Probability Theory Review

Brendan O'Connor

10-601 Recitation

Sept 11 & 12, 2012

Mathematical Tools for Machine Learning

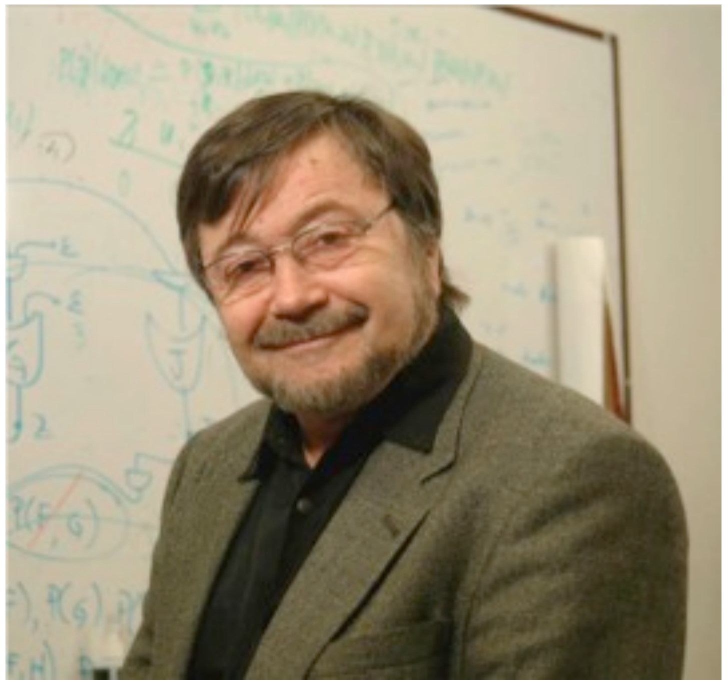
- Probability Theory
- Linear Algebra
- Calculus

- Wikipedia is great reference

Probability Theory in AI

- Fundamental problem: Reasoning about uncertainty
- "Uncertainty in Artificial Intelligence": history of non-probabilistic approaches.
  fuzzy logic, nonmonotonic reasoning, Dempster-Shafer, three-valued logics...
- Pearl 1988 (among others): probability theory is the way to go
- Probabilitistic approaches completely dominate modern machine learning

Judea Pearl
2011 Turing Award winner



Why Probability Theory?

- "Dutch book"
  - If you violate these rules, you will lose money.
- Kolmogorov Axioms
- Cox Axioms
  - Probabilities of **boolean variables**. Logic.

Bruno de Finetti
1906-1985

Andrey Kolmogorov
1903-1987

Cox axioms

- Probability theory as an extension of boolean logic.
- Let B(X | Z) be a *degree of belief* function.
  - Belief in proposition *X*
  - Conditional on background knowledge *Z*
- Assume
  1. Degrees-of-belief are well-ordered
  2. B(~X) = f[B(X)]
  3. B(X and Y) = g(B(X), B(Y | X))

note controversies -- see references on http://en.wikipedia.org/wiki/Cox's_theorem

Cox axioms

- Probability theory as an extension of boolean logic.
- Let B(X | Z) be a *degree of belief* function.
  - Belief in proposition *X*
  - Conditional on background knowledge *Z*
- Assume
  1. Degrees-of-belief are well-ordered
  2. B(~X) = f[B(X)]
  3. B(X and Y) = g(B(X), B(Y | X))
- Then can map B(.) to P(.) where
  1. P(true) = 1
  2. P(false) = 0
  3. P(X and Y) = P(X) P(Y | X)
  4. P(~X) = 1 - P(X)

note controversies -- see references on http://en.wikipedia.org/wiki/Cox's_theorem

Cox axioms

- Probability theory as an extension of boolean logic.
- Let B(X | Z) be a *degree of belief* function.
  - Belief in proposition *X*
  - Conditional on background knowledge *Z*
- Assume
  1. Degrees-of-belief are well-ordered
  2. B(~X) = f[B(X)]
  3. B(X and Y) = g(B(X), B(Y | X))
- Then can map B(.) to P(.) where
  1. P(true) = 1
  2. P(false) = 0
  3. P(X and Y) = P(X) P(Y | X)
  4. P(~X) = 1 - P(X)

1. G(true) = 0
2. G(false) = -infinity
3. G(X and Y) = G(X) + G(Y | X)
4. G(~X) = log(1 - exp[G(X)])

note controversies -- see references on http://en.wikipedia.org/wiki/Cox's_theorem

Many views of probability

- Meaning of a probability
  - Epistemic: knowledge/belief; vs
  - Long-term frequencies
- Related: Bayesian vs Frequentist statistics

Discrete Random Variables

- Views of RVs: (1) outcome of a random process, or (2) has a true value that you don't know
- Binary (boolean) variable
  - Values(A) = {0, 1}
  - Values(A) = {-1, +1}
  - Values(A) = {☼, ☾}
- Categorical variable
  - Values(A) = {0, 1, 2}
  - Values(A) = {55, 32, 10}
  - Values(A) = {♧, ♡, ♢}

Probability (Mass) Function

P(A=a) --> [0, 1]
P(A=1) = 0.7
P(A=0) = 0.3

Many notations
$P_A(a)$
P(A)
P(a)

Crucial definitions/laws

They hold for *all* RV's. Make sure you know them!

Conditional Probability and Chain Rule

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

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

Law of Total Probability

$$P(A) = \sum_b P(A, B = b)$$

$$P(A) = \sum_b P(A|B = b)P(B = b)$$

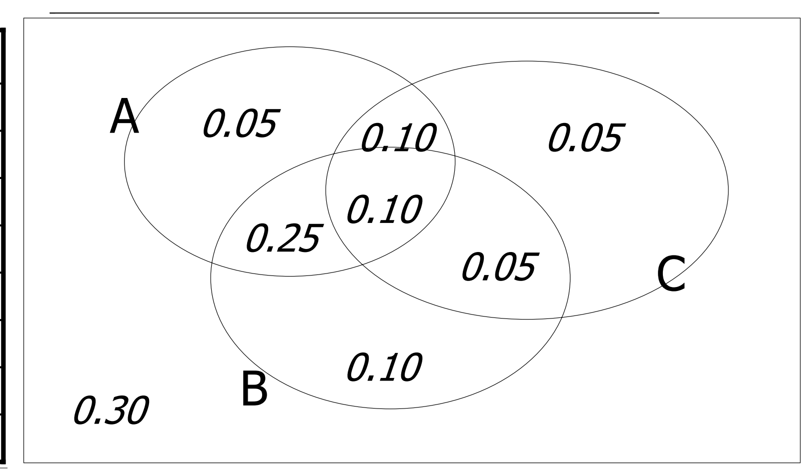
Disjunction (Union)

