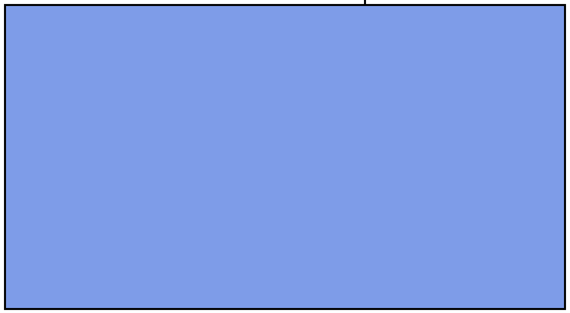
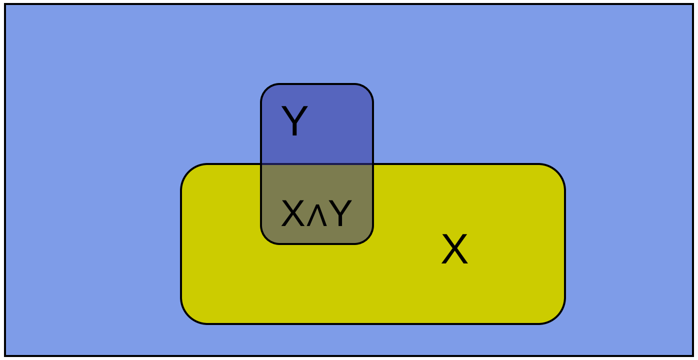
Elementary manipulations of probabilities

- Set probability of multi-valued r.v.



  - $P(\{x=Odd\}) = P(1)+P(3)+P(5) = 1/6+1/6+1/6 = ½$
  - $$P(X = x_1 \vee X = x_2, \ldots, \vee X = x_i) = \sum_{j=1}^{i} P(X = x_j)$$

- Multi-variant distribution:



  - **Joint probability**: $P(X = true \wedge Y = true)$

$$P(Y \wedge \{X = x_1 \vee X = x_2, \ldots, \vee X = x_i\}) = \sum_{j=1}^{i} P(Y \wedge X = x_j)$$

  - **Marginal Probability:** $P(Y) = \sum_{j \in S} P(Y \wedge X = x_j)$

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:

	B	¬B
F	0.005	0.02
¬F	0.195	0.78

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

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

Conditional Probability

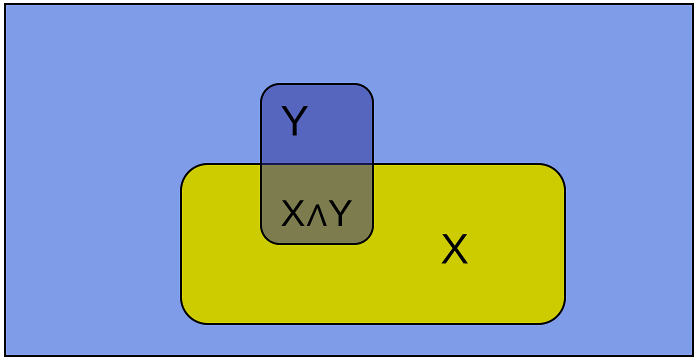
- P(X|Y) = Fraction of worlds in which X is true that also have Y true
  - H = "having a headache"
  - F = "coming down with Flu"
    - P(H)=1/10
    - P(F)=1/40
    - P(H|F)=1/2
  - P(H|F) = fraction of flu-inflicted worlds in which you have a headache

    = P(H∧F)/P(F)
- Definition:

$$P(X \mid Y) = \frac{P(X \wedge Y)}{P(Y)}$$

  - Corollary: The Chain Rule

$$P(X \wedge Y) = P(X \mid Y)P(Y)$$



MLE

- Objective function:

$$\ell(\theta; D) = \log P(D \mid \theta) = \log \theta^{n_h} (1-\theta)^{n_t} = n_h \log \theta + (N - n_h) \log(1-\theta)$$

- We need to maximize this w.r.t. θ
- Take derivatives wrt θ

$$\frac{\partial \ell}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1-\theta} = 0 \quad \Rightarrow \quad \widehat{\theta}_{MLE} = \frac{n_h}{N} \quad \text{or} \quad \widehat{\theta}_{MLE} = \frac{1}{N} \sum_i x_i$$

