

Overview of Probability Theory

Adapted from

[http://www.cs.cmu.edu/~tom/10701_sp11/
slides/Overfitting_ProbReview-1-13-2011-
ann.pdf](http://www.cs.cmu.edu/~tom/10701_sp11/slides/Overfitting_ProbReview-1-13-2011-ann.pdf)

Probability Overview

- **Events**
 - Discrete random variables,
 - continuous random variables,
 - compound events.
- **Axioms of probability**
 - What defines a reasonable theory of uncertainty
- **Independent events**
- **Conditional probabilities**
- **Bayes rule and beliefs**
- **Joint probability distribution**
- **Expectations**
- **Independence, Conditional independence**

Random Variables

- **Informally , A is random variable if**
 - A denotes something about which we are uncertain
 - Perhaps the outcome of a randomized experiment
- **Examples**
 - A = True if randomly drawn person from our class is female
 - A = The hometown of a randomly drawn person from our class
 - A = True if two randomly drawn persons from our classs have same birthday
- **Define P(A)** as “*the fraction of possible worlds in which A s true*” or “*the fraction of times A holds, in repeated runs of the random experiment*”
 - The set of possible worlds is called the sample space, S
 - A random variable A is a function defined over S

$$A : S \rightarrow \{0, 1\}$$

A Little Formalism

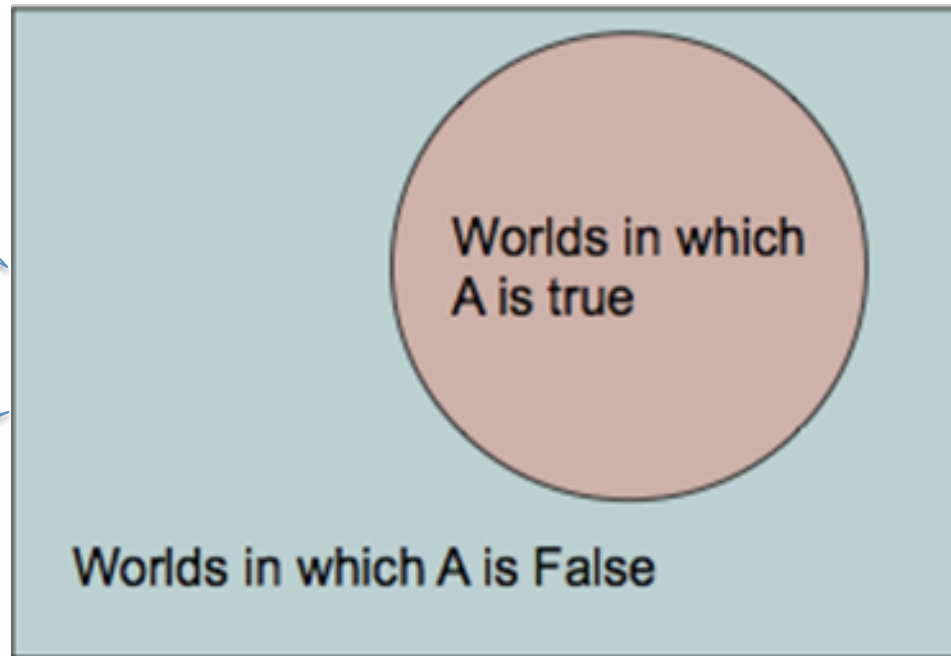
More formally, we have

- A sample space S (e.g., set of students in our class)
 - aka the set of possible worlds
- a random variable is a function defined over the sample space
 - Gender : $S \rightarrow \{m, f\}$
 - Height : $S \rightarrow \text{Reals}$
- an event is subset of S
 - e.g., the subset of S for which ***Gender=f***
 - e.g., the subset of S for which ***(Gender=m) AND (eyeColor=blue)***
- we're often interested in probabilities of specific events
- and of specific events conditioned on other specific events

Visualizing A

Sample space
of all possible
worlds

Its area is 1



$P(A)$ = Area of
reddish oval

The Axioms of Probability

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Interpreting The Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$



The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Interpreting The Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

