Probability Distributions

Thanks to R Parr, C Guesterin
Outline

- Probability 101
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Mean, Variance

- Estimation
  - Binomial
  - Gaussian
  - Estimation: MLE, MAP
  - Conjugate Distributions
Probability: Who needs it?

- Learning without probabilities is possible
  - Version spaces
  - Explanation-based learning

- Learning almost always involves
  - Noise in data
  - Prediction about the future

- Learning systems that don’t use probability in some way tend to be very, very brittle
Probabilities

- Natural way to represent uncertainty
- People have intuitive notions about probabilities
  - Many of these are wrong or inconsistent
  - Most people don’t get what probabilities mean
- Finer details of this question still debated
Understanding Probabilities

- Probabilities have dual meanings
  - Relative frequencies (frequentist view)
  - Degree of belief (Bayesian view)

- Neither is entirely satisfying
  - No two events are truly the same (reference class problem)
  - Statements should be grounded in reality in some way
Why Probabilities are Good … despite difficulties

- Subjectivists: *probabilities are degrees of belief*

- Is any *degree of belief* \(\equiv\) *probability*?
  - AI has used many notions of belief:
    - Certainty Factors
    - Fuzzy Logic

- **NO!!**
  - Dutch book
  - If you follow doesn’t follow probability theory, you will lose… see below.
Terms from Probability Theory

- **Random Variable:**
  \( \text{Weather} \in \{ \text{Sunny, Rain, Cloudy, Snow} \} \)

- **Domain:** Possible values a random variable can take.
  \( \ldots \text{finite set, } \mathbb{R}, \text{functions} \ldots \)  

- **Probability distribution:**
  mapping from domain to values \( \in [0, 1] \)

- \( P( \text{Weather} ) = \langle 0.7, 0.2, 0.08, 0.02 \rangle \)
  \[ \begin{align*}
  P( \text{Weather} = \text{Sunny} ) &= 0.7 \\
  P( \text{Weather} = \text{Rain} ) &= 0.2 \\
  P( \text{Weather} = \text{Cloudy} ) &= 0.08 \\
  P( \text{Weather} = \text{Snow} ) &= 0.02 
  \end{align*} \]
  means

- **Event:**
  Each assignment (e.g., \( \text{Weather} = \text{Rain} \)) is “event”
Hepatitis? Hepatitis, not Jaundiced but +BloodTest? Jaundiced BloodTest
Typical Task

- Given observations \( \{O_1=v_1, \ldots, O_k=v_k\} \)
  
  (J=No, B=Yes [symptoms, history, test results, …])

  what is best DIAGNOSIS \( D_{x_i} \) for patient?
  
  (Hep=Yes, Hep=No)

- Compute Probabilities of \( D_{x_i} \)

  given observations \( \{O_i=v_i, \ldots, O_k=v_k\} \)

\[
P(\ D_{x} = u \mid O_1 = v_1, \ldots, O_k = v_k )
\]
General Events

- **Atomic Event**: “Complete specification"
  Conjunction of assignments to EVERY variable [PossibleWorld]

- **Joint Probability Distribution**:
  Probability of every possible atomic event

---

$n$ binary variables: $2^n$ entries
($2^n - 1$ independent values, as sum = 1)
A huge table!

<table>
<thead>
<tr>
<th>J</th>
<th>B</th>
<th>H</th>
<th>P(j,b,h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03395</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.0095</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.0003</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.1805</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.01455</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.038</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.00045</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.722</td>
</tr>
</tbody>
</table>
Inference by Enumeration

- Using only joint probability distribution:

- For any proposition \( \varphi \), add the atomic events where it is true:
  \[
  P(\varphi) = \sum_{\omega: \omega \models \varphi} P(\omega)
  \]

- \( P( +j ) \)
  \[
  = 0.01455 + 0.038 + 0.00045 + 0.722 \\
  = 0.775
  \]
Cost of Marginalization

- Called “marginal”

\[
P(X_n) = \sum_{x_1, \ldots, x_{n-1}} P(x_1, \ldots, x_{n-1}, X_n)
\]

- To compute marginal distribution \(P(X_n)\):
  - If all binary, \(2^{n-1}\) additions
  - One term for each value of \(x_1, \ldots, x_{n-1}\)
Inference by Enumeration

- Using only joint probability distribution:

- For any proposition $\varphi$, add the atomic events where it is true:

\[
P(\varphi) = \sum_{\omega : \omega \models \varphi} P(\omega)
\]

- $P(-j \lor +b)$

\[
= 0.03395 + 0.0095 + 0.0003 + 0.1805 + 0.0045 + 0.722 = 0.9467
\]
Conditional Probabilities

- After learning that $\beta$ is true, how do we feel about $\alpha$?
- If roll EVEN, what is chance of rolling 2?
- If have hepatitis, what is chance of jaundice?