Frequency as sample mean

- Sufficient statistics
  - The counts, n_h, where $n_k = \sum_i x_i$, are **sufficient statistics** of data D

The Bayes Rule

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

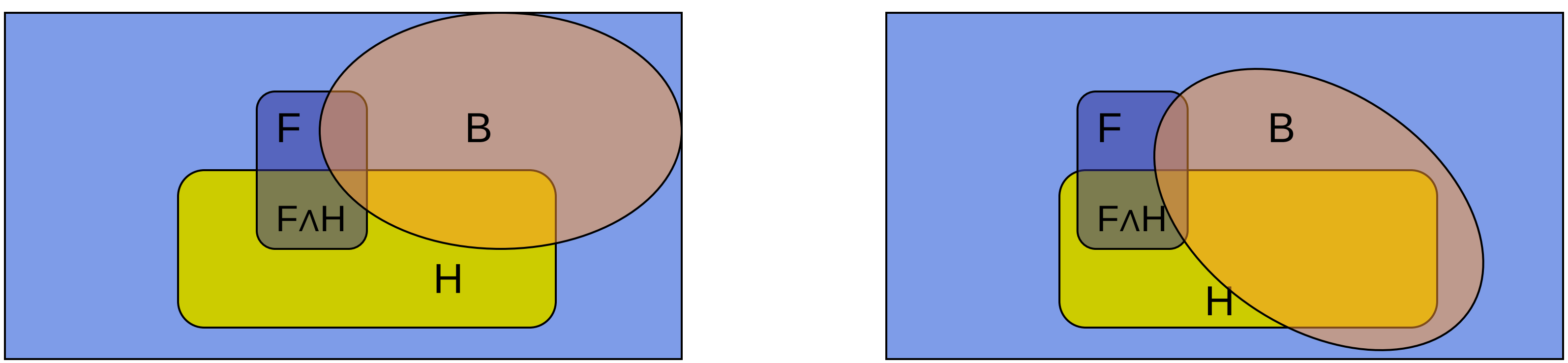
$$P(y \mid X) = \frac{P(X \mid y)\,p(y)}{P(X)}$$

This is Bayes Rule

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**

More General Forms of Bayes Rule

- $$P(Y|X) = \frac{P(X|Y)p(Y)}{P(X|Y)p(Y) + P(X|\neg Y)p(\neg Y)}$$
- $$P(Y = y_i | X) = \frac{P(X|Y)p(Y)}{\sum_{i \in S} P(X|Y = y_i)p(Y = y_i)}$$
- $$P(Y|X \wedge Z) = \frac{P(X|Y \wedge Z)p(Y \wedge Z)}{P(X \wedge Z)} = \frac{P(X|Y \wedge Z)p(Y \wedge Z)}{P(X|\neg Y \wedge Z)p(\neg Y \wedge Z) + P(X|\neg Y \wedge Z)p(\neg Y \wedge Z)}$$
  - P(Flu | Headhead ∧ DrankBeer)



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”

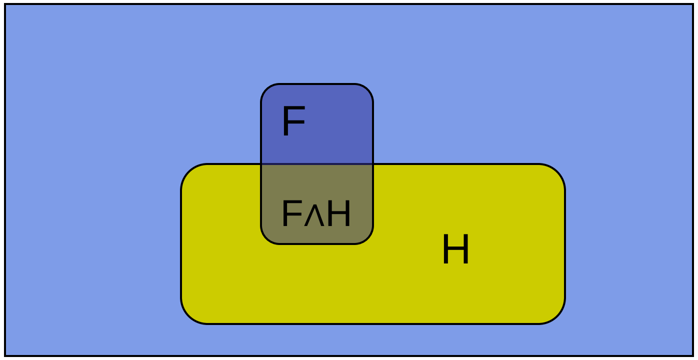
  Is this reasoning correct?