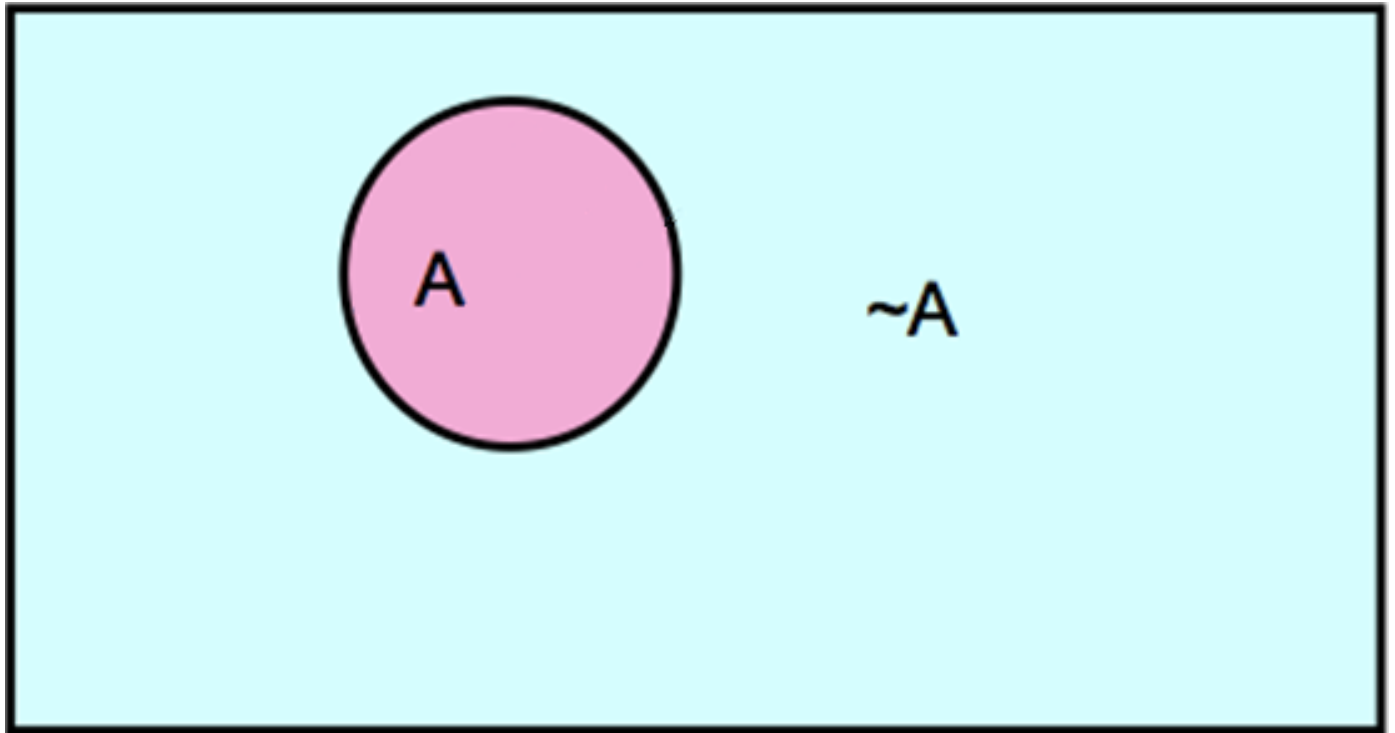


The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true

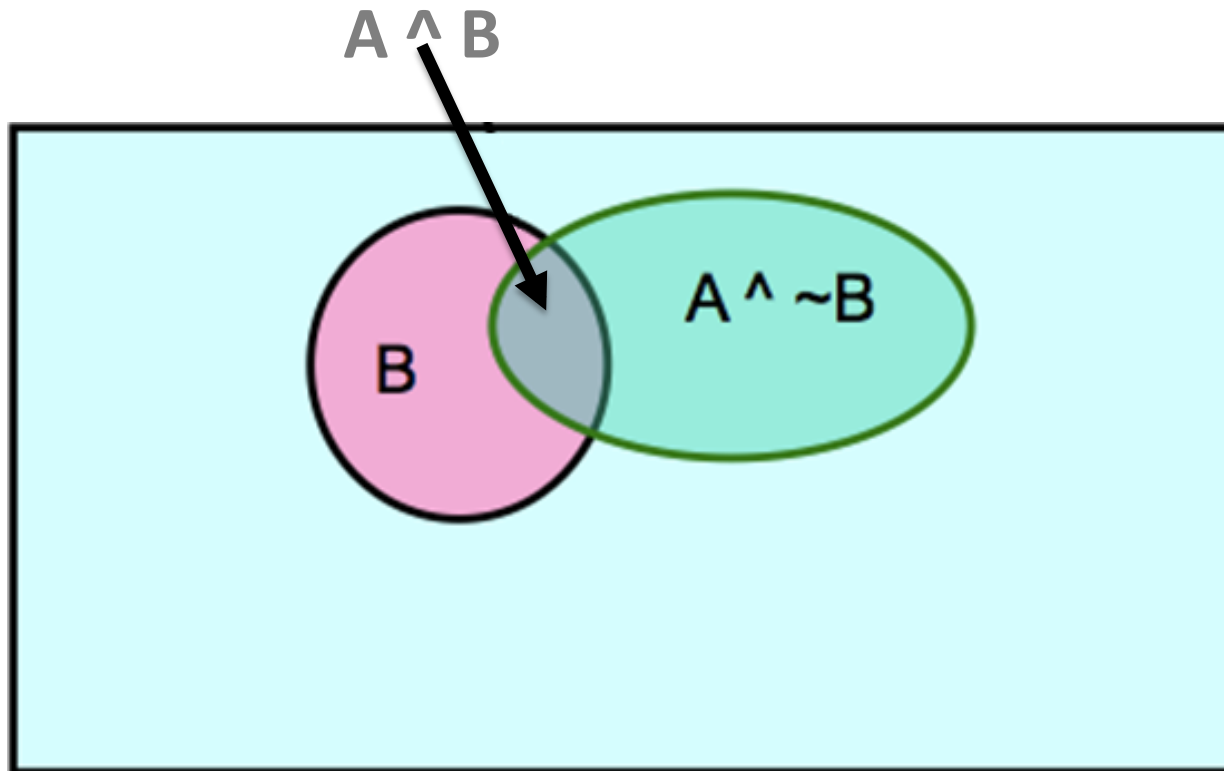
Elementary Probability in Pictures

- $P(\sim A) + P(A) = 1$



Elementary Probability in Pictures

- $P(A) = P(A \wedge B) + P(A \wedge \sim B)$



Multivalued Discrete Random Variables