$$P(A \vee B) = P(A) + P(B) - P(AB)$$

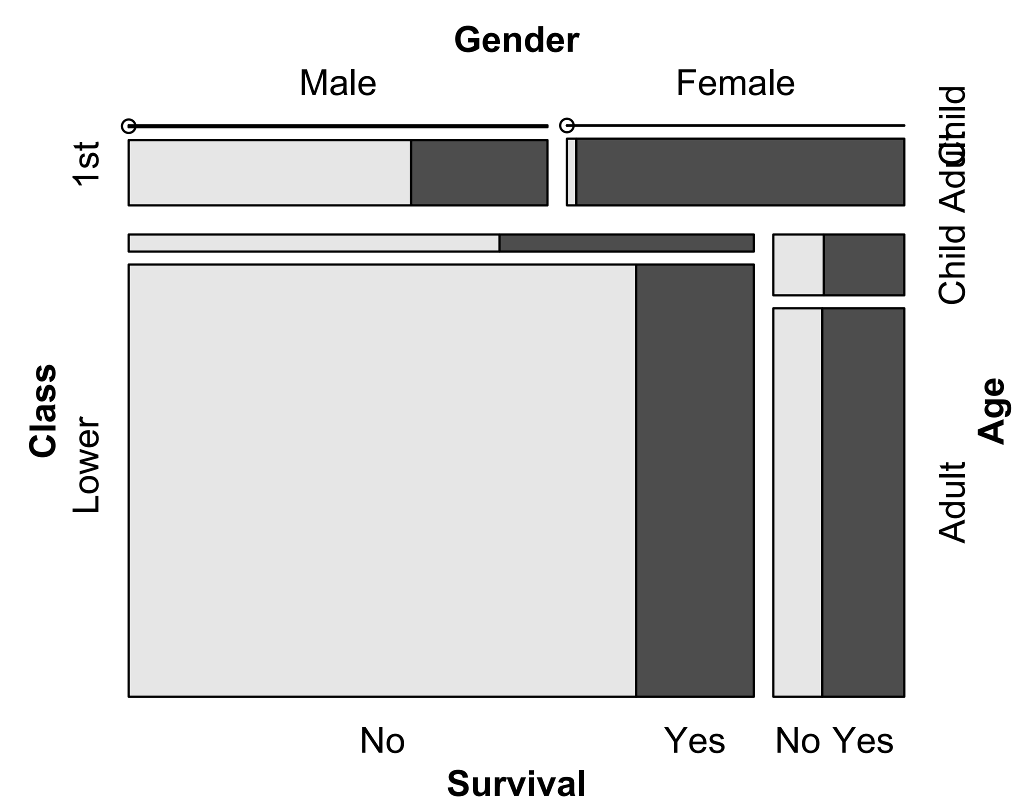
Negation (Complement)

$$P(\neg A) = 1 - P(A)$$

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



Class	Gender	Age	No	Yes	Total
1st	Male	Child	0	5	5
1st	Male	Adult	118	57	175
1st	Female	Child	0	1	1
1st	Female	Adult	4	140	144
Lower	Male	Child	35	24	59
Lower	Male	Adult	1211	281	1492
Lower	Female	Child	17	27	44
Lower	Female	Adult	105	176	281
		totals:	1490	711	2201



Any close conditional independences?
P(Survive | adult lower female) \approx P(Survive | child lower female)
Survive _|_ Age | lower female

“Data”: 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

P(A=☃) =

P(A=☹) =

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

$$P(A=☃) = \frac{15 + 20}{1035}$$

$$P(A=☹) = \frac{250 + 750}{1035}$$

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

$$P(A=☃) = \frac{15 + 20}{1035}$$

$$P(A=☹) = \frac{250 + 750}{1035}$$

Sum Rule

$$\sum_{a \in Values(A)} P(A = a) = 1$$

$$P(A=☃) + P(A=☹) = 1$$

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

P(A=☃) = $\frac{15 + 20}{1035}$

P(A=☹) = $\frac{250 + 750}{1035}$

Sum Rule

$$\sum_{a \in Values(A)} P(A = a) = 1$$

P(A=☃) + P(A=☹) = 1

Negation (Complement)

P(A!=☃) = 1 - P(A=☃) = $\frac{1035}{1035} - \frac{15 + 20}{1035}$

Disjunctions

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

P(A=☃ or B=☼) =

$$\frac{\#\{A=☃ \text{ or } B=☼\}}{N} = \frac{15+20+250}{1035}$$

$$P(A \vee B) = P(A) + P(B) - P(AB)$$

P(A=☃ or B=☼) =

= P(A=☃) + P(B=☼) - P(A=☃,B=☼)

$$= \frac{15+20}{1035} + \frac{15+250}{1035} - \frac{15}{1035}$$

Joint and Conditional Probs

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

P(A=☃, B=☼) =

P(A=☃ | B=☼) =

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

Joint and Conditional Probs

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

$$P(A=\text{☃}, B=\text{☼}) = \frac{\#\{A=\text{☃}, B=\text{☼}\}}{N} = \frac{15}{1035}$$

$$P(A=\text{☃} \mid B=\text{☼}) =$$

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

Joint and Conditional Probs

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

$$P(A=☃, B=☼) = \frac{\#\{A=☃, B=☼\}}{N} = \frac{15}{1035}$$

$$P(A=☃ \mid B=☼) = \frac{\#\{A=☃, B=☼\}}{\#\{B=☼\}} = \frac{15}{15 + 250}$$

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

Joint and Conditional Probs

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

$$P(A=☃, B=☼) = \frac{\#\{A=☃, B=☼\}}{N} = \frac{15}{1035}$$

$$P(A=☃ \mid B=☼) = \frac{\#\{A=☃, B=☼\}}{\#\{B=☼\}} = \frac{15}{15 + 250}$$

Every probability
has a "universe"
or background knowledge

P(A=☃, B=☼ | in the table)

P(A=☃ | B=☼, and in the table)

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

Joint and Conditional Probs

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

$$P(A=☃, B=☼) = \frac{\#\{A=☃, B=☼\}}{N} = \frac{15}{1035}$$

$$P(A=☃ \mid B=☼) = \frac{\#\{A=☃, B=☼\}}{\#\{B=☼\}} = \frac{15}{15 + 250}$$

"Fraction of worlds in which B is true that also have A true"

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

$$P(A=☃ \mid B=☼) = \frac{P(A=☃, B=☼)}{P(B=☼)} = \frac{\;\frac{15}{1035}\;}{\frac{15 + 250}{1035}}$$

“Data”: 1035 (A,B) pairs

Joint and Conditional Probs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

P(A=☃, B=☼) = $\frac{\#\{A=☃, B=☼\}}{N}$ = $\frac{15}{1035}$

P(A=☃ | B=☼) = $\frac{\#\{A=☃, B=☼\}}{\#\{B=☼\}}$ = $\frac{15}{15 + 250}$

Def’n conditional prob.

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

Chain rule

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

"Data": 1035 (A,B) pairs

Law of Total Probability

$$P(A) = \sum_b P(A, B=b)$$

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

Break A into B subgroups
Add joint probs across subgroups

$$P(A=☃) = P(A=☃, B=☼) + P(A=☃, B=☾) = \frac{15}{1035} + \frac{20}{1035}$$

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

Law of Total Probability

"Data": 1035 (A,B) pairs

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

Break A into B subgroups
Add joint probs across subgroups

$$P(A) = \sum_b P(A, B=b)$$

$$P(A=☃) = P(A=☃, B=☼) + P(A=☃, B=☾) = \frac{15}{1035} + \frac{20}{1035}$$

	A=☃	A=☹
B=☼	15	250
B=☾	20	750

Break universe into B subgroups
Average cond probs across subgroups

$$P(A) = \sum_b P(A|B=b)P(B=b)$$

$$P(A=☃) = P(A=☃ | B=☼)\, P(B=☼) + P(A=☃ | B=☾)\, P(B=☾)$$

$$= \frac{15}{15+250}\,\frac{15+250}{1035} + \frac{20}{20+750}\,\frac{20+750}{1035}$$