Probabilistic Inference

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

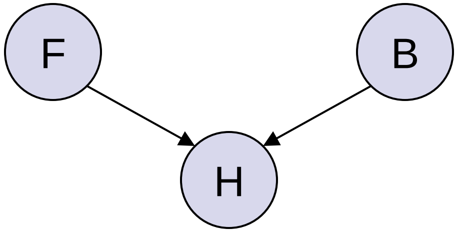
- The Problem:

P(F|H) = ?



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 =true*) = 0.025 and *P*(*DrinkBeer =true*) = 0.2

  correspond to belief prior to arrival of any (new) evidence
- A probability distribution gives values for all possible assignments:
  - **P**(*DrinkBeer*) =[0.01,0.09*,* 0.1, 0.8]
  - (normalized, i.e., sums to 1)

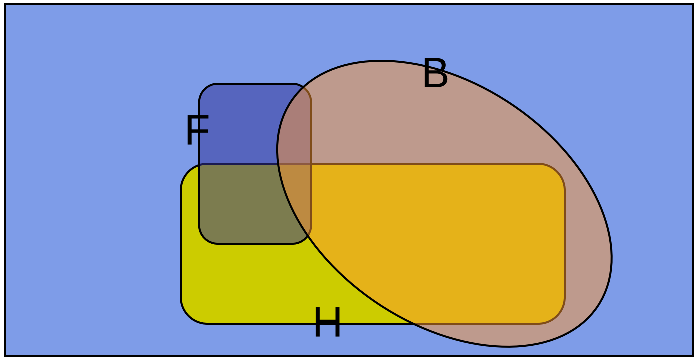
Posterior conditional probability

- 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:
  - $\mathbf{P}(Flu|Headache)$ = 2-element vector of 2-element vectors
- If we know more, e.g., DrinkBeer is also given, then we have
  - $\mathbf{P}(Flu|Headache,DrinkBeer) = 0.070$ **This effect is known as explain away!**
  - $\mathbf{P}(Flu|Headache,Flu) = 1$
  - 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.,
  - $\mathbf{P}(Flu|Headache,StealerWin) = \mathbf{P}(Flu|Headache)$
  - 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

F	B	H	Prob
0	0	0	0.4
0	0	1	0.1
0	1	0	0.17
0	1	1	0.2
1	0	0	0.05
1	0	1	0.05
1	1	0	0.015
1	1	1	0.015

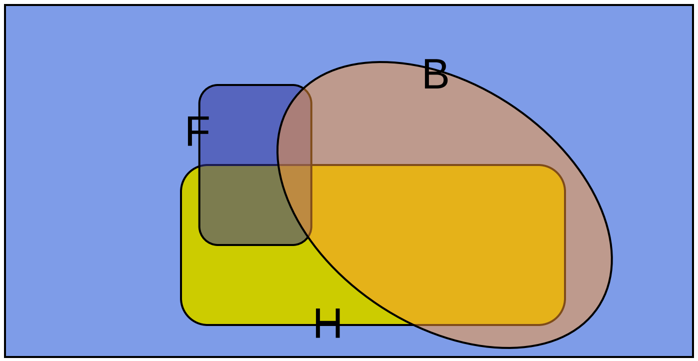


Inference with the Joint

- One you have the JD you can ask for the probability of any atomic event consistent with you query

$$P(E) = \sum_{i \in E} P(row_i)$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

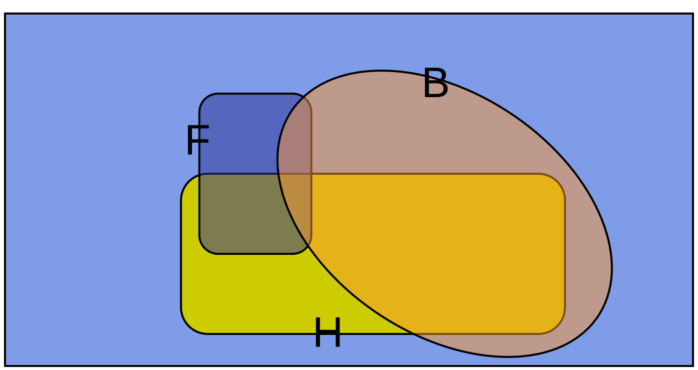


Inference with the Joint

- Compute Marginals

$$P(\text{Flu} \wedge \text{Headache}) =$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