$$P(\alpha \mid \beta)$$
Conditional Probability

- Conditional Probability:
  \[ P(\alpha | \beta) = \text{Probability of event } \alpha, \text{ given that event } \beta \text{ has happened} \]
- \[ P(\text{Jaundice} | \text{Hepatitis}) = 0.8 \]

- In gen'l:
  \[
P(\alpha | \beta) = \frac{P(\alpha \& \beta)}{P(\beta)}
\]
  \[P(\alpha \& \beta) = P(\alpha | \beta) \cdot P(\beta)\]
Conditional Probability

Unconditional (prior) Probability:
- Probability of event before evidence is presented
- \( P(\text{Jaundice}) = 0.04 \)
  - prob that someone (from this population) is jaundiced is 4 in 100

Evidence: Percepts that affects degree of belief in event

Conditional (posterior) Probability:
- Probability of event after evidence is presented
- N.b., posterior prob can be COMPLETELY different than prior prob!

\[
P(\alpha | \beta) = \frac{P(\alpha \& \beta)}{P(\beta)}
\]

\[
P(\alpha \& \beta) = P(\alpha | \beta) \cdot P(\beta)
\]
Inference by Enumeration

- Using only joint probability distribution:

- Can compute \textit{conditional probabilities}:

\[
P(-\text{b} \mid +\text{j}) = \frac{P(-\text{b} \land +\text{j})}{P(+\text{j})} = \frac{0.01455 + 0.038}{0.01455 + 0.038 + 0.00045 + 0.722} \approx 0.0678
\]
Useful Rule #1: The chain rule

- \( P(\alpha, \beta) = P(\alpha) P(\beta|\alpha) \)

- More generally:

\[
P(\alpha_1, \ldots, \alpha_k) = P(\alpha_1) P(\alpha_2|\alpha_1) \cdots P(\alpha_k|\alpha_1, \ldots, \alpha_{k-1})
\]

- ... any order ...

\[
P(\alpha_1, \ldots, \alpha_k) = P(\alpha_3) P(\alpha_7|\alpha_3) P(\alpha_{14}|\alpha_3, \alpha_7) \cdots
\]
Useful Rule #2: Bayes rule

\[ P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)} \]

More generally, external event \( \gamma \):

\[ P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma)P(\alpha \mid \gamma)}{P(\beta \mid \gamma)} \]
Bayes' Rule and its Use

- **Diagnosis** typically involves computing $P( \text{Hypothesis} \mid \text{Symptoms} )$
  
  What is $P( \text{Meningitis} \mid \text{StiffNeck} )$?
  
  ≡ prob that patient A has meningitis, given that A has stiff neck?

- Typically have . . .
  - Prior prob of meningitis $P( +m ) = 1/50,000$
  - Prior prob of having a stiff neck $P( +s ) = 1/20$
  - Prob that meningitis causes a stiff neck $P( +s \mid +m ) = 1/2$

- Bayes' Rule:

  $$P( M \mid SN ) = \frac{P( SN \mid M ) P( M )}{P( SN )}$$

- Eg: $P( +m \mid +s ) = P(+s \mid +m) P(+m) / P(+s) = 0.5 \times 0.00002 / 0.05 = 0.0002$

- Only 1 in 5000 stiff necks have meningitis... even though SN is major symptom of M...
Factoids

\[ P(\ +c\ ) = \sum_a P(\ +c, A = a ) \]
Important concept: (a) Independence

- Coin tosses:
  - \( H_1 \): the first toss is a head;
  - \( T_2 \): the second toss is a tail
  - \( P(T_2 \mid H_1) = P(T_2) \)

- \( \alpha \) and \( \beta \) independent iff \( P(\beta \mid \alpha) = P(\beta) \)
  - In distribution \( P \), \( \alpha \) indep of \( \beta \)

**Proposition:** \( \alpha \) and \( \beta \) independent if and only if
\[
P(\alpha, \beta) = P(\alpha)P(\beta)
\]
Independence

- Events $\alpha$ and $\beta$ are independent iff
  - $P(\alpha, \beta) = P(\alpha) P(\beta)$
  - $P(\alpha | \beta) = P(\alpha)$
  - $P(\alpha \lor \beta) = 1 - (1 - P(\alpha)) (1 - P(\beta))$

- Variables independent
  $\iff$ independent for all values
  $\forall a, b \; P( A = a, B = b ) = P(A = a) \; P(B = b)$
**Conditional Independence**

- **Reading Ability** and **Shoe Size** are dependent,
  \[ P(\text{ReadAbility} \mid \text{ShoeSize}) \neq P(\text{ReadAbility}) \]

- but become independent, given **Age**
  \[ P(\text{ReadAbility} \mid \text{ShoeSize, Age}) = P(\text{ReadAbility} \mid \text{Age}) \]
Conditional Independence

- Events $A$ and $B$ are conditionally independent given $E$ iff
  \[ P(A \mid E, B) = P(A \mid E) \]

