the measured location of an aircraft $X_{obs}$

Discrete Prob. Distribution

- (In the discrete case), a probability distribution $P$ on $S$ (and hence on the domain of $X$) is an assignment of a non-negative real number $P(s)$ to each $s\in S$ (or each valid value of $x$) such that $\Sigma_{s\in S}P(s)=1$. ($0\leq P(s) \leq 1$)
  - intuitively, $P(s)$ corresponds to the frequency (or the likelihood) of getting $s$ in the experiments, if repeated many times
  - call $\theta= P(s)$ the parameters in a discrete probability distribution
- A probability distribution on a sample space is sometimes called a probability model, in particular if several different distributions are under consideration
  - write models as $M_1$, $M_2$, probabilities as $P(X|M_1)$, $P(X|M_2)$
  - e.g., $M_1$ may be the appropriate prob. dist. if $X$ is from "fair dice", $M_2$ is for the "loaded dice".
  - $M$ is usually a two-tuple of (dist. family, dist. parameters)
Bernoulli distribution: \( \text{Ber}(p) \)

\[
\begin{align*}
\rho(x) &= \begin{cases} 
1-p & \text{for } x=0 \\
p & \text{for } x=1
\end{cases} \\
\Rightarrow \rho(x) &= p^x (1-p)^{1-x}
\end{align*}
\]

Multinominal distribution: \( \text{Mult}(1, \theta) \)

- Multinominal (indicator) variable:

\[
X = \begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_k
\end{bmatrix}, \quad \text{where} \quad \sum_{j=1}^{k} X_j = 1 \quad \text{and} \quad \sum_{j=1}^{k} \theta_j = 1.
\]

\[
\rho(x(j)) = \rho(\{X_j = 1, \text{where } j \text{ index the dice - face}\}) = \theta_j \times \theta_j \cdots \theta_j \times \theta_k \times \cdots \theta_k = \prod_{j} \theta_j = \theta^x
\]

Multinominal distribution: \( \text{Mult}(n, \theta) \)

- Count variable:

\[
X = \begin{bmatrix}
X_1 \\
\vdots \\
X_k
\end{bmatrix}, \quad \text{where} \quad \sum_{j=1}^{k} X_j = n
\]

\[
\rho(x) = \frac{n!}{x_1!x_2!\cdots x_k!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k} = \frac{n!}{x_1!x_2!\cdots x_k!} \theta^x
\]
Continuous Prob. Distribution

- A continuous random variable $X$ can assume any value in an interval on the real line or in a region in a high dimensional space.
  - $X$ usually corresponds to a real-valued measurements of some property, e.g., length, position, ...
  - It is not possible to talk about the probability of the random variable assuming a particular value --- $P(x) = 0$
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval

  - $P(X = [x_1, x_2])$,
  - $P(X < x) = P(X \in [-\infty, x])$

- Arbitrary Boolean combination of basic propositions

The probability of the random variable assuming a value within some given interval from $x_1$ to $x_2$ is defined to be the area under the graph of the probability density function between $x_1$ and $x_2$.

- Probability mass: $P(X \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x) dx$,
  - note that $\int_{-\infty}^{\infty} p(x) dx = 1$.
- Cumulative distribution function (CDF):
  - $P(X) = P(X < x) = \int_{-\infty}^{x} p(x') dx'$
- Probability density function (PDF):
  - $p(x) = \frac{d}{dx} P(x)$
  - $\int_{-\infty}^{\infty} p(x) dx = 1$; $p(x) > 0, \forall x$
What is the intuitive meaning of \( p(x) \)?

- If \( p(x_1) = a \) and \( p(x_2) = b \), then when a value \( X \) is sampled from the distribution with density \( p(x) \), you are \( \frac{a}{b} \) times as likely to find that \( X \) is “very close to” \( x_1 \) than that \( X \) is “very close to” \( x_2 \).