Crucial definitions/laws

They hold for *all* RV's

Conditional Probability and Chain Rule

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

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

Law of Total Probability

$$P(A) = \sum_b P(A, B = b)$$

$$P(A) = \sum_b P(A|B = b)P(B = b)$$

Disjunction (Union)

$$P(A \vee B) = P(A) + P(B) - P(AB)$$

Negation (Complement)

$$P(\neg A) = 1 - P(A)$$

Background conditional variants

$$P(\neg A) = 1 - P(A)$$

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

etc...

picture

Background conditional variants

$$P(\neg A) = 1 - P(A)$$

$$P(\neg A|Z) = 1 - P(A|Z)$$

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

etc...

picture

Background conditional variants

$$P(\neg A) = 1 - P(A) \qquad P(\neg A|Z) = 1 - P(A|Z)$$

$$P(AB) = P(A|B)P(B) \qquad P(AB|Z) = P(A|BZ)P(B|Z)$$

etc...

picture

Independence

- A and B are independent when
  $P(A|B) = P(A)$
  (Alternate formulation...)
- $P(AB) = ?$

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

Independence

- A and B are independent when
  $P(A|B) = P(A)$
  (Alternate formulation...)

- $P(AB) = ?$

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

Independence

- Chain rule: in general
  P(ABC) =
- But when A, B, C are independent
  P(ABC) = ?
- If you have independence -- can estimate joint probability of thousands of RV's

[1] = P(A|BC) P(B|C) P(C)
? = P(A) P(B) P(C)

Independence

- Chain rule: in general
  P(ABC) = P(A|BC) P(B|C) P(C)
- But when A, B, C are independent
  P(ABC) = ?
- If you have independence -- can estimate joint probability of thousands of RV's

[1] = P(A|BC) P(B|C) P(C)
? = P(A) P(B) P(C)

Notation

$$P(AB)$$
$$P(A \wedge B)$$
$$P(A \cap B)$$

$$P(A = a, B = b)$$
$$P(A = a \ \wedge \ B = b)$$

$$P(A) = \sum_b P(A, B = b)$$

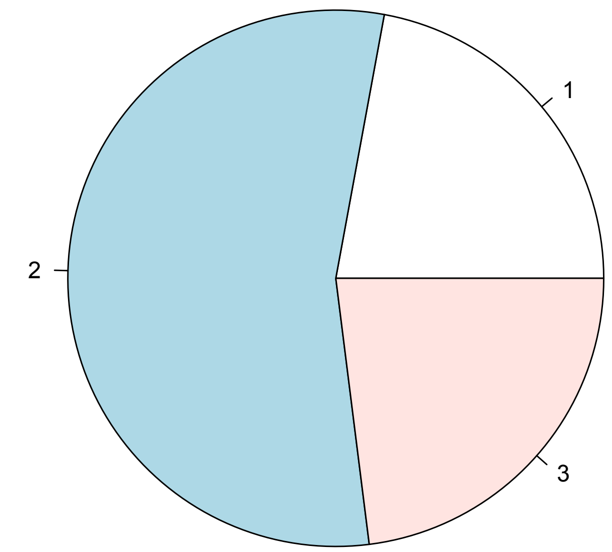
$$P(A = a) = \sum_b P(A = a, B = b)$$

Discrete vs Continuous RV's

Discrete: Prob Mass Fn

$$P(X) \longrightarrow [0,1]$$

$$\sum_{x \in Values_X} P(X=x) = 1$$



Continuous: Prob Density Fn

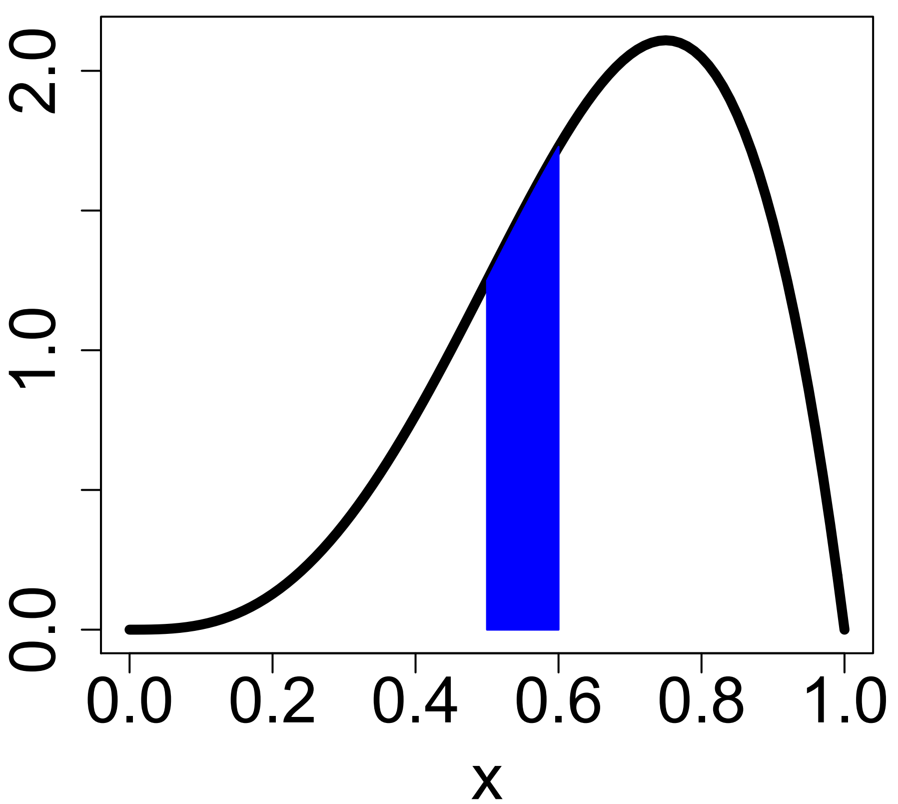
There is no probability at one point!

$$P(X \in [a,b]) = \int_a^b p(x)dx$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Quiz: can p(x)>1 ?