- Given $E$, knowing $B$ does not change the probability of $A$

- Equivalent formulations:
  \[
  P(A, B \mid E) = P(A \mid E) \cdot P(B \mid E) \\
  P(B \mid E, A) = P(B \mid E)
  \]
Probability Theory

- **Axioms:**
  - \( 0 \leq P(A) \leq 1 \)
  - \( P(\text{True}) = 1, \quad P(\text{False}) = 0 \)
  - \( P(A \lor B) = P(A) + P(B) - P(A \land B) \)
  - \( P(A) + P(\neg A) = 1 \)

- **Not arbitrary:**
  - If Agent1 use probabilities that violate axioms, then
    - betting strategy s.t.
      - Agent1 guaranteed to lose $
The Three-Card Problem

Three cards

- RR = red on both sides
- WW = white on both sides
- RW = red on one side, white on the other

Draw single card randomly and toss it into the air.

What is the probability …

a. … of drawing red-red?  $P(D_{RR})$

b. … that the drawn cards lands white side up?  $P(W_{up})$

c. … that the red-red card was not drawn, assuming that the drawn card lands red side up.
   $P(\text{not-D}_{RR} | R_{up})$
A bet is fair to an individual B if,
- according to B’s probability assessment,
- the bet will break even in the long run.

B thinks these 3 bets are fair:

Bet (a): Win $4.20 if \( D_{RR} \);
   lose $2.10 otherwise. [B believes \( P(D_{RR}) = 1/3 \)]

Bet (b): Win $2.00 if \( W_{up} \);
   lose $2.00 otherwise. [B believes \( P(W_{up}) = 1/2 \)]

Bet (c): Win $4.00 if \( R_{up} \) and not \( D_{RR} \);
   lose $4.00 if \( R_{up} \) and \( D_{RR} \);
   win $0 if not-\( R_{up} \).
   [B believes \( P(\text{not-}D_{RR} | R_{up}) = 1/2 \)]

B believes
- \( P( D_{RR} ) = 1/3 \)
- \( P( W_{up} ) = 1/2 \)
- \( P( \text{not-}D_{RR} | R_{up} ) = 1/3 \)
Possible Outcomes

1. \( W_{up} \) & not-\( D_{RR} \): Some card other than RR is drawn, which lands white side up.
2. \( R_{up} \) & not-\( D_{RR} \): Some card other than RR is drawn, which lands red side up.
3. \( R_{up} \) & \( D_{RR} \): RR is drawn, which lands (of course) red side up.

\[ \begin{array}{ccc}
(a) & \text{Win } $4.20 \text{ if } D_{RR}; \\
& \text{lose } $2.10 \text{ otherwise.} \\
(b) & \text{Win } $2.00 \text{ if } W_{up}; \\
& \text{lose } $2.00 \text{ otherwise.} \\
(c) & \text{Win } $4.00 \text{ if } R_{up} \text{ and not } D_{RR}; \\
& \text{lose } $4.00 \text{ if } R_{up} \text{ and } D_{RR}; \\
& \text{win } $0 \text{ if not } R_{up}. \\
\end{array} \]

\( B \) is always guaranteed to lose money…

- whichever card is drawn &
- however it lands!
The Dutch Book Theorem

- Spse B accepts any bet it thinks is fair. Then...

- a Dutch book can be made against B iff

  B's assessment of probability violates Bayesian axiomatization.
Outline

- Probability 101
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Mean, Variance

- Estimation
  - Binomial
  - Gaussian
  - Estimation: MLE, MAP
  - Conjugate Distributions
Expected Value

- **Discrete**

  - \[ E(X) = \sum_x x \cdot P(x) \]
  - \( \approx \) “average”, “mean”, arithmetic mean
  - \( P(X=1) = 1/6, \ P(X=2)=1/6, \ldots, \ P(X=6) = 1/6 \)
  - \[ E[X] = (1 \times 1/6) + (2 \times 1/6) + \ldots + (6 \times 1/6) \]
  - \[ = \ 21/6 \ = \ 3.5 \]

- **Continuous**

  - \[ E(X) = \int_x x \cdot P(x) \, dx \]
Properties of Expectation

\[ E( f(X) ) = \sum_x f(x) P(x) \]

\[ E( aX ) = a \ E( X ) \]
\[ E(aX+b) = a \ E(X) + b \]
\[ E(X + Y) = E(X) + E(Y) \]
\[ E(XY) = ??? \]

If \( X \perp Y \), then \( E(X) \ E(Y) \)
Variance

“How much to trust the mean”
... hard to define in words...

\[
\text{Var}(X) = \mathbb{E}[ (X - \mathbb{E}(X))^2 ] = \mathbb{E}(X^2) - \mathbb{E}(X)^2
\]
Properties of Variance

\[ \text{Var}( f(X) ) = \text{E}[ X - \text{E}(X))^2 ] \]

\[ \text{Var}( aX ) = a^2 \text{Var}(X) \]

\[ \text{Var}(aX+b) = a^2 \text{Var}(X) \]

\[ \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{E}[X-\text{E}(X)) (Y-\text{E}(Y)] \]