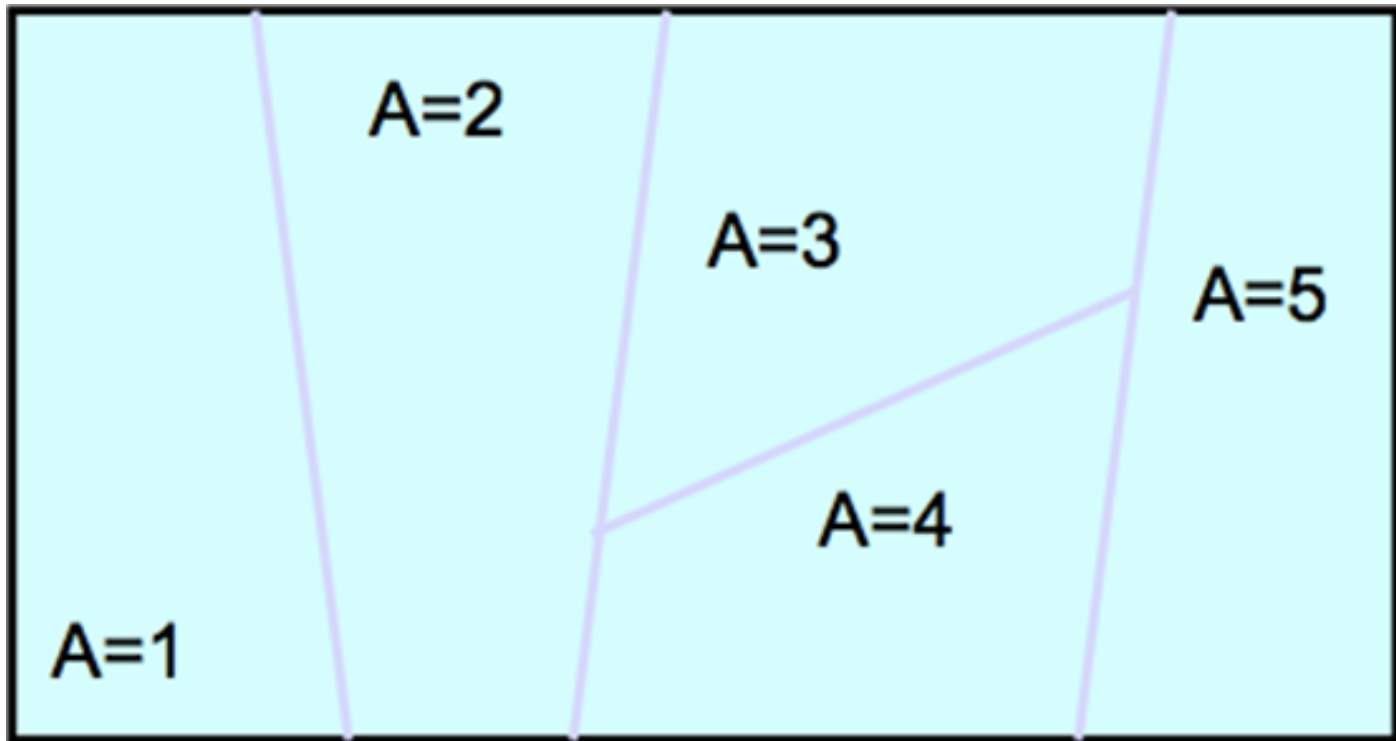
- Suppose A can take on more than 2 values
- A is random variable with arity k if it can take on exactly one value out of $\{v_1, v_2, \dots, v_k\}$
- Thus ...

$$P(A = v_i \wedge A = v_j) = 0 \quad \text{if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \dots \vee A = v_k) = 1$$