$p_{\mathrm{Beta}}(x;4,2)$

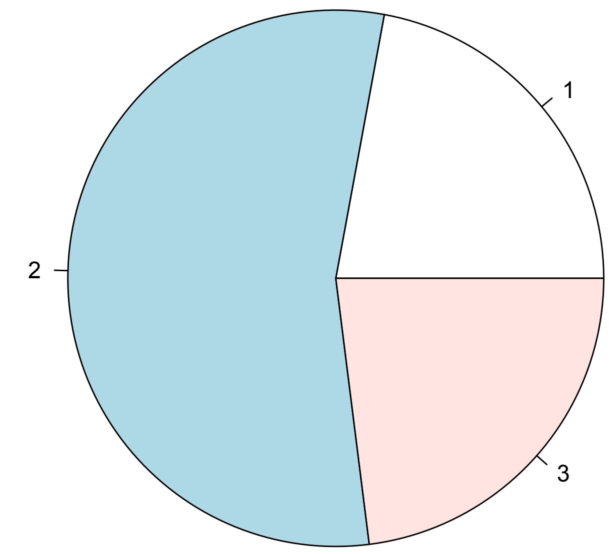


Discrete vs Continuous RV's

Discrete: Prob Mass Fn

$$P(X) \longrightarrow [0, 1]$$

$$\sum_{x \in Values_X} P(X = x) = 1$$



Continuous: Prob Density Fn

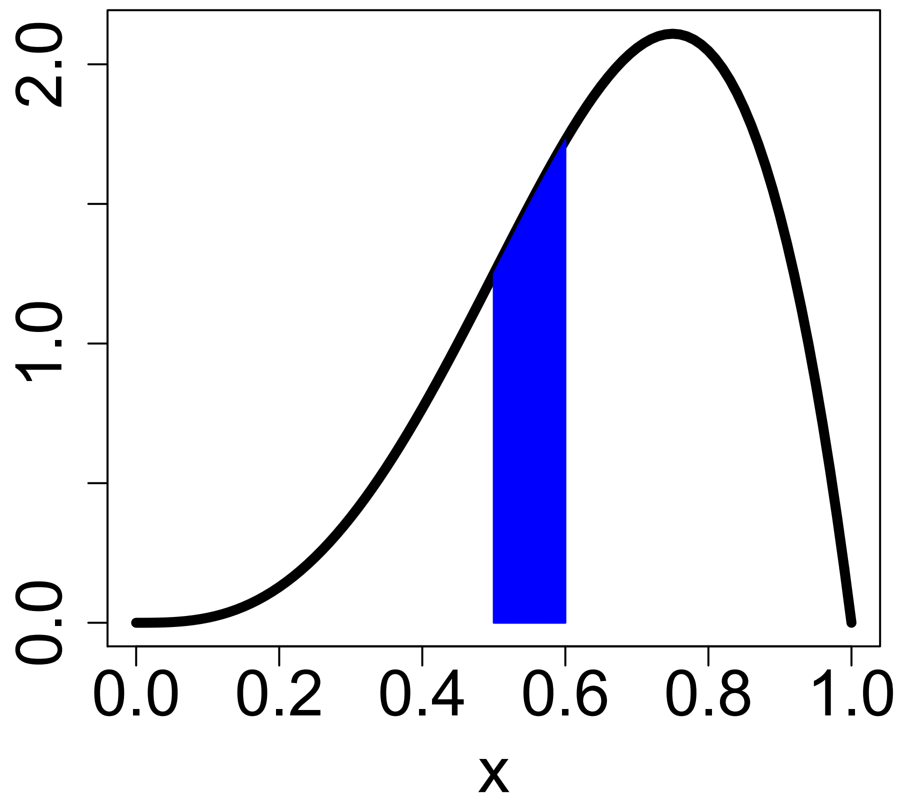
There is no probability at one point!

$$P(X \in [a, b]) = \int_a^b p(x) dx$$

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

$$p(x) \longrightarrow [0, \infty)$$

$p_{\mathrm{Beta}}(x; 4, 2)$



Bayes Rule

H: have disease?

E: outcome of medical test

Want P(H|E) but only have P(E|H)

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

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Why? e.g. P(E|H) has been measured, or H causes E ...

Likelihood

Prior

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Posterior

Normalizer

Pierre-Simon Laplace
1749-1827

Rev. Thomas Bayes
c. 1701-1761

Bayes Rule and its pesky denominator

Likelihood Prior

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|h)P(H)}{\sum_{h'} P(E|h')P(h')}$$

$$P(H|E) = \frac{1}{Z} P(E|H)P(H)$$

Z: whatever lets the posterior, when summed across h, to sum to 1
Zustandssumme, "*sum over states*"

$$P(H|E) \propto \underline{P(E|H)P(H)}$$

"Proportional to"

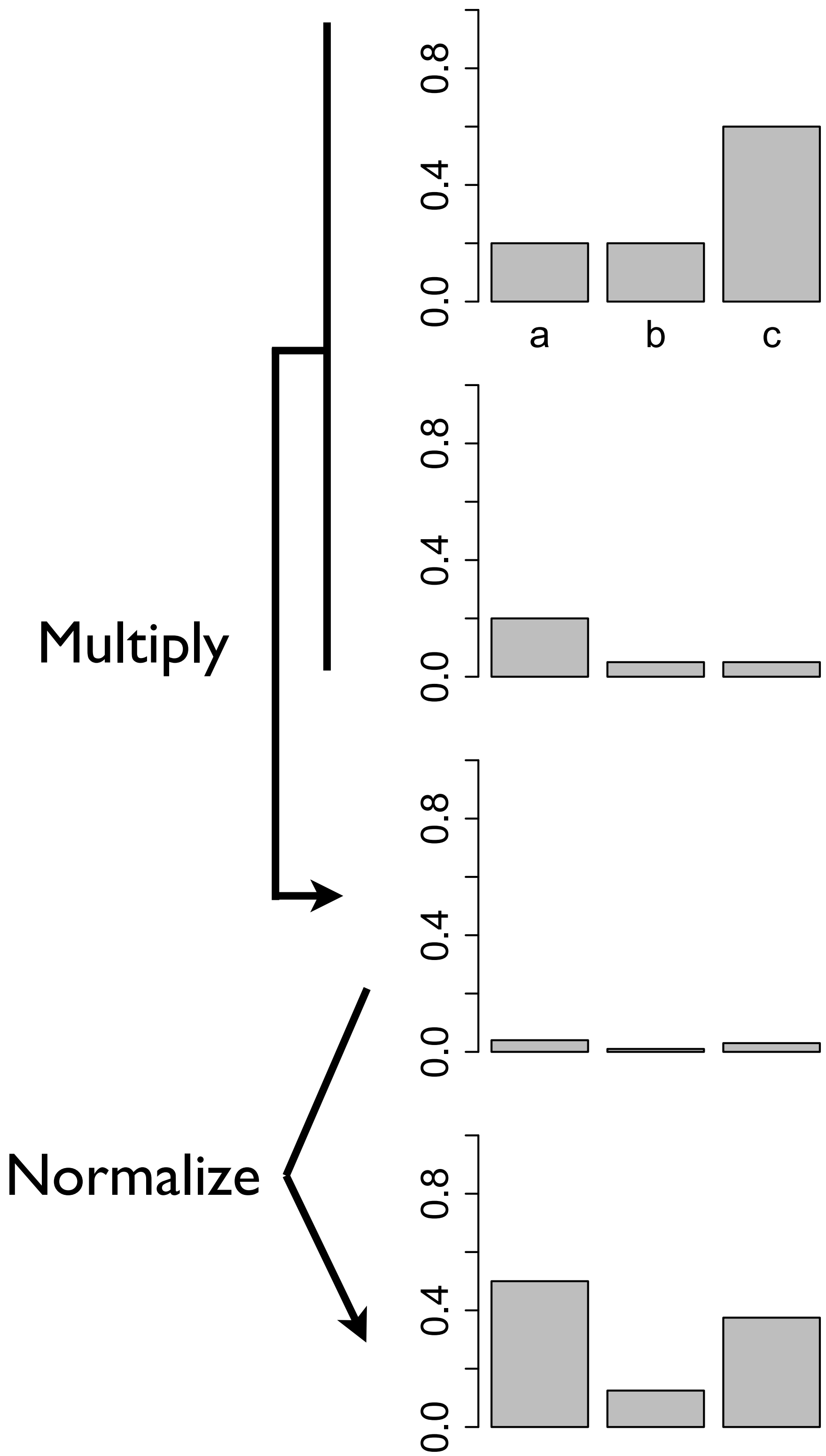
Unnormalized posterior
By itself does not sum to 1!

Zustandssumme, "sum over states" -- from 19th century statistical physics.