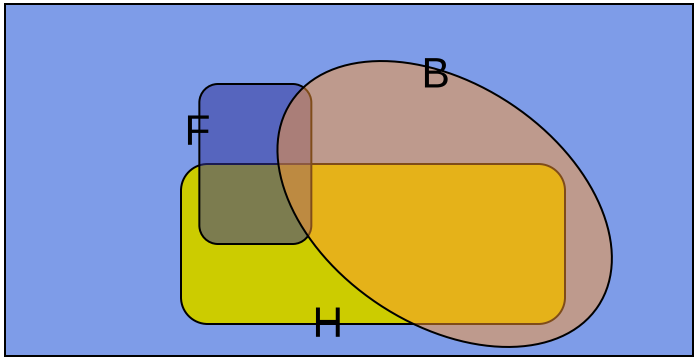


Inference with the Joint

- Compute Marginals

$$P(\text{Headache}) =$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015



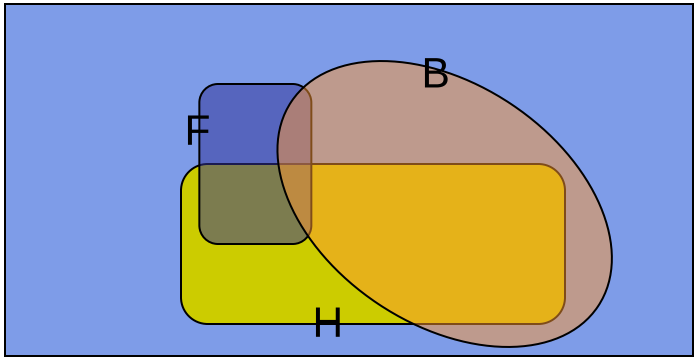
Inference with the Joint

- Compute Conditionals

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)}$$

$$= \frac{\sum_{i \in E_1 \cap E_2} P(row_i)}{\sum_{i \in E_2} P(row_i)}$$

¬F	¬B	¬H	0.4	
¬F	¬B	H	0.1	
¬F	B	¬H	0.17	
¬F	B	H	0.2	
F	¬B	¬H	0.05	
F	¬B	H	0.05	
F	B	¬H	0.015	
F	B	H	0.015	



Inference with the Joint

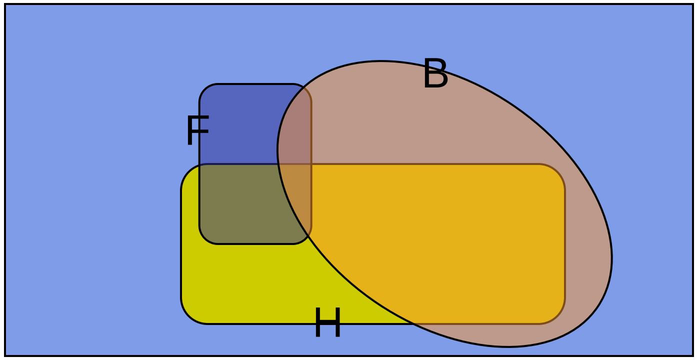
- Compute Conditionals

$$P(\text{Flu} \mid \text{Headhead}) = \frac{P(\text{Flu} \wedge \text{Headhead})}{P(\text{Headhead})}$$

$$=$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

- 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:

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

- 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(Headache;Flu;Virus;DrinkBeer)

  = P(Headache | Flu;Virus;DrinkBeer) P(Flu;Virus;DrinkBeer)

  = P(Headache | Flu;Virus;DrinkBeer) P(Flu | Virus;DrinkBeer) P(Virus | DrinkBeer) P(DrinkBeer)

  Assume independence and conditional independence

  = **P(Headache|Flu;DrinkBeer) P(Flu|Virus) P(Virus) P(DrinkBeer)**

  I.e., ? independent parameters

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from **exponential** in *n* to **linear** in *n*.
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Rules of Independence --- by examples