Elementary Probability in Pictures

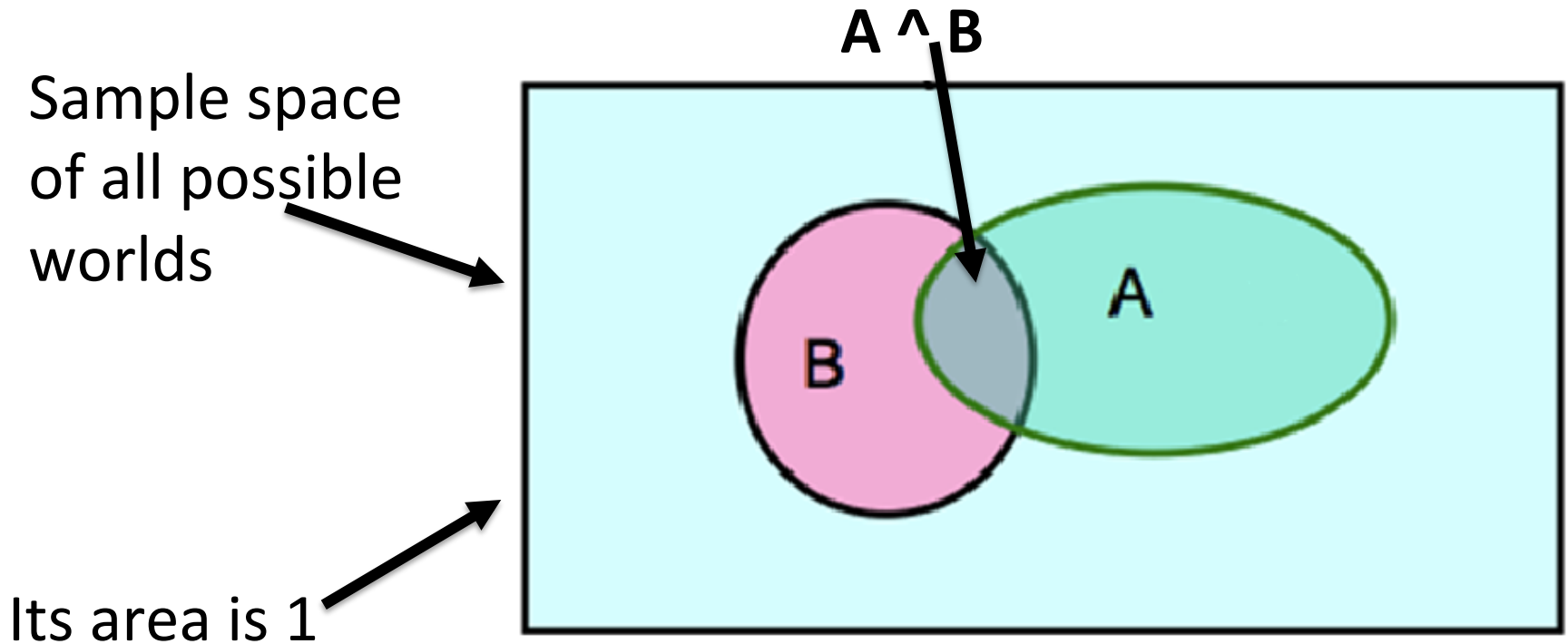
$$\sum_{j=1}^k P(A = v_j) = 1$$



Independent Events

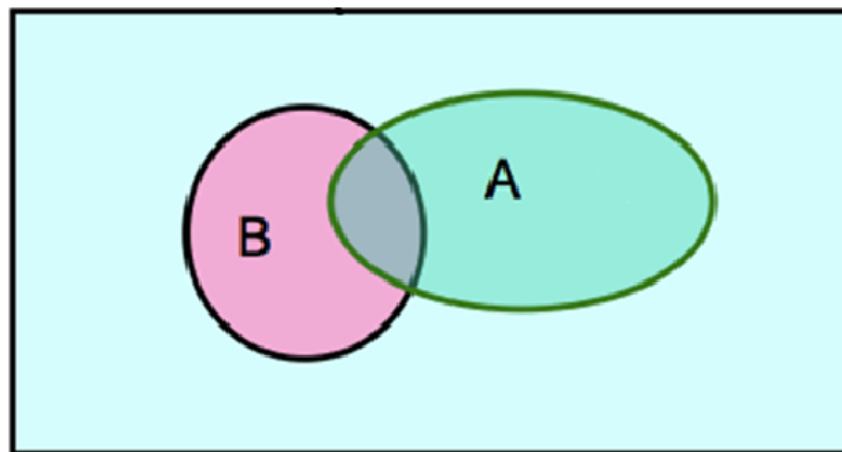
- Definition : two events **A** and **B** are *independent* if **$\Pr(\mathbf{A \text{ and } B}) = \Pr(\mathbf{A}) * \Pr(\mathbf{B})$**
- Intuition : knowing A tells us nothing about the value of B (and vice versa)