That is:

\[
\lim_{h \to 0} \frac{\int_{x_1-h}^{x_1+h} p(x) \, dx}{\int_{x_2-h}^{x_2+h} p(x) \, dx} = \frac{a}{b} \]

Continuous Distributions

- Uniform Probability Density Function

\[
p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}
\]

- Normal (Gaussian) Probability Density Function

\[
p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
- The distribution is symmetric, and is often illustrated as a bell-shaped curve.
- Two parameters, \( \mu \) (mean) and \( \sigma \) (standard deviation), determine the location and shape of the distribution.
- The highest point on the normal curve is at the mean, which is also the median and mode.
- The mean can be any numerical value: negative, zero, or positive.

- Exponential Probability Distribution

\[
density: \quad p(x) = \frac{1}{\mu} e^{-x/\mu}, \quad CDF: \quad P(X \leq x_0) = 1 - e^{-x_0/\mu}
\]
Statistical Characterizations

- **Expectation**: the centre of mass, mean value, first moment:
  \[ E(X) = \begin{cases} 
  \sum_{i=5} x_i p(x_i) & \text{discrete} \\
  \int_{-\infty}^{\infty} x p(x) dx & \text{continuous}
  \end{cases} \]

- **Sample mean**:
  \[ \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \]

- **Variance**: the spreadness:
  \[ Var(X) = \begin{cases} 
  \sum_{i=5} (x_i - E(X))^2 p(x_i) & \text{discrete} \\
  \int_{-\infty}^{\infty} (x - E(X))^2 p(x) dx & \text{continuous}
  \end{cases} \]

- **Sample variance**
  \[ \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \]

Gaussian (Normal) density in 1D

- If \( X \sim N(\mu, \sigma^2) \), the probability density function (pdf) of \( X \) is defined as
  \[ p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

- We will often use the precision \( \lambda = 1/\sigma^2 \) instead of the variance \( \sigma^2 \).
- Here is how we plot the pdf in matlab
  ```
  xs=-3:0.01:3;
  plot(xs,normpdf(xs,mu,sigma))
  ```

- Note that a density evaluated at a point can be bigger than 1!
Gaussian CDF

- If $Z \sim N(0, 1)$, the cumulative density function is defined as

$$\Phi(x) = \int_{-\infty}^{x} p(z)\,dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2}\,dz$$

- This has no closed form expression, but is built in to most software packages (e.g., normcdf in Matlab Stats Toolbox).

Use of the cdf

- If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$.

- How much mass is contained inside the $[-1.98\sigma, 1.98\sigma]$ interval?

$$P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

- Since

$$p(Z \leq -1.96) = \text{normcdf}(-1.96) = 0.025$$

we have

$$P(-2\sigma < X - \mu < 2\sigma) = 1 - 2 \times 0.025 = 0.95$$
Central limit theorem

- If \((X_1, X_2, \ldots, X_n)\) are i.i.d. (we will come back to this point shortly) continuous random variables
- Then define
  \[ \overline{X} = f(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \]
- As \( n \to \infty \),
  \( p(\overline{X}) \to \) Gaussian with mean \( E[X_j] \) and variance \( \text{Var}[X_j] \)

Elementary manipulations of probabilities

- Set probability of multi-valued r.v.
  - \( P(\{x=\text{Odd}\}) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \)
  - \( P(X = x_1 \vee X = x_2, \ldots, \vee X = x_j) = \sum_{j=1}^{n} P(X = x_j) \)

- Multi-variant distribution:
  - Joint probability: \( P(X = \text{true} \wedge Y = \text{true}) \)
    \[ P(Y \wedge \{X = x_1 \vee X = x_2, \ldots, \vee X = x_j\}) = \sum_{j=1}^{n} P(Y \wedge X = x_j) \]
  - Marginal Probability: \( P(Y) = \sum_{j=1}^{n} P(Y \wedge X = x_j) \)
Conditional Probability