- P(Virus | DrinkBeer) = **P(Virus)**

  iff Virus is independent of DrinkBeer

- P(Flu | Virus;DrinkBeer) = **P(Flu|Virus)**

  iff Flu is independent of DrinkBeer, given Virus

- P(Headache | Flu;Virus;DrinkBeer) = **P(Headache|Flu;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 $P(E|H)$, 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

  $$P(E) = P(E|H)$$

  (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

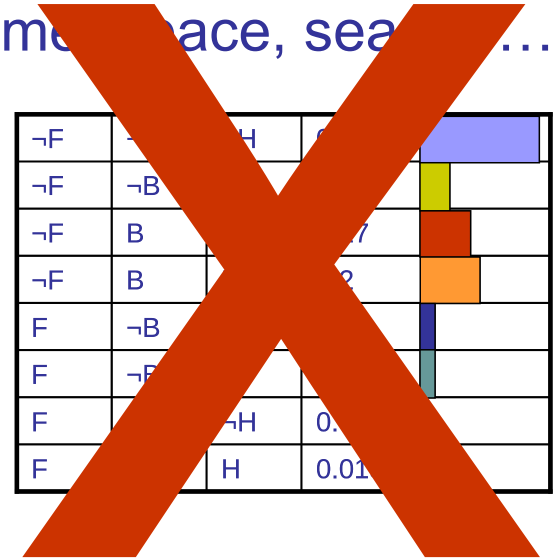
  $$P(E|H,F) = P(E|H)$$

  or equivalently

  $$P(E,F|H) = P(E|H)P(F|H)$$

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|¬F,B)

P(H|F,¬B)

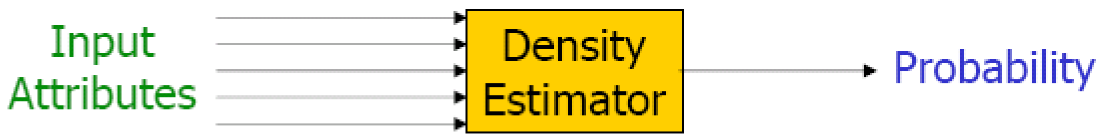
…

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

- 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



- 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, …)
  - Evaluation scheme (likelihood on test data, predictability, consistency, …)

Parameter Learning from iid data

- Goal: estimate distribution parameters θ 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:

$$\begin{aligned} L(\theta) &= P(x_1, x_2, \ldots, x_N; \theta) \\ &= P(x;\theta)P(x_2;\theta), \ldots, P(x_N;\theta) \\ &= \prod_{i=1}^{N} P(x_i;\theta) \end{aligned}$$

  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, …, 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 $D=\{x_1, \ldots, x_N\}$:

$$P(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \left(\theta^{x_i} (1-\theta)^{1-x_i} \right) = \theta^{\sum_{i=1}^{N} x_i} (1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}} (1-\theta)^{\#\text{tails}}$$

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!

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

Example 2: univariate normal

- Data:
  - We observed N ***iid*** real samples:

    D={-0.1, 10, 1, -5.2, …, 3}
- Model: $$P(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-(x-\mu)^2/2\sigma^2\right\}$$
- Log likelihood:

$$\ell(\theta; D) = \log P(D \mid \theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{\sigma^2}$$

- MLE: take derivative and set to zero:

$$\frac{\partial \ell}{\partial \mu} = (1/\sigma^2)\sum_n (x_n - \mu)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_n (x_n-\mu)^2$$

$\Rightarrow$

$$\mu_{MLE} = \frac{1}{N}\sum_n (x_n)$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_n (x_n - \mu_{ML})^2$$

Overfitting

- Recall that for Bernoulli Distribution, we have

$$\hat{\theta}_{ML}^{head} = \frac{n^{head}}{n^{head} + n^{tail}}$$

- What if we tossed too few times so that we saw zero head?

  We have $\hat{\theta}_{ML}^{head} = 0,$ 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

$$\hat{\theta}_{ML}^{head} = \frac{n^{head} + n'}{n^{head} + n^{tail} + n'}$$

  - But can we make this more formal?