Visualizing Probabilities



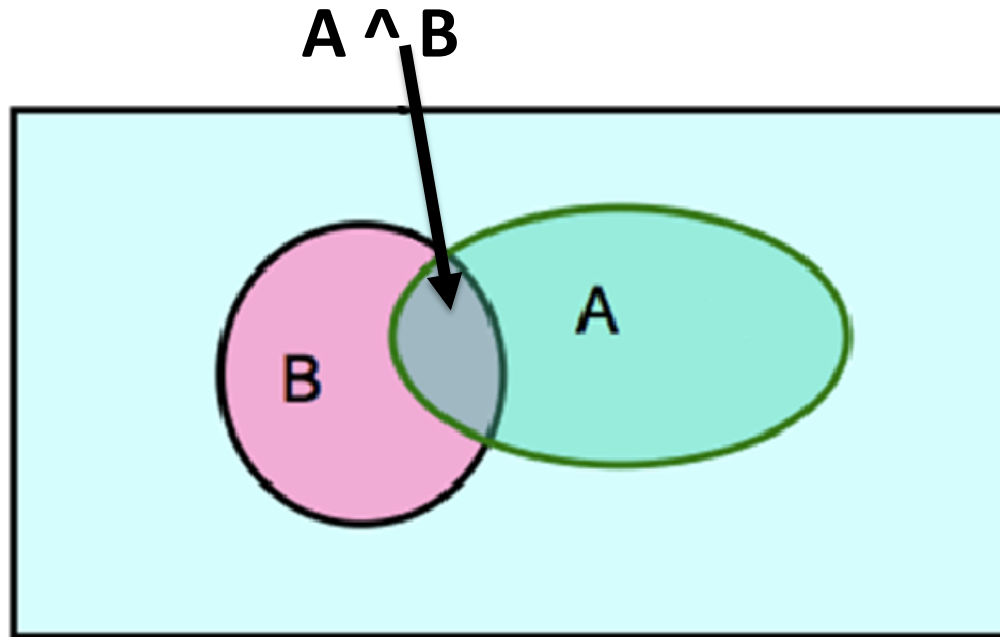
Definition of Conditional Probability

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$



Bayes Rule

- Let's write 2 expressions for $P(A \wedge B)$



$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad \text{Bayes' rule}$$

We call $P(A)$ the “prior”

and $P(A|B)$ the “posterior”



Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370-418

... by no means merely a curious speculation in the doctrine of chance, but necessary to be solved in order to a sure foundation for all our reasoning concerning past facts, and what is likely to be hereafter Necessary to be considered by any that would give a clear account of the strength of analogical or inductive reasoning ...

Definition of Conditional Probability

$$P(A | B) = \frac{P(B | A) * P(A)}{P(B)}$$

Corollary : The Chain Rule

$$P(A \wedge B) = P(A | B) P(B)$$

$$P(C \wedge A \wedge B) = P(C | A \wedge B) P(A | B) P(B)$$

Other Forms of Bayes Rule

$$P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | \sim A)P(\sim A)}$$

$$P(A | B \wedge X) = \frac{P(B | A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

Applying Bayes Rule

- A = You have the flu
- B = You just coughed
- The chance of getting flue is 0.05
- The chance of coughed given you have the flue is 0.8
- The chance of coughed given you do not have the flue is 0.2
- What is the chance of getting the flue given the fact the you just coughed

Applying Bayes Rule

$$P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | \sim A)P(\sim A)}$$

A = you have the flu, B = you just coughed

Assume :

$$P(A) = 0.05$$

$$P(B | A) = 0.80$$

$$P(B | \sim A) = 0.2$$

What is $P(\text{flu} | \text{cough}) = P(A | B)$?

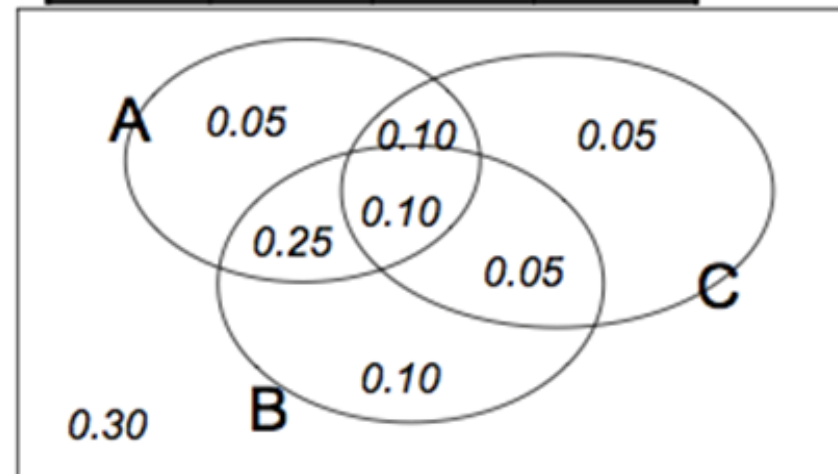
**what does all this have to do with
function approximation ?**