(Called a "partition function" for certain types of models, "Markov random fields".)

Computing this can be very difficult in high dimensions -- tons of work in statistical machine learning is essentially devoted to computing Z or variants of it.

Bayes Rule: Discrete



<u>Sum to 1?</u>

$$P(H = h)$$

Prior

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

Likelihood

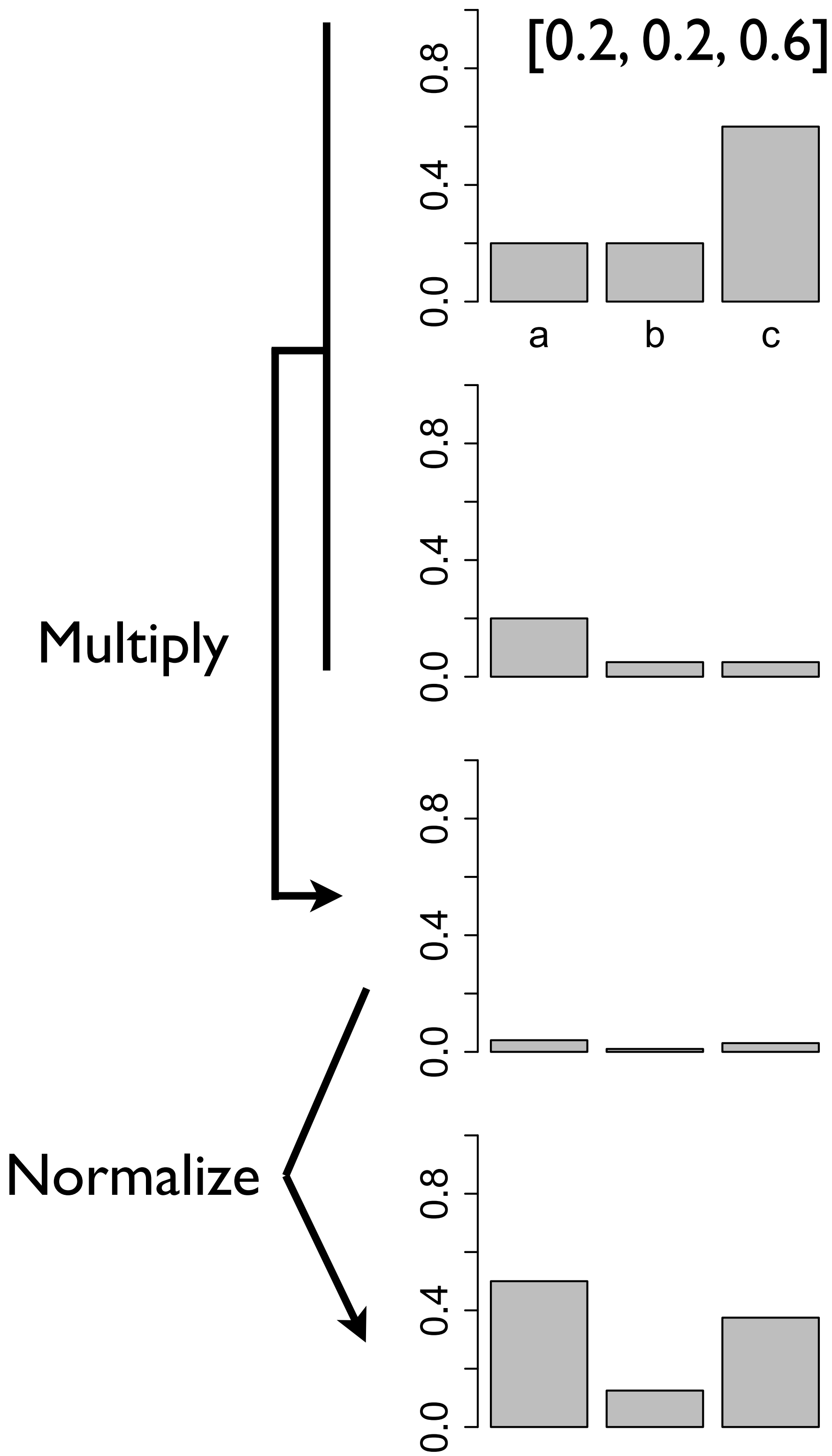
$$P(E|H = h)P(H = h)$$

Unnorm. Posterior

$$\frac{1}{Z}P(E|H = h)P(H = h)$$

Posterior

Bayes Rule: Discrete



<u>Sum to 1?</u>

$$P(H = h)$$

Prior

Yes

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

Likelihood