- \( P(X|Y) = \) Fraction of worlds in which \( X \) is true that also have \( Y \) true
  - \( H = "having a headache" \)
  - \( F = "coming down with Flu" \)
    - \( P(H)=1/10 \)
    - \( P(F)=1/40 \)
    - \( P(H|F)=1/2 \)
  - \( P(H|F) = \) fraction of flu-inflicted worlds in which you have a headache
    = \( P(H \land F)/P(F) \)
- Definition:
  \[
  P(X|Y) = \frac{P(X \land Y)}{P(Y)}
  \]
- Corollary: The Chain Rule
  \[
  P(X \land Y) = P(X|Y)P(Y)
  \]

Probabilistic Inference

- \( H = "having a headache" \)
- \( F = "coming down with Flu" \)
  - \( P(H)=1/10 \)
  - \( P(F)=1/40 \)
  - \( P(H|F)=1/2 \)
- One day you wake up with a headache. You come with the following reasoning: "since 50% of flues are associated with headaches, so I must have a 50-50 chance of coming down with flu"

  Is this reasoning correct?
Probabilistic Inference

- \( H = \) "having a headache"
- \( F = \) "coming down with Flu"
  - \( P(H) = 1/10 \)
  - \( P(F) = 1/40 \)
  - \( P(H|F) = 1/2 \)

- The Problem:

\[
P(F|H) = ?
\]

The Bayes Rule

- What we have just did leads to the following general expression:

\[
P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}
\]

This is Bayes Rule

More General Forms of Bayes Rule

- \( P(Y | X) = \frac{P(X | Y)p(Y)}{P(X | Y)p(Y) + P(X | \neg Y)p(\neg Y)} \)
- \( P(Y = y_i | X) = \frac{P(X | Y)p(Y)}{\sum_{i=1}^{s} P(X | Y = y_i)p(Y = y_i)} \)
- \( P(Y | X \land Z) = \frac{P(X | Y \land Z)p(Y \land Z)}{P(X \land Z)} = \frac{P(X | Y \land Z)p(Y \land Z)}{P(X | Y \land Z)p(Y \land Z) + P(X | \neg Y \land Z)p(\neg Y \land Z) + P(X | Y \land Z)p(Y \land Z) + P(X | \neg Y \land Z)p(\neg Y \land Z)} \)

Prior Distribution

- Support that our propositions about the possible has a "causal flow"
  - e.g.,

- Prior or unconditional probabilities of propositions
  - e.g., \( P(Flu = \text{true}) = 0.025 \) and \( P(\text{DrinkBeer} = \text{true}) = 0.2 \)
  - correspond to belief prior to arrival of any (new) evidence

- A probability distribution gives values for all possible assignments:
  - \( P(\text{DrinkBeer}) = [0.01, 0.09, 0.1, 0.8] \)
  - (normalized, i.e., sums to 1)
Joint Probability

- A joint probability distribution for a set of RVs gives the probability of every atomic event (sample point)
  
  - $P(Flu, DrinkBeer) = a 2 \times 2$ matrix of values:

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>¬B</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0.005</td>
<td>0.02</td>
</tr>
<tr>
<td>¬F</td>
<td>0.195</td>
<td>0.78</td>
</tr>
</tbody>
</table>

- $P(Flu, DrinkBeer, Headache) = ?$

- Every question about a domain can be answered by the joint distribution, as we will see later.