If \( X \perp Y \), then \( \text{Var}(X) + \text{Var}(Y) \)
CoVariance

\[ \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{E}[X-\text{E}(X)) (Y-\text{E}(Y)] \]

- CoVariance captures the “leftover”

\[ \text{Cov}(X,Y) = \text{E}[X-\text{E}(X)) (Y-\text{E}(Y)] \]

\[ \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X,Y) \]

- If \( X \perp Y \), then \( \text{Cov}(X, Y) = 0 \)
Standard Deviation

\[ SD(X) = \sqrt{Var(X)} \]

- Sometimes more natural than variance:
  - \( SD(a \times X) = a \times SD(X) \)

- Sometimes, not:
  - \( X \perp Y, \) then \( SD(X + Y) = \sqrt{SD(X)^2 + SD(Y)^2} \)
Outline

- Probability 101
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Mean, Variance

- Estimation
  - Binomial
  - Gaussian
  - Estimation: MLE, MAP
  - Conjugate Distributions
ML involves Estimation

- Consider flipping a Thumbtack. What is the probability it will land with the nail up?

- Try flipping it a few times… observe $H,H,T,T,H$

- What is your BEST GUESS?
Binomial Distribution

- Model:
  - P(Heads) = \(\theta\), \(P(Tails) = 1 - \theta\)
  - Flips are i.i.d.:
    - Independent events
    - Identically distributed according to distribution

- \(P(H, H, T, T, H) = \theta \theta (1 - \theta) (1 - \theta) \theta = \theta^3 (1 - \theta)^2\)

- Sequence \(D\) of \(\alpha_H\) Heads and \(\alpha_T\) Tails:
  \[P(D \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}\]
Maximum Likelihood Estimation

- **Data**: Observed set $D$ of $\alpha_H$ Heads and $\alpha_T$ Tails

- **Hypothesis Space**: Binomial distributions

Learning “best” $\theta$ is an *optimization problem*
- What’s the objective function?

- **MLE**: Choose $\theta$ that maximizes the probability of observed data:

$$
\theta = \arg \max_\theta P(D | \theta) = \arg \max_\theta \ln P(D | \theta)
$$
Simple “Learning” Algorithm

\[ \hat{\theta} = \arg \max_\theta \ln P(\mathcal{D} | \theta) \]

\[ = \arg \max_\theta \ln \theta^\alpha H (1 - \theta)^\alpha T \]

- Set derivative to zero:

\[
\frac{d}{d\theta} \ln P(\mathcal{D} | \theta) = 0
\]

\[
\frac{1}{\theta} \ln [\theta^h (1 - \theta)^t] = \frac{h}{\theta} [h \ln \theta + t \ln (1 - \theta)^t] = \frac{h}{\theta} + \frac{\partial_t}{(1 \partial \theta)}
\]

\[
\frac{h}{\theta} + \frac{-t}{(1 - \theta)} = 0 \quad \hat{\theta} \leftarrow \frac{t}{t + h}
\]

So just average!!!
How many flips are “needed”?

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

- Given 3 heads and 2 tails, \( \theta_{MLE} = \frac{3}{5} = 0.6 \)
- But…
  - Given 30 heads and 20 tails, \( \theta_{MLE} = 0.6 \)
- SAME!!

Which is better? … more precise?
Using Variance

- Variance measures "spread" around mean
- For Binomial(h, t)
  - Mean: $\mu = \frac{h}{h+t}$
  - Variance: $\sigma = \frac{\mu(1-\mu)}{h+t}$
- Binomial(3H, 2T)  $\mu=0.6$  $\sigma=0.048$
- Binomial(30H, 20T)  $\mu=0.6$  $\sigma=0.0048$
Binomial Distribution

$P(D \mid \theta)$ for fixed $\theta = 0.6$

Prob that $p=0.6$ coin generates $k/n$ heads, in $n$ flips
Probability Functions

$P(D | \theta)$ for fixed $D$

Prob that $p=\theta$ coin generates $h$ heads, $t$ tails
Hoeffding’s Equality

Def'n: \[ S_m = \frac{1}{m} \sum_{i=1}^{m} X_i \] be observed average over \( m \) r.v.s in \( \{0, 1\} \)

- \( \Pr[|S_m - \mu| < \lambda] \geq 1 - 2e^{-2m(\lambda/\Gamma)^2} \)

- Holds \( \forall \) (bounded) distributions ... not just Bernoulli...