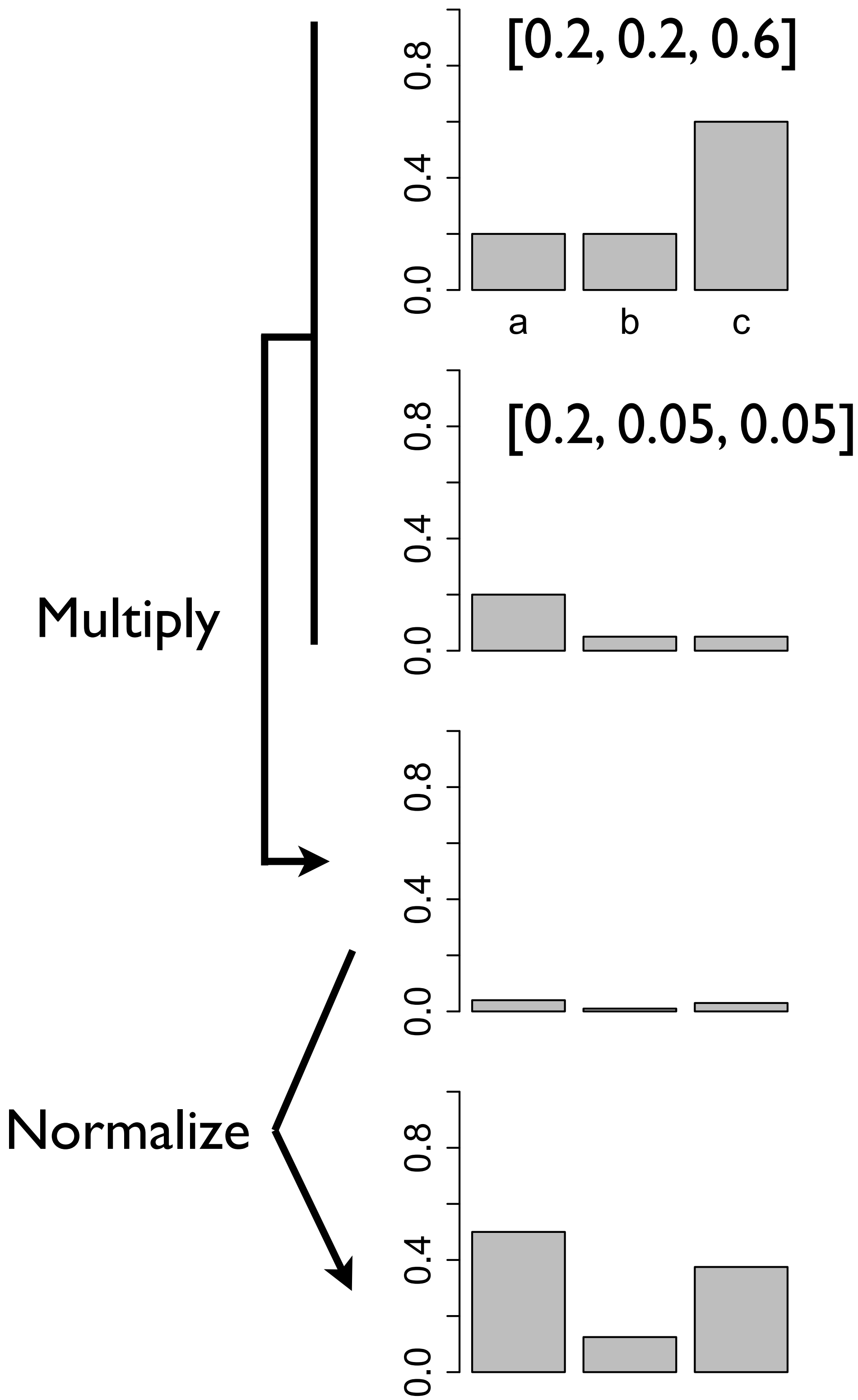
$$P(E|H = h)P(H = h)$$

Unnorm. Posterior

$$\frac{1}{Z}P(E|H = h)P(H = h)$$

Posterior

Bayes Rule: Discrete



<u>Sum to 1?</u>

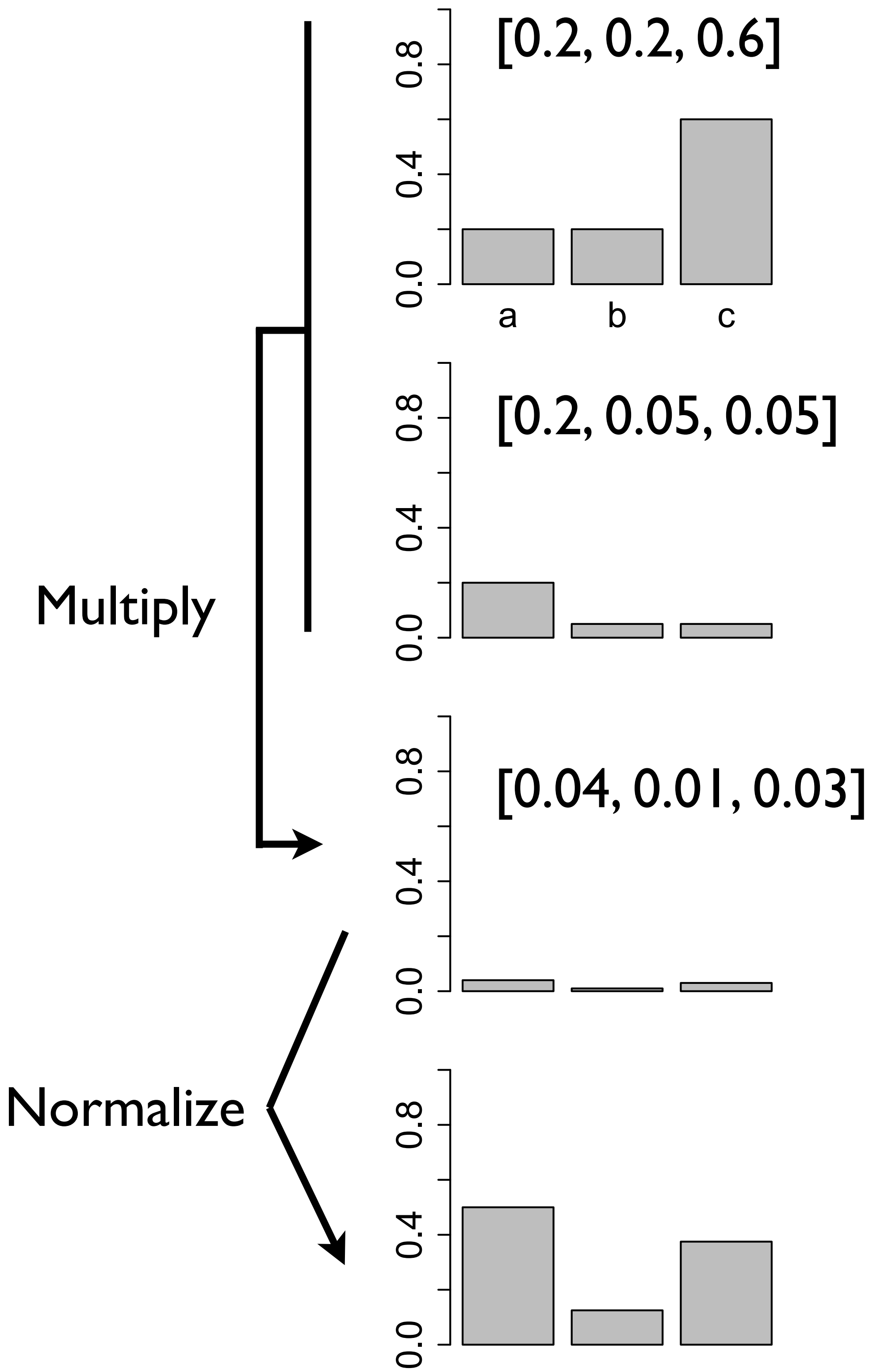
$$P(H = h)$$
Prior — Yes

$$P(E|H = h)$$
Likelihood — No

$$P(E|H = h)P(H = h)$$
Unnorm. Posterior

$$\frac{1}{Z}P(E|H = h)P(H = h)$$
Posterior

Bayes Rule: Discrete



<u>Sum to 1?</u>

$$P(H = h)$$

Prior — Yes

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

Likelihood — No

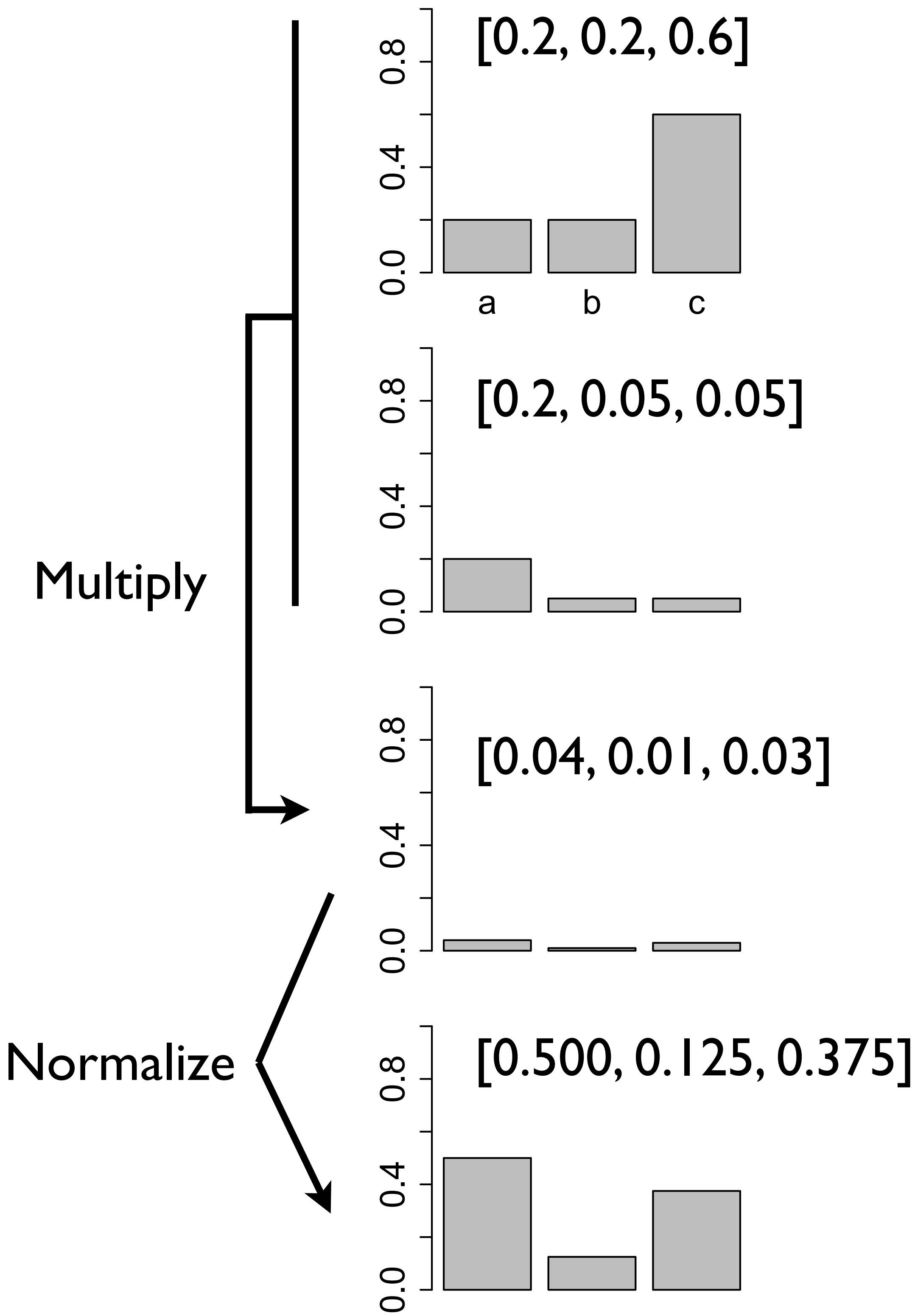
$$P(E|H = h)P(H = h)$$

Unnorm. Posterior — No

$$\frac{1}{Z}P(E|H = h)P(H = h)$$

Posterior

Bayes Rule: Discrete



Sum to 1?