Posterior conditional probability

- Conditional or posterior (see later) probabilities
  
  - e.g., $P(Flu|Headache) = 0.178$
  
  → given that flu is all I know
  
  NOT “if flu then 17.8% chance of Headache”

- Representation of conditional distributions:
  
  - $P(Flu|Headache) = 2$-element vector of 2-element vectors

- If we know more, e.g., DrinkBeer is also given, then we have
  
  - $P(Flu|Headache, DrinkBeer) = 0.070$ \textbf{This effect is known as explain away!}$
  
  - $P(Flu|Headache, Flu) = 1$
  
  - Note: the less or more certain belief remains valid after more evidence arrives, but is not always useful

- New evidence may be irrelevant, allowing simplification, e.g.,
  
  - $P(Flu|Headache, StealerWin) = P(Flu|Headache)$
  
  - This kind of inference, sanctioned by domain knowledge, is crucial
Inference by enumeration

- Start with a Joint Distribution
- Building a Joint Distribution of M=3 variables
  - Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have $2^M$ rows).
  - For each combination of values, say how probable it is.
  - Normalized, i.e., sums to 1

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>B</th>
<th>H</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.17</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.015</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Inference with the Joint

- One you have the JD you can ask for the probability of any atomic event consistent with your query
  $$P(E) = \sum_{i \in E} P(\text{row}_i)$$
Inference with the Joint

- Compute Marginals

\[ P(\text{Flu} \land \text{Headache}) = \]

\[ P(\text{Headache}) = \]
Inference with the Joint

- Compute Conditionals

\[ P(E_1|E_2) = \frac{P(E_1 \land E_2)}{P(E_2)} \]

\[ = \frac{\sum_{i \in E_1 \land E_2} P(\text{row}_i)}{\sum_{i \in E_2} P(\text{row}_i)} \]

- General idea: compute distribution on query variable by **fixing** evidence variables and **summing** over hidden variables
Summary: Inference by enumeration

- Let X be all the variables. Typically, we want
  - the posterior joint distribution of the query variables Y
  - given specific values e for the evidence variables E
  - Let the hidden variables be H = X-Y-E

- Then the required summation of joint entries is done by summing out the hidden variables:

\[ P(Y|E=e) = \alpha \sum_h P(Y, E=e, H=h) \]

- The terms in the summation are joint entries because Y, E, and H together exhaust the set of random variables

- Obvious problems:
  - Worst-case time complexity \( O(d^n) \) where d is the largest arity
  - Space complexity \( O(d^n) \) to store the joint distribution
  - How to find the numbers for \( O(d^n) \) entries???

Conditional independence

- Write out full joint distribution using chain rule:

\[
P(\text{Headache};\text{Flu};\text{Virus};\text{DrinkBeer}) = P(\text{Headache} | \text{Flu};\text{Virus};\text{DrinkBeer}) P(\text{Flu};\text{Virus};\text{DrinkBeer})
= P(\text{Headache} | \text{Flu};\text{Virus};\text{DrinkBeer}) P(\text{Flu} | \text{Virus};\text{DrinkBeer}) P(\text{Virus} | \text{DrinkBeer}) P(\text{DrinkBeer})
\]

Assume independence and conditional independence

\[
P(\text{Headache}|\text{Flu};\text{DrinkBeer}) P(\text{Flu}|\text{Virus}) P(\text{Virus}) P(\text{DrinkBeer})
\]

i.e., \( ? \) independent parameters

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from \textit{exponential} in \( n \) to \textit{linear} in \( n \).

- Conditional independence is our most basic and robust form of knowledge about uncertain environments.
### Rules of Independence --- by examples

- \( P(\text{Virus} \mid \text{DrinkBeer}) = P(\text{Virus}) \)
  
  iff Virus is independent of DrinkBeer

- \( P(\text{Flu} \mid \text{Virus};\text{DrinkBeer}) = P(\text{Flu} \mid \text{Virus}) \)
  
  iff Flu is independent of DrinkBeer, given Virus

- \( P(\text{Headache} \mid \text{Flu};\text{Virus};\text{DrinkBeer}) = P(\text{Headache} \mid \text{Flu};\text{DrinkBeer}) \)
  
  iff Headache is independent of Virus, given Flu and DrinkBeer

### Marginal and Conditional Independence

- Recall that for events \( E \) (i.e. \( X=x \)) and \( H \) (say, \( Y=y \)), the conditional probability of \( E \) given \( H \), written as \( P(E \mid H) \), is
  
  \[
P(E \text{ and } H)/P(H)
  