- Sample average likely to be close to true value as \#samples (\( m \)) increases...
Simple bound
(using Hoeffding’s Inequality)

Here…

- #flips \( m = \alpha_H + \alpha_T \)
- Sample average = \( \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \)
- Let \( \theta^* \) be the true parameter

For any \( \varepsilon > 0 \cdot \)

\[
P\left( | \hat{\theta} - \theta^* | \geq \epsilon \right) \leq 2e^{-2N\epsilon^2}
\]
PAC Learning

- PAC: Probably Approximately Correct

\[ P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2N\epsilon^2} \]

- To know the thumbtack parameter \( \theta \),
  - within \( \epsilon = 0.1 \),
  - with probability \( \geq 1-\delta = 0.95 \)
  
require \#flips \( m > (\ln 2/\delta)/2\epsilon^2 \) \approx 460.2
What about prior knowledge?

- Spse you know the thumbtack $\theta$ is “close” to 50-50
- You can estimate it the Bayesian way…
- Rather than estimate a single $\theta$, obtain a
  
  *distribution over possible values of $\theta$*
Two (related) Distributions: Parameter, Instances

\[ \Theta = 0.1 \]

\[ \Theta = 0.5 \]

\[ \Theta = 0.8 \]

Uniform density
Bayesian Learning

- Use Bayes rule:

\[ P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \]

- Or equivalently (wrt \(\arg\max_{\theta} P(\theta\mid\mathcal{D})\))

\[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]
Bayesian Learning for Thumbtack

\[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta) \]

- Likelihood function is simply Binomial:
  \[ P(\mathcal{D} \mid \theta) = \theta^m_H (1 - \theta)^m_T \]

- What about prior?
  - Represent expert knowledge
  - Simple posterior form

- Conjugate priors:
  - Closed-form representation of posterior (more details soon)
  - **For Binomial, conjugate prior is Beta distribution**
Beta prior distribution – \( P(\theta) \)

- Prior: 
  \[
P(\theta) = \frac{\theta^{\alpha_H-1} (1 - \theta)^{\alpha_T-1}}{B(\alpha_H, \alpha_T)} \sim \text{Beta}(\alpha_H, \alpha_T)
\]

- Likelihood function: 
  \[
P(\mathcal{D} \mid \theta) = \theta^{m_H} (1 - \theta)^{m_T}
\]

- Given \( X \sim \text{Beta}(a, b) \):
  - Mean: \( a/(a + b) \)
  - Unimodal if \( a, b > 1 \)… here mode: \( (a-1) / (a+b-2) \)
  - Variance: \( a b / (a+b)^2 (a+b-1) \)
Posterior distribution… from Beta

\[ P(\theta \mid D) \propto P(\theta) P(D \mid \theta) \]

\[ = \theta^{\alpha_H - 1} (1 - \theta)^{\alpha_T - 1} \times \theta^{m_H} (1 - \theta)^{m_T} \]

\[ = \theta^{\alpha_H + m_H - 1} (1 - \theta)^{\alpha_T + m_T - 1} \]

\[ \sim \text{Beta}(\alpha_M + m_H, \alpha_T + m_T) \]

So Posterior is same form as Prior!! Conjugate!
Posterior Distribution

- Prior: $\theta \sim \text{Beta}(\alpha_H, \alpha_T)$
- Data $\mathcal{D}$: $m_H$ heads, $m_T$ tails
- Posterior distribution:
  $\theta | \mathcal{D} \sim \text{Beta}(m_H + \alpha_H, m_T + \alpha_T)$

Prior
+ observe 1 head
+ observe 27 more heads; 18 tails
Conjugate Prior

- **Given**
  - Prior: \( \Theta \sim \text{Beta}(\alpha_H, \alpha_T) \)
  - Data: \( \mathcal{D} \) with \( m_H \) heads and \( m_T \) tails
    (binomial likelihood)

- **Posterior distribution:**
  \[
  \Theta | \mathcal{D} \sim \text{Beta}(\alpha_H + m_H, \alpha_T + m_T)
  \]

- (Parametric) prior \( P(\theta|\alpha) \) is **conjugate** to likelihood function if **posterior** is of the same parametric family, and can be written as:
  \[
  P(\theta|\alpha'), \quad \text{for some new set of parameters } \alpha'
  \]
Using Bayesian posterior

- Posterior distribution:

\[ P(\theta \mid \mathcal{D}) \sim \text{Beta}(m_H + \alpha_H, m_T + \alpha_T) \]

- Bayesian inference … want \( f(\theta) \)
  - No longer single parameter
  - Can use Expected value:

\[
E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) \, d\theta
\]

… but integral is often hard to compute
MAP: Maximum a posteriori approximation

\[ P(\theta \mid D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- As more data is observed, dist. is more peaked… more of distribution is at MAP:
  \[ \hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta \mid D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]
  
  - Like \( \text{MLE} = \arg \max_\theta P(D \mid \theta) \)
  - but after “observing” prior \( \approx (\beta_H-1, \beta_T-1) \) extra flips

- MAP: use most likely parameter:

\[ E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid D) \, d\theta \approx f(\hat{\theta}_{MAP}) \]
MAP for Beta distribution

\[ P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- **MAP**: use most likely parameter:

  \[ \hat{\theta}_{\text{MAP}} = \arg\max_\theta P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