The Bayesian Theory

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

$$P(M|D) = P(D|M)P(M)/P(D)$$

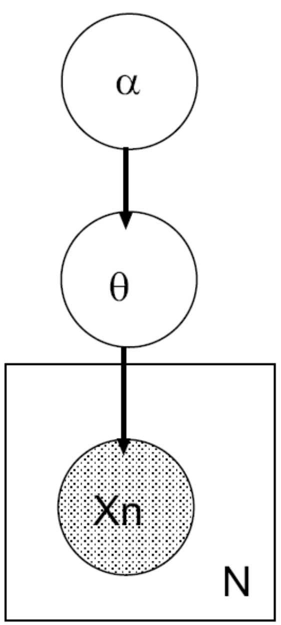
  - the **posterior** equals to the **likelihood** times the **prior**, up to a constant.

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

Hierarchical Bayesian Models

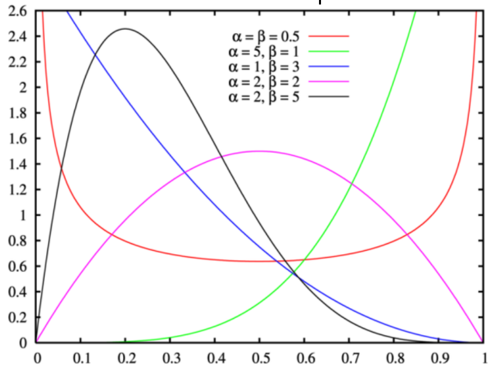
- θ are the parameters for the likelihood $p(x|\theta)$
- α are the parameters for the prior $p(\theta|\alpha)$.
- We can have hyper-hyper-parameters, etc.
- We stop when the choice of hyper-parameters makes no difference to the marginal likelihood; typically make hyper-parameters constants.
- Where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

$$\widehat{\vec{\alpha}}_{MLE} = \arg\max_{\vec{\alpha}} = p(\vec{n} \mid \vec{\alpha})$$



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 θ:

$$P(\theta\mid x_1,\ldots,x_N)=\frac{p(x_1,\ldots,x_N\mid\theta)\,p(\theta)}{p(x_1,\ldots,x_N)}\propto\theta^{n_h}(1-\theta)^{n_t}\times\theta^{\alpha-1}(1-\theta)^{\beta-1}=\theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

  - 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 θ:

$$P(\theta \mid x_1,\ldots,x_N) = \frac{p(x_1,\ldots,x_N \mid \theta)\, p(\theta)}{p(x_1,\ldots,x_N)} \propto \theta^{n_h}(1-\theta)^{n_t} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

- Maximum *a posteriori* (MAP) estimation:

$$\theta_{MAP} = \arg\max_{\theta} \log P(\theta \mid x_1,\ldots,x_N)$$

- Posterior mean estimation:

Bata parameters can be understood as pseudo-counts

$$\theta_{Bayes} = \int \theta p(\theta \mid D)\, d\theta = C \int \theta \times \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}\, d\theta = \frac{n_h+\alpha}{N+\alpha+\beta}$$

- Prior strength: A=α+β
  - A can be interoperated 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 $\vec{n}=(n_h=2, n_t=8)$
- Weak prior A = 2. Posterior prediction:

$$p(x=h \mid n_h=2, n_t=8, \vec{\alpha}=\vec{\alpha}'\times 2)=\frac{1+2}{2+10}=0.25$$

- Strong prior A = 20. Posterior prediction:

$$p(x=h \mid n_h=2, n_t=8, \vec{\alpha}=\vec{\alpha}'\times 20)=\frac{10+2}{20+10}=0.40$$

- However, if we have enough data, it washes away the prior. e.g., $\vec{n}=(n_h=200, n_t=800)$. 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:

$$P(\mu) = \left(2\pi\tau^2\right)^{-1/2} \exp\left\{-(\mu-\mu_0)^2/2\tau^2\right\}$$

- Joint probability:

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

- Posterior:

$$P(\mu\,|\,x) = \left(2\pi\tilde{\sigma}^2\right)^{-1/2} \exp\left\{-(\mu-\tilde{\mu})^2/2\tilde{\sigma}^2\right\}$$

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

Sample mean