The Joint Distribution

Recipe for making a joint distribution of M variables :

Example : Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



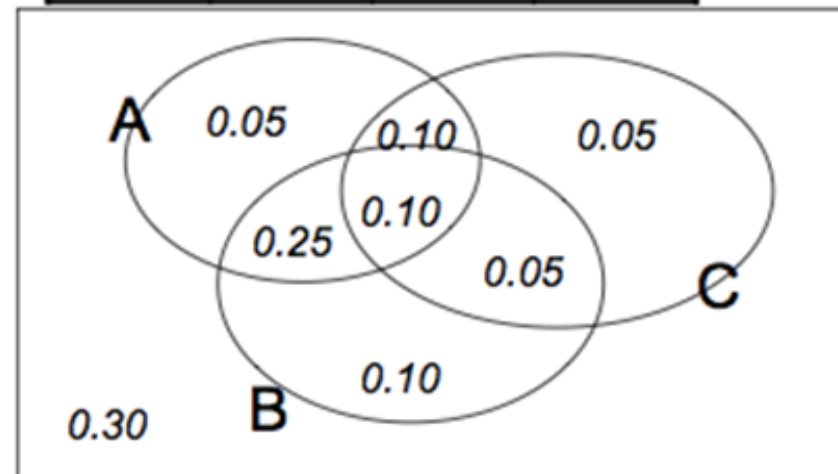
The Joint Distribution

Recipe for making a joint distribution of M variables :

1. 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)

Example : Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



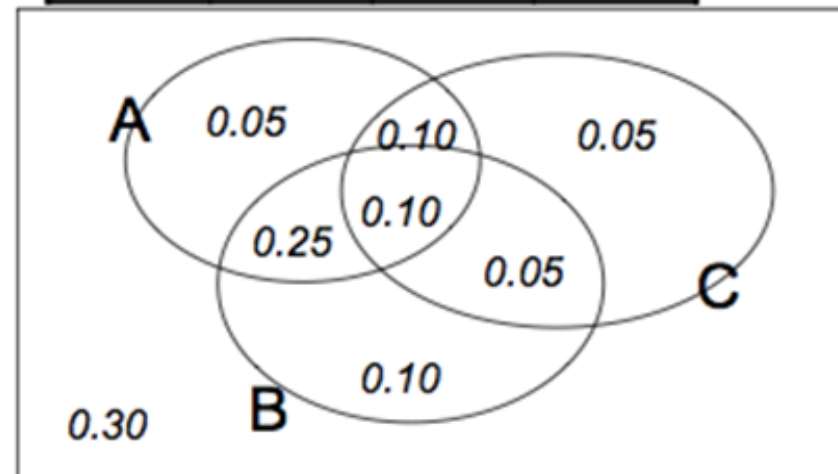
The Joint Distribution

Recipe for making a joint distribution of M variables :

1. 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)
2. For each combination of values, say how probable it is.

Example : Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



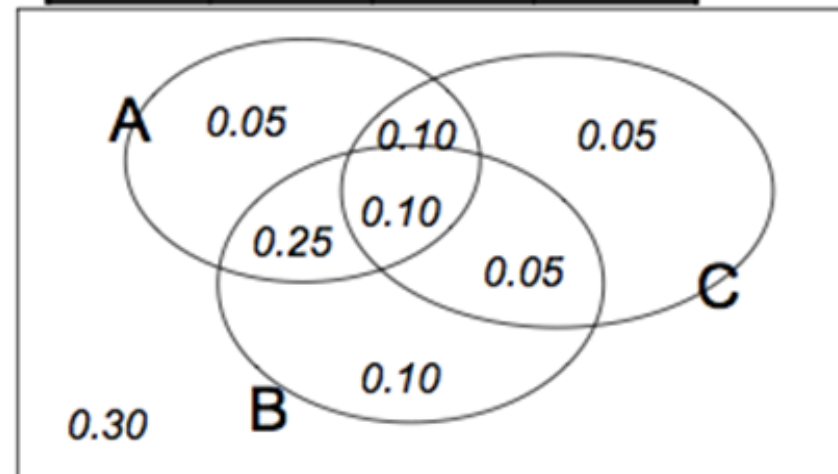
The Joint Distribution

Recipe for making a joint distribution of M variables :

1. 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)
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

Example : Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10











Using the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Once you have the JD
you can ask for the
probability of any logical
expression involving your
attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{ROW})$$



Using the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(ROW)$$

Using the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

$$P(\text{Poor}) = 0.7604$$

$$P(E) = \sum_{\text{rows matching } E} P(\text{ROW})$$

Learning and the Joint Distribution

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Suppose we want to learn the function $f : \langle G, H \rangle \rightarrow W$

Equivalently, $P(W | G, H)$

Solution : learn joint distribution from data, calculate $P(W | G, H)$

e.g., $P(W=\text{rich} \mid G = \text{female}, H = 40.5-) =$

You should know

- Event
 - Discrete random variables, continuous random variables, compound events
- Axioms of probability
 - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs

**sounds like the solution to
learning $F: X \rightarrow Y$,
Or $P(Y|X)$.**

Are we done ?