  (= the probability of both \( E \) and \( H \) are true, given \( H \) is true)

- \( E \) and \( H \) are (statistically) independent if
  
  \[P(E) = P(E \mid H)\]
  
  (i.e., prob. \( E \) is true doesn't depend on whether \( H \) is true); or equivalently
  
  \[P(E \text{ and } H) = P(E)P(H).\]

- \( E \) and \( F \) are conditionally independent given \( H \) if
  
  \[P(E \mid H,F) = P(E \mid H)\]
  
  or equivalently
  
  \[P(E,F \mid H) = P(E \mid H)P(F \mid H).\]
Why knowledge of Independence is useful

- Lower complexity (time, space, search …)

- Motivates efficient inference for all kinds of queries
  
  Stay tuned!!

- Structured knowledge about the domain
  - easy to learning (both from expert and from data)
  - easy to grow

Where do probability distributions come from?

- Idea One: Human, Domain Experts
- Idea Two: Simpler probability facts and some algebra
  
  e.g., \( P(F) \)
  \( P(B) \)
  \( P(H|\neg F,B) \)
  \( P(H|F,\neg B) \)
  …

- Idea Three: Learn them from data!
  
  - A good chunk of this course is essentially about various ways of learning various forms of them!
Density Estimation

- A Density Estimator learns a mapping from a set of attributes to a Probability

| Input Attributes | Density Estimator | → Probability |

- Often know as parameter estimation if the distribution form is specified
  - Binomial, Gaussian ...

- Three important issues:
  - Nature of the data (iid, correlated, …)
  - Objective function (MLE, MAP, …)
  - Algorithm (simple algebra, gradient methods, EM, …)
  - Evolution scheme (likelihood on test data, predictability, consistency, …)

Parameter Learning from iid data

- Goal: estimate distribution parameters $\theta$ from a dataset of $N$ independent, identically distributed (iid), fully observed, training cases

$$D = \{x_1, \ldots, x_N\}$$

- Maximum likelihood estimation (MLE)
  1. One of the most common estimators
  2. With iid and full-observability assumption, write $L(\theta)$ as the likelihood of the data:

$$L(\theta) = P(x_1, x_2, \ldots, x_N; \theta)$$
$$= P(x_1; \theta) P(x_2; \theta) \cdots P(x_N; \theta)$$
$$= \prod_{i=1}^{N} P(x_i; \theta)$$

  3. pick the setting of parameters most likely to have generated the data we saw:

$$\theta^* = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \log L(\theta)$$
Example 1: Bernoulli model

- Data:
  - We observed \( N \) iid coin tossing: \( D = \{1, 0, 1, \ldots, 0\} \)

- Representation:
  Binary r.v.
  \( x_n = \{0, 1\} \)

- Model:
  \[
P(x) = \begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = \theta^x (1-\theta)^{1-x}
  \]

- How to write the likelihood of a single observation \( x_i \)?
  \[
P(x_i) = \theta^{x_i} (1-\theta)^{1-x_i}
  \]

- The likelihood of dataset \( D = \{x_1, \ldots, x_N\} \):
  \[
P(x_1, x_2, \ldots, x_N | \theta) = \prod_{i=1}^{N} P(x_i | \theta) = \prod_{i=1}^{N} \left( \theta^{x_i} (1-\theta)^{1-x_i} \right) = \theta^{\sum_{i=1}^{N} x_i} (1-\theta)^{\sum_{i=1}^{N} (1-x_i)} = \theta^{\text{counts}} (1-\theta)^{\text{tails}}
  \]

---

MLE

- Objective function:
  \[
  \ell(\theta; D) = \log P(D | \theta) = \log \theta^{n_h} (1-\theta)^{n_t} = n_h \log \theta + (N - n_h) \log(1-\theta)
  \]