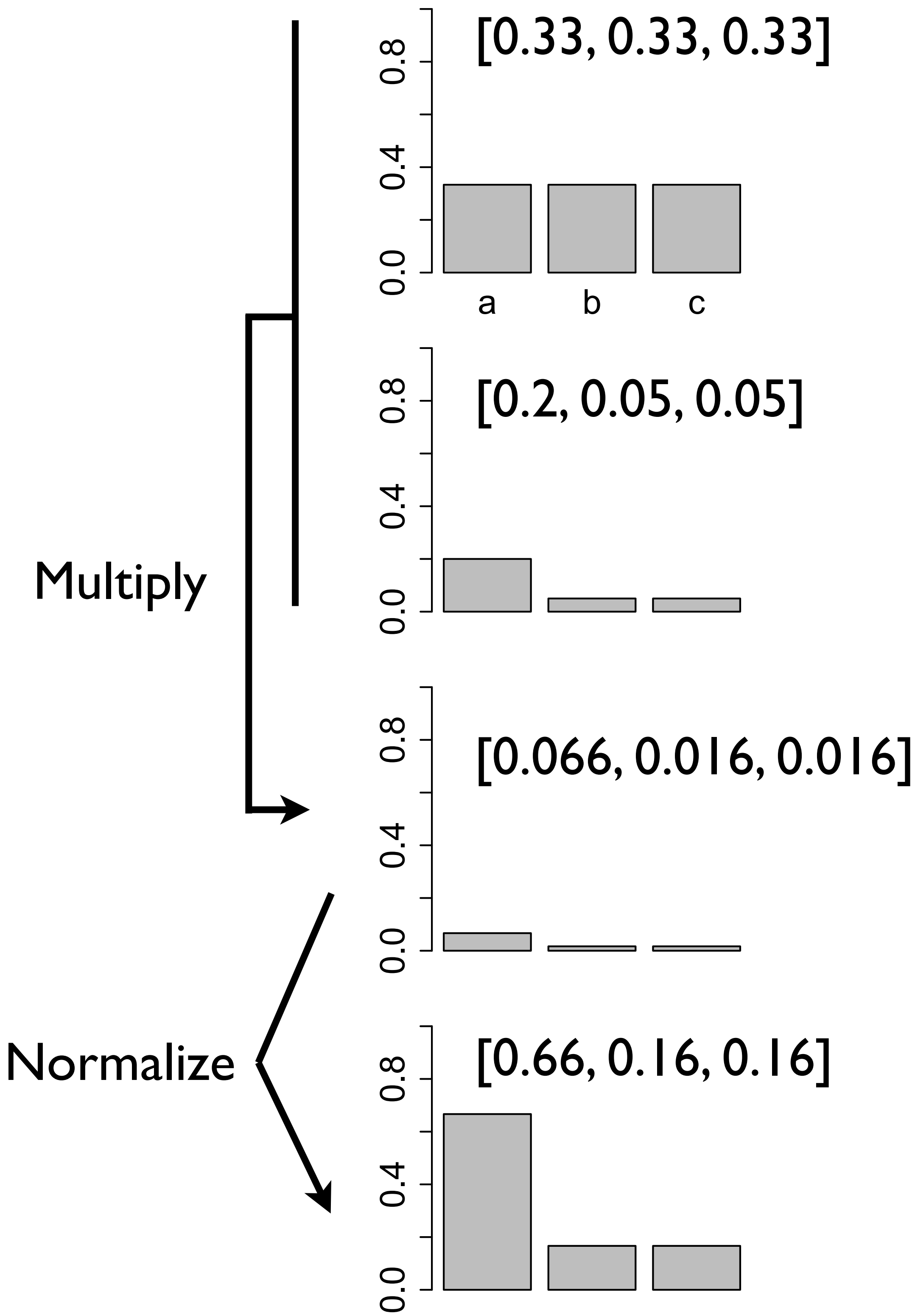
$$P(H = h)$$
Prior — Yes

$$P(E|H = h)$$
Likelihood — No

$$P(E|H = h)P(H = h)$$
Unnorm. Posterior — No

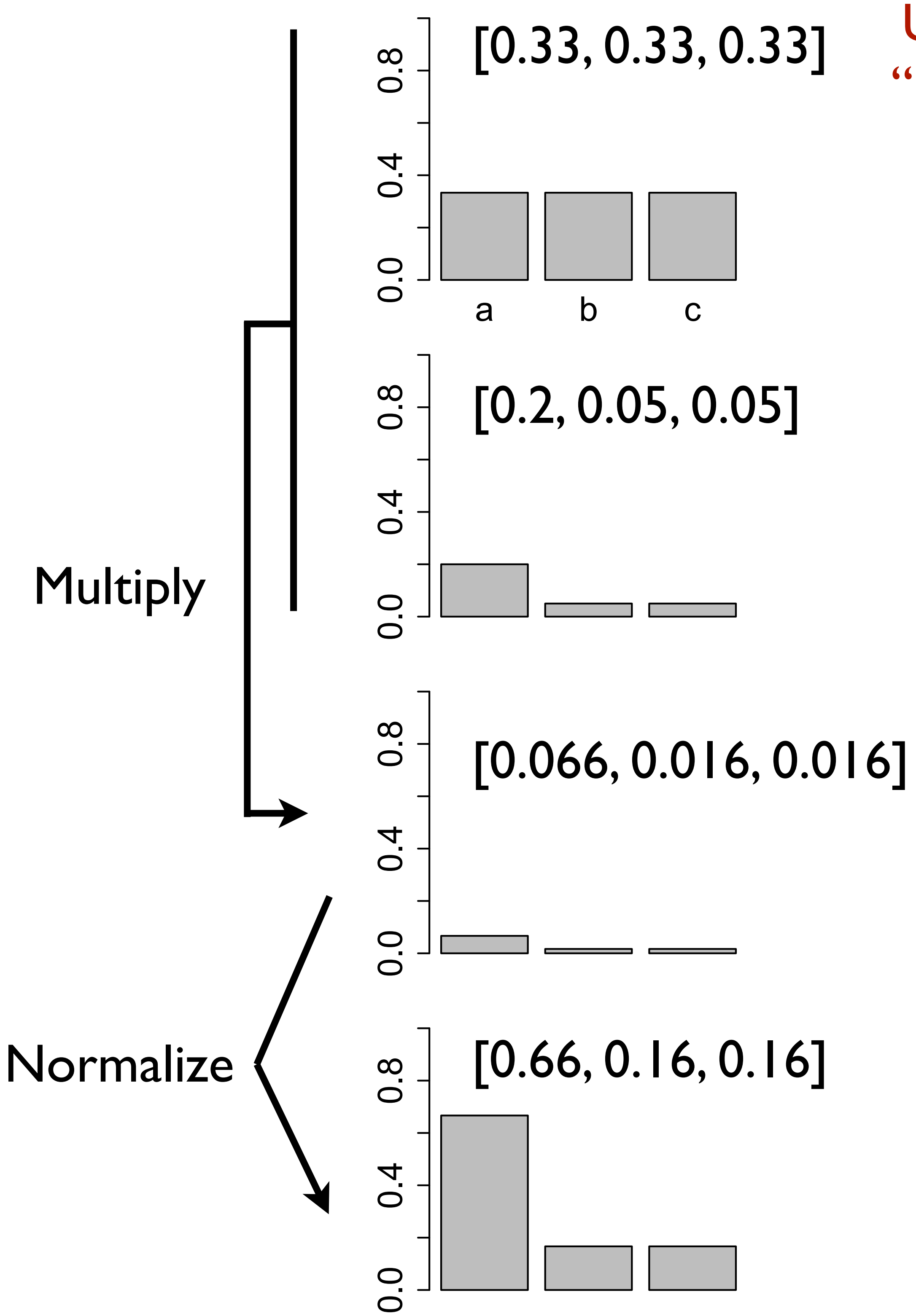
$$\frac{1}{Z}P(E|H = h)P(H = h)$$
Posterior — Yes

Bayes Rule: Discrete, uniform prior



		Sum to 1?
$P(H = h)$	Prior	Yes
$P(E\|H = h)$	Likelihood	No
$P(E\|H = h)P(H = h)$	Unnorm. Posterior	No
$\frac{1}{Z}P(E\|H = h)P(H = h)$	Posterior	Yes