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question :
 - He says : I have thumbtack, if I flip it, what's the probability it will fall the nail up ?
 - You say : Please flip it a few times :
 - You say : The probability is :
 - **He says : Why ???**
 - You say : Because

Thumbtack - Binomial Distribution

- $P(\text{Heads}) = \theta$, $P() = 1 - \theta$
- Flips are i.i.d. :
 - independent events
 - Identically distributed according to Binomial distribution
- Sequence D of α_H Heads and α_T Tails.

$$P(D \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- Data : Observed set D of α_H Heads and α_T Tails.
- Hypothesis : Binomial distribution
- Learning θ is an optimization problem
 - What's the objective function ?
- MLE : choose θ that maximize the probability of observed data :

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(D | \theta) \\ &= \arg \max_{\theta} \ln P(D | \theta)\end{aligned}$$

Maximum Likelihood For θ

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(D | \theta) \\ &= \arg \max_{\theta} \ln P(D | \theta)\end{aligned}$$

- Set derivation to zero : $\frac{d}{d\theta} \ln P(D | \theta) = 0$

Maximun Likelihood For θ

- Set derivation to zero : $\frac{d}{d\theta} \ln P(D | \theta) = 0$

$$\hat{\theta} = \arg \max_{\theta} P(D | \theta)$$

$$= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

How many flips do I need ?

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

Bayesian Learning

Use Bayes rule :

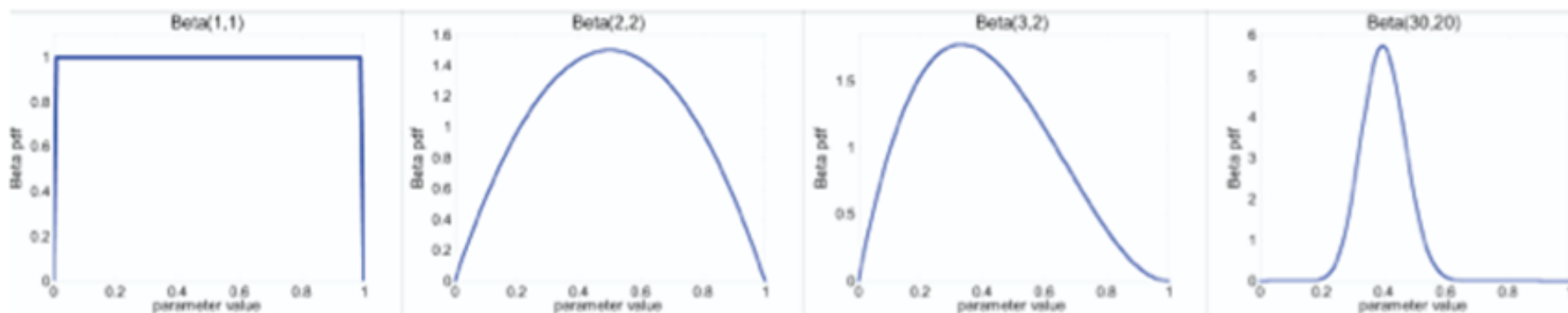
$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

Or equivalently

$$P(\theta | D) \propto P(D | \theta)P(\theta)$$

Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

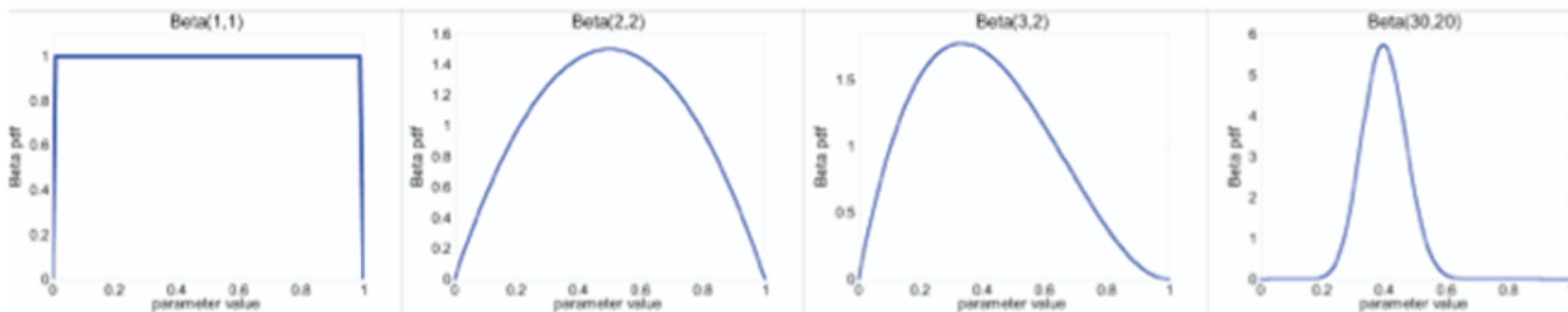


- Likelihood function : $P(D | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$
- Posterior : $P(\theta | D) \propto P(D | \theta)P(\theta)$