- We need to maximize this w.r.t. \( \theta \)

- Take derivatives wrt \( \theta \)
  \[
  \frac{\partial \ell}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1-\theta} = 0 \quad \Rightarrow \quad \hat{\theta}_{MLE} = \frac{n_h}{N} \quad \text{or} \quad \hat{\theta}_{MLE} = \frac{1}{N} \sum_i x_i
  \]

- Sufficient statistics
  - The counts, \( n_h \), where \( n_h = \sum x_i \), are sufficient statistics of data \( D \)
MLE for discrete (joint) distributions

- More generally, it is easy to show that

\[ P(\text{event}_i) = \frac{\# \text{records in which event}_i \text{ is true}}{\text{total number of records}} \]

- This is an important (but sometimes not so effective) learning algorithm!

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>B</td>
<td>H</td>
<td>0.4</td>
</tr>
<tr>
<td>F</td>
<td>H</td>
<td>H</td>
<td>0.1</td>
</tr>
<tr>
<td>F</td>
<td>B</td>
<td>H</td>
<td>0.17</td>
</tr>
<tr>
<td>F</td>
<td>B</td>
<td>H</td>
<td>0.2</td>
</tr>
<tr>
<td>F</td>
<td>B</td>
<td>H</td>
<td>0.05</td>
</tr>
<tr>
<td>F</td>
<td>B</td>
<td>H</td>
<td>0.05</td>
</tr>
<tr>
<td>F</td>
<td>B</td>
<td>B</td>
<td>0.015</td>
</tr>
<tr>
<td>F</td>
<td>B</td>
<td>H</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Example 2: univariate normal

- Data:
  - We observed \( N \) iid real samples:
    \( D = \{-0.1, 10, 1, -5.2, \ldots, 3\} \)
  - Model:
    \( P(x) = \left(2\pi \sigma^2\right)^{-1/2} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} \)

- Log likelihood:
  \[
  \ell(\theta; D) = \log P(D | \theta) = -\frac{N}{2} \log(2\pi \sigma^2) - \frac{1}{2} \sum_{n=1}^{N} \left(x_n - \mu\right)^2
  \]

- MLE: take derivative and set to zero:
  \[
  \frac{\partial \ell}{\partial \mu} = \left(1/\sigma^2\right) \sum_n (x_n - \mu) \quad \Rightarrow \quad \mu_{\text{MLE}} = \frac{1}{N} \sum_n x_n
  \]
  \[
  \frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_n (x_n - \mu)^2 \quad \Rightarrow \quad \sigma^2_{\text{MLE}} = \frac{1}{N} \sum_n (x_n - \mu_{\text{MLE}})^2
  \]
Overfitting

- Recall that for Bernoulli Distribution, we have

\[ \hat{\theta}_{ML}^{\text{head}} = \frac{n_{\text{head}}}{n_{\text{head}} + n_{\text{tail}}} \]

- What if we tossed too few times so that we saw zero head? We have \( \hat{\theta}_{ML}^{\text{head}} = 0 \), and we will predict that the probability of seeing a head next is zero!!!

- The rescue:
  - Where \( n' \) is know as the pseudo- (imaginary) count

\[ \hat{\theta}_{ML}^{\text{head}} = \frac{n_{\text{head}} + n'}{n_{\text{head}} + n_{\text{tail}} + n'} \]

- But can we make this more formal?

The Bayesian Theory

- The Bayesian Theory: (e.g., for date \( D \) and model \( M \))

\[ P(M|D) = \frac{P(D|M)P(M)}{P(D)} \]

- the posterior equals to the likelihood times the prior, up to a constant.