- Beta prior equivalent to extra thumbtack flips
- As \( N \to \infty \), prior is “forgotten”
- For small sample size, prior is important!
Bayesian prediction of a new thumbtack flip

- Prior: $\Theta \sim \text{Beta}(\alpha_H, \alpha_T)$
- Observed $m_H$ heads, $m_T$ tails
- What is probability that next ($m+1$st) flip is heads?

\[ p_H = \frac{\alpha_H + m_H}{\alpha_H + \alpha_T + m_H + m_T} \]
Alternative “Encoding”

- **Beta( a, b ) \equiv B( m, \mu )**
  - where
    - m = (a+b)  
      - ... effective sample size
    - \mu = a/(a+b)

- Eg...
  - Beta(1,1)  = B( 2, 0.5)
  - Beta(10,10) = B( 20, 0.5)
  - Beta( 7, 3)  = B( 10, 0.7)
  - ...

![Graphs showingBeta distributions](image)
Asymptotic behavior and equivalent sample size

- Beta prior equivalent to extra flips:
  \[ E[\theta] = \frac{m_H + \alpha_H}{m_H + \alpha_H + m_T + \alpha_T} \]

- As \( m \to 1 \), prior is “forgotten”

- But, for small sample size, prior is important!
  \[ E[\theta] = \frac{m_H + \alpha m'}{m_H + m_T + m'} \]

- Equivalent sample size:
  - Prior parameterized by \( \alpha_H, \alpha_T \), or
  - \( m' \) (equivalent sample size) and \( \alpha \)
Bayesian learning $\approx$ Smoothing

$$E[\theta] = \frac{m_H + \alpha m'}{m_H + m_T + m'}$$

- $m=0 \Rightarrow$ prior parameter
- $m \geq 1 \Rightarrow$ MLE

$\alpha m'$
Bayesian learning for \textit{Multinomial}

- What if you have a k-sided thumbtack???
  - … still just ONE thumbtack (so just one event)
- Likelihood function if \textit{multinomial}:
  - $P(X = i) = \theta_i, \quad i = 1..k$
  - $\sum_i \theta_i = 1, \quad \theta_i \geq 0$
- \textbf{Conjugate} prior for multinomial is \textit{Dirichlet}:
  - $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i-1}$

- Observe \textit{m} data points, $m_i$ from assignment i, posterior:
  - $\text{Dirichlet}(\alpha_1 + m_i, \ldots, \alpha_k + m_k)$

- Prediction:
  $$P(X_{m+1} = i | D) = \frac{\alpha_i + m_i}{\sum_j (\alpha_j + m_j)}$$
Learning a Gaussian

- Collect a set of data, $D$ of real-valued i.i.d. instances  
  - e.g., exam scores

- Learn parameters  
  - Mean, $\mu$  
  - Variance, $\sigma$

\[
P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
Multivariate Normal Distributions: A tutorial

- **univariate normal** (Gaussian), with mean $\mu$; variance $\sigma^2$
- PDF (probability distribution function)

$$p(x) = \frac{1}{(2\pi)^{1/2} \sigma} \exp \left[ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right]$$
Some properties of Gaussians

- **Affine transformation (multiplying by scalar and adding a constant)**
  - $X \sim \mathcal{N}(\mu, \sigma^2)$
  - $Y = aX + b \Rightarrow Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

- **Sum of Gaussians**
  - $X \sim \mathcal{N}(\mu_X, \sigma^2_X)$
  - $Y \sim \mathcal{N}(\mu_Y, \sigma^2_Y)$
  - $Z = X + Y \Rightarrow Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma^2_X + \sigma^2_Y)$
The Multivariate Gaussian

A 2-dimensional Gaussian is defined by

- a mean vector \( \mu = [ \mu_1, \mu_2 ] \)
- a covariance matrix:

\[
\Sigma = \begin{bmatrix}
\sigma_{1,1}^2 & \sigma_{2,1}^2 \\
\sigma_{1,2}^2 & \sigma_{2,2}^2
\end{bmatrix}
\]

where \( \sigma_{i,j}^2 = \mathbb{E}[ (x_i - \mu_i) (x_j - \mu_j) ] \) is (co)variance

- Note: \( \Sigma \) is symmetric,

“positive semi-definite”: \( \forall x: x^T \Sigma x \geq 0 \)
Standard Independent Gaussian

- **Standard independent normal:**

  \[ \mu = \langle 0, 0 \rangle \text{ and } \Sigma = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]

Here: \( \Sigma^{-1} = I_2, |\Sigma| = 1; n = 2 \)

\[
P(\langle 3, -2 \rangle \mid N(\langle 0, 0 \rangle, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}))
\]

\[
= \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^\frac{1}{2}} \exp \left[ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right]
\]

\[
= \frac{1}{(2\pi)^{\frac{1}{2}} 1^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} ((3, -2) - \langle 0, 0 \rangle)^\top I_2 ((3, -2) - \langle 0, 0 \rangle) \right]
\]