Posterior distribution

- Prior : $Beta(\beta_H, \beta_T)$
- Data : α_H heads and α_T tails
- Posterior distribution:

$$P(\theta | D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



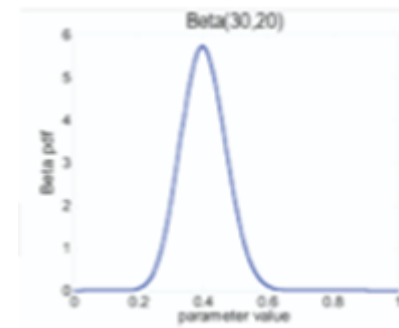
Inference with the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

$$P(\text{Male} | \text{Poor}) = 0.4654 / 0.7604 = 0.612$$

MAP for Beta distribution



$$P(\theta | D) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- **MAP : use most likely parameter**

$$\hat{\theta} = \arg \max_{\theta} P(\theta | D) =$$

- **Beta prior equivalent to extra thumbtack flips**
- **As $N \rightarrow \infty$, prior is “forgotten”**
- **But, for small sample size, prior is important!**

Dirichlet distribution

- **Number of heads in N flips of a two-sided coin**
 - Follows a binomial distribution
 - Beta is a good prior (conjugate prior for binomial)
- **What its' not two-sided, but k-sided?**
 - Follows a multinomila distribution
 - Dirichlet distribution is the conjugate prior

$$P(\theta_1, \theta_2, \dots, \theta_K) = \frac{1}{B(\alpha)} \prod_i^K \theta_i^{(\alpha_i - 1)}$$

Lejeune Dirichlet



Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	 Germany
Nationality	 German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

Estimating Parameters

- **Maximum Likelihood Estimate (MLE)** : choose θ that maximizes probability of observed data D

$$\hat{\theta} = \arg \max_{\theta} P(D | \theta)$$

- **Maximum a Posteriori (MAP) estimate** : choose θ that is most probable given prior probability and the data

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(D | \theta) \\ &= \arg \max_{\theta} = \frac{P(D | \theta)P(\theta)}{P(D)}\end{aligned}$$

You should know

- **Probability basics**

- random variables, events, sample space, conditional probs, ...
- Independence of random variables
- Bayes rule
- Joint probability distributions
- Calculating probabilities from the joint distribution

- **Point estimation**

- Maximum likelihood estimates
- Maximum a posteriori estimates
- Distributions – binomial, Beta, Dirichlet, ...

Expected values

- Given discrete random variable X , the expected value of X , written $E[X]$ is

$$E[X] = \sum_{x \in X} xP(X = x)$$

- We also can talk about the expected value offunction of X

$$E[f(X)] = \sum_{x \in X} f(x)P(X = x)$$

Covariance

- Given two discrete r.v.'s X and Y , we define the covariance of X and Y as

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

e.g., X = gender, Y =plays Football

Or X = gender, Y =leftHanded

Remember $E[X] = \sum_{x \in X} xP(X = x)$

Example : Bernoulli model

- **Data :**
 - We observed N iid coin tossing : $D=\{1,0,1,...,0\}$
- **Representation**
 - Binary r.v. : $x_n=\{0,1\}$
- **Model :**

$$P(x) = \begin{cases} 1-\theta & \text{for } x=0 \\ \theta & \text{for } x=1 \end{cases} \Rightarrow P(x) = \theta^x (1-\theta)^{1-x}$$

- **How to write the likelihood of a single observation**

$$P(x_i) = \theta^{x_i} (1-\theta)^{1-x_i}$$

- **The likelihood of dataset $D=\{X_1, \dots, X_N\}$**

$$P(x_1, x_2, \dots, x_K \mid \theta) = \prod_{i=1}^N \left(\theta^{x_i} (1-\theta)^{1-x_i} \right) = \theta^{\sum_{i=1}^N x_i} (1-\theta)^{\sum_{i=1}^N 1-x_i} = \theta^{\alpha_{head}} (1-\theta)^{\alpha_{tails}}$$

You should know

- **Probability basics**

- random variables, events, sample space, conditional probs, ...
- Independence of random variables
- Bayes rule
- Joint probability distributions
- Calculating probabilities from the joint distribution

- **Point estimation**

- Maximum likelihood estimates
- Maximum a posteriori estimates
- Distributions – binomial, Beta, Dirichlet, ...