- This allows us to capture uncertainty about the model in a principled way
Hierarchical Bayesian Models

- $\theta$ are the parameters for the likelihood $p(x|\theta)$
- $\alpha$ are the parameters for the prior $p(\theta|\alpha)$
- We can have hyper-hyper-parameters, etc.
- We stop when the choice of hyper-parameters makes no difference to the marginal likelihood; typically make hyper-parameters constants.
- Where do we get the prior?
  - Intelligent guesses
  - Empirical Bayes (Type-II maximum likelihood)

\[ \alpha_{\text{MLE}} = \arg \max_\alpha p(\hat{\alpha}) \]

Bayesian estimation for Bernoulli

- Beta distribution:
  
  \[ P(\theta;\alpha,\beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} = B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1} \]

- Posterior distribution of $\theta$:
  
  \[ P(\theta|x_1,...,x_N) = \frac{p(x_1,...,x_N|\theta)p(\theta)}{p(x_1,...,x_N)} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \]

  - Notice the isomorphism of the posterior to the prior,
  - such a prior is called a **conjugate prior**
Bayesian estimation for Bernoulli, con'd

- Posterior distribution of $\theta$:
  \[ P(\theta | x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n | \theta) p(\theta)}{p(x_1, \ldots, x_n)} = \theta^\alpha (1 - \theta)^\beta \times \theta^{n \alpha - 1} (1 - \theta)^{n \beta - 1} = \theta^{n \alpha + \alpha - 1} (1 - \theta)^{n \beta + \beta - 1} \]

- Maximum a posteriori (MAP) estimation:
  \[ \theta_{\text{MAP}} = \arg \max_\theta \log P(\theta | x_1, \ldots, x_n) \]

- Posterior mean estimation:
  \[ \theta_{\text{mean}} = \int \theta p(\theta | x_1, \ldots, x_n) d\theta = C \int \theta \theta^{n \alpha - 1} (1 - \theta)^{n \beta - 1} d\theta = \frac{n + \alpha}{N + \alpha + \beta} \]

- Prior strength: $A = \alpha + \beta$
  - $A$ can be interpreted as the size of an imaginary data set from which we obtain the pseudo-counts

Effect of Prior Strength

- Suppose we have a uniform prior ($\alpha = \beta = 1/2$), and we observe $\tilde{n} = (n_h = 2, n_t = 8)$
- Weak prior $A = 2$. Posterior prediction:
  \[ p(x = h | n_h = 2, n_t = 8, \tilde{\alpha} = \tilde{\alpha} \times 2) = \frac{1 + 2}{2 + 10} = 0.25 \]

- Strong prior $A = 20$. Posterior prediction:
  \[ p(x = h | n_h = 2, n_t = 8, \tilde{\alpha} = \tilde{\alpha} \times 20) = \frac{10 + 2}{20 + 10} = 0.40 \]

- However, if we have enough data, it washes away the prior. e.g., $\tilde{n} = (n_h = 200, n_t = 800)$. Then the estimates under weak and strong prior are $\frac{1 + 200}{2 + 1000}$ and $\frac{10 + 200}{20 + 1000}$, respectively, both of which are close to 0.2
Bayesian estimation for normal distribution

- Normal Prior:
  \[ P(\mu) = \left(2\pi\tau^2\right)^{-1/2} \exp\left\{ -\left(\mu - \mu_0\right)^2 / 2\tau^2 \right\} \]

- Joint probability:
  \[
P(x, \mu) = \left(2\pi\sigma^2\right)^{N/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\} \times \left(2\pi\tau^2\right)^{1/2} \exp\left\{ -\frac{1}{2\tau^2} (\mu - \mu_0)^2 \right\}
  \]

- Posterior:
  \[
P(\mu | x) = \left(2\pi\sigma^2\right)^{1/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \bar{\mu})^2 \right\}
  \]

where \( \bar{\mu} = \frac{N/\sigma^2}{N/\sigma^2 + 1/\tau} \bar{x} + \frac{1/\tau}{N/\sigma^2 + 1/\tau} \mu_0 \), and \( \sigma^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\tau}\right)^{-1} \)

Sample mean