- \((3, -2) - \langle 0, 0 \rangle)^\top I_2 ((3, -2) - \langle 0, 0 \rangle)\)

\[
= [3, -2] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix}
\]

\[
= (3 \times 3) + (-2 \times -2) = 13
\]

So \(P(\langle -3, 2 \rangle \mid \ldots) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} 13 \right] = \ldots \)
The Multivariate Gaussian: Ex 2

\[ E \mu = \langle 2, 3 \rangle \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \]

\[ P(\langle 3, -2 \rangle \mid \mathcal{N}(\langle 2, 3 \rangle, \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix})) = \frac{1}{(2\pi)^{2/2}2^{1/2}} \exp \left[ \frac{1}{2} (\langle 3, -2 \rangle - \langle 2, 3 \rangle)^\top \Sigma^{-1} (\langle 3, -2 \rangle - \langle 2, 3 \rangle) \right] \]

\[ = \frac{1}{(2\pi)^{2^{1/2}}} \exp \left( \frac{1}{2} \begin{bmatrix} 1 \\ -5 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -5 \end{bmatrix} \right) \]

\[ = \frac{1}{\alpha} \exp \left( -\frac{1}{2} \left[ \frac{1}{2} \times 1^2 + 1 \times (-5)^2 \right] \right) \]
Independent Variables

- Variables independent \( \equiv \) Covariance matrix is Diagonal
  Lines of equal probability \( \equiv \) ellipses parallel to axes

\[
P(\langle x, y \rangle = \langle 3, -2 \rangle | \langle x, y \rangle \sim \mathcal{N}(\langle 0, 0 \rangle, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}))
\]

\[
= P(x = 3 | x \sim \mathcal{N}(0, 1)) \times P(y = -2 | y \sim \mathcal{N}(0, 1))
\]

- \( P(\langle x, y \rangle = \langle 3, -2 \rangle | \langle x, y \rangle \sim \mathcal{N}(\langle 2, 3 \rangle, \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix})) \)

\[
= P(x = 3 | x \sim \mathcal{N}(2, 2)) \times P(y = -2 | y \sim \mathcal{N}(3, 1))
\]
The Multivariate Gaussian: Ex 3

- If $\Sigma$ is arbitrary,
  then $x_1$ and $x_2$ are dependent

Lines of equal probability are “tilted” ellipses

$Eg$ For $\mu = \langle 2, 3 \rangle$ and $\Sigma = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.
Examples of Gaussians

\[ \sum = \alpha I \]
\[ = \text{Diag}(\alpha, \ldots, \alpha) \]

\[ \sum = \text{Diag}(\alpha_1, \ldots, \alpha_k) \]

General \( \sum \)

Marginal…
MLE for Gaussian

- Prob. of i.i.d. instances $D = \{x_1, \ldots, x_N\}$:

$$P(D | \mu, \sigma) = \prod_{i=1}^{N} P(x_i | \mu, \sigma) = \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\ln P(D | \mu, \sigma) = \ln \left[ \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right]$$

$$= -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$
MLE for mean of a Gaussian

What’s ML estimate $\mu_{MLE}$ for mean $\mu$?

$$
\frac{d}{d \mu} \ln P(D | \mu, \sigma) = \frac{d}{d \mu} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\
= -\sum_{i=1}^{N} \frac{d}{d \mu} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right] = \frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(x_i - \mu) = \frac{1}{\sigma^2} \left[ \sum_{i=1}^{N} x_i - N\mu \right]
$$

$$
\frac{d}{d \mu} \ln P(D | \mu, \sigma) = 0 \Rightarrow \left[ \sum_{i=1}^{N} x_i - N\mu \right] = 0
$$

$\Rightarrow \mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$

Just empirical mean!!
MLE for Variance

\[
\frac{d}{d\sigma} \ln P(D | \mu, \sigma) = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

\[
= \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\sigma} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

\[
= -\frac{N}{\sigma} - \sum_{i} \frac{-2(x_i - \mu)^2}{2 \sigma^3}
\]

\[
\Rightarrow \quad \sigma^2_{MLE} = \frac{1}{N} \sum_{i} (x_i - \mu)^2
\]

Just empirical variance!!
\begin{align*}
\mu^*_{MLE} \text{ is unbiased} \\
\text{Estimator } \hat{y} \text{ of } y \text{ is unbiased } \text{ iff } E[\hat{y}] = y \\
\text{Observe } \{ x_1, \ldots, x_n \} \\
\quad \text{drawn iid (independent and identically distributed)} \\
\quad \text{… with common mean } E[x_i] = \mu \\
E[\mu^*_{MLE}] = E\left[\frac{1}{N} \sum_{i=1}^{N} x_i\right] = \frac{1}{N} \sum_{i=1}^{N} E[x_i] = \frac{1}{N} \sum_{i=1}^{N} \mu = \mu
\end{align*}
Learning Gaussian parameters

- **MLE:**
  \[ \hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i \]
  \[ \hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \]

- But… MLE for Gaussian variance is **biased**
  - Expected result of estimation ≠ true parameter!
  - Unbiased variance estimator:
    \[ \hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \]

Homework#1!!
Why is it Biased?

- Bias is wrt Mean; MLE is wrt Mode
  ... Mean ≠ Mode

- Consider...
Estimating a Multivariate Gaussian

- Given data set \(\{x_1, \ldots, x_m\}\), MLE is...

\[
\hat{\mu}_{MLE} = \frac{1}{N} \sum_i x_i
\]

\[
\hat{\Sigma}_{MLE} = \frac{1}{N} \sum_i (x_i - \hat{\mu}) \cdot (x_i - \hat{\mu})^T
\]

- Recall...

\[
x \cdot y^\top = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cdot [y_1 \ y_2 \ y_3] = \begin{bmatrix} x_1 y_1 & x_1 y_2 & x_1 y_3 \\ x_2 y_1 & x_2 y_2 & x_2 y_3 \\ x_3 y_1 & x_3 y_2 & x_3 y_3 \end{bmatrix}
\]
Bayesian learning of Gaussian parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

\[
P(\mu | \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu - \eta)^2}{2\lambda^2}}
\]
MAP for mean of Gaussian

\[
P(\mu \mid D, \sigma, \eta, \lambda) \propto P(D \mid \mu, \sigma) P(\mu \mid \eta, \lambda)
\]

\[
P(D \mid \mu, \sigma) = \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad P(\mu \mid \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}}
\]

\[
\frac{d}{d\mu} \ln P(D \mid \mu) P(\mu) = \frac{d}{d\mu} \ln P(D \mid \mu) + \frac{d}{d\mu} \ln P(\mu)
\]

\[
= - \sum_i \frac{(\mu-x_i)}{\sigma^2} - \frac{(\mu-\eta)}{\lambda^2}
\]

\[
\ldots = 0 \Rightarrow \mu_{\text{MAP}} = \frac{\left(\sum_i \frac{x_i}{\sigma^2}\right) + \frac{\eta}{\lambda^2}}{\left[\frac{N}{\sigma^2} + \frac{1}{\lambda^2}\right]}
\]
MAP for mean of Gaussian

\[
\hat{\mu}_{MAP} = \left[ \left( \sum_i \frac{x_i}{\sigma^2} \right) + \frac{\eta}{\lambda^2} \right] / \left[ \frac{N}{\sigma^2} + \frac{1}{\lambda^2} \right]
\]

- If know nothing, \( \lambda^2 \to \infty \)
  \[\Rightarrow \text{MAP estimate is same as MLE!} \]
- But if \( \lambda^2 < \infty \),
  then MAP is WEIGHTed AVERAGE of MLE and “prior” \( \eta \)
What you need to know

- Probability 101
- Point Estimation
  - MLE
  - Hoeffding inequality (PAC)
  - Bayesian learning
    - Beta, Dirichlet distributions
    - Gaussian, ...
  - MAP
Factoids…

- \( \ln a^b = b \ln a \)
- \( \ln (a \times b) = \ln a + \ln b \)

\[
\frac{\partial}{\partial \ln \theta} \frac{\ln \theta}{\theta} = \frac{1}{\theta}
\]

\[
\frac{\partial}{\partial \ln (1-\theta)} = \frac{\partial 1}{(1 \partial \theta)}
\]
Basic concepts for random variables

- Atomic outcome: assignment $x_1, \ldots, x_n$ to $X_1, \ldots, X_n$

- Conditional probability: $P(X,Y) = P(X) \cdot P(Y|X)$

- Bayes rule: $P(X|Y) = P(Y|X) \cdot P(X) / P(Y)$

- Chain rule:
  
  $P(X_1, \ldots, X_n) = P(X_1) \cdot P(X_2|X_1) \cdots P(X_k|X_1, \ldots, X_{k-1}) \cdots P(X_n|X_1, \ldots, X_{n-1})$
Chebyshev’s Inequality

- $X$ with finite mean, variance

\[ P(|X - E(X)| \geq c) \leq \frac{\text{Var}(X)}{c^2} \]

- Variance governs chance of missing mean
Convergence of Sample Mean

- Apply Chebyshev’s Inequality to sample mean:

\[ P(|X - E(X)| \geq c) \leq \frac{Var(X)}{c^2} \]

\[ Var(\bar{X}) = Var\left(\sum_i \frac{X_i}{n}\right) = \sum_i \frac{1}{n^2} Var(X_i) = \frac{Var(X)}{n} \]

\[ \lim_{n \to \infty} P(|X - E(X)| \geq c) \leq \lim_{n \to \infty} \frac{Var(X)}{nc^2} = 0 \]