Bayes Rule: Discrete, uniform prior



Uniform distribution:
"Uninformative prior"

<u>Sum to 1?</u>

$$P(H = h)$$

Prior — Yes

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

Likelihood — No

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

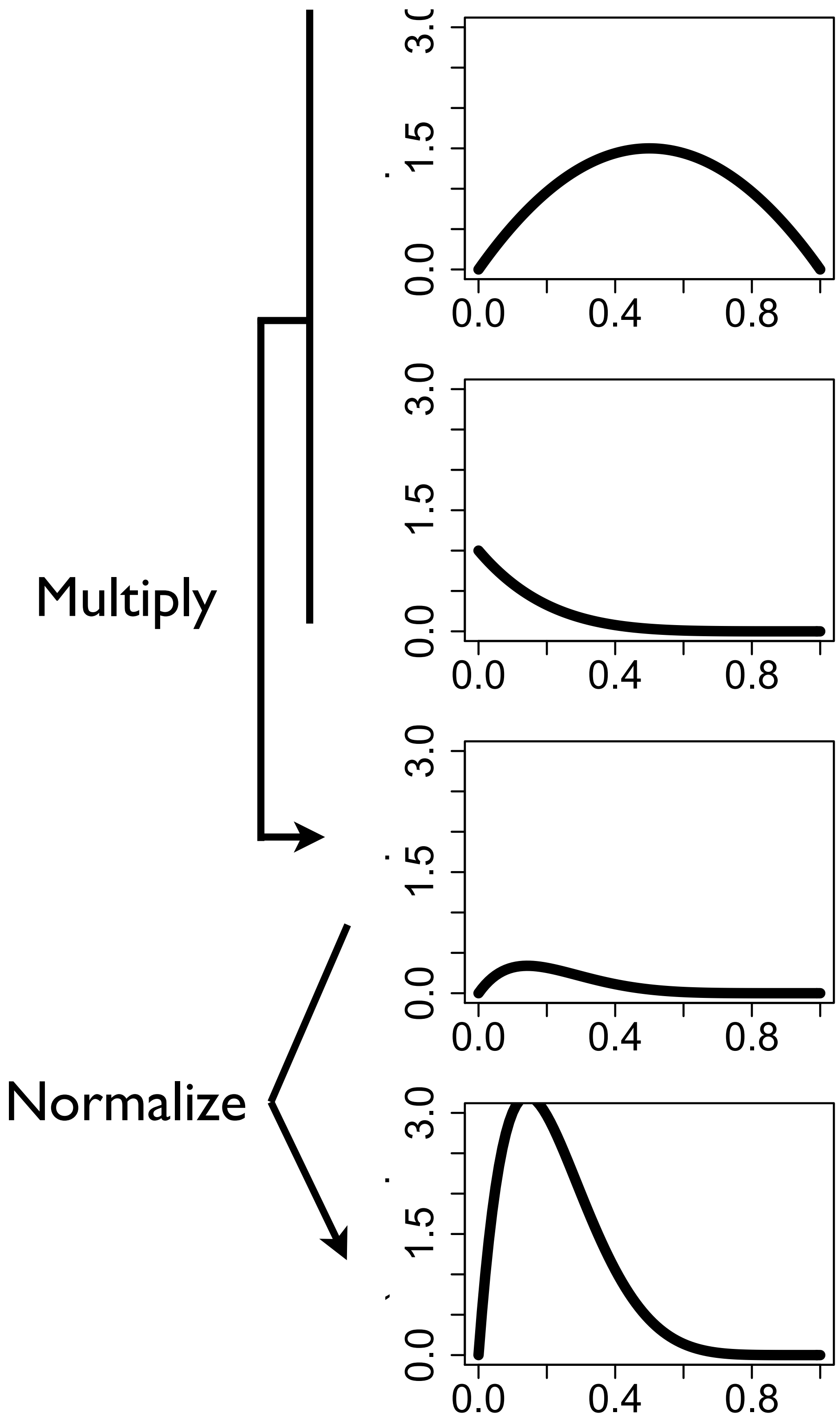
Unnorm. Posterior — No

Uniform prior implies that
posterior is just renormalized likelihood

$$\frac{1}{Z}P(E|H = h)P(H = h)$$

Posterior — Yes

Bayes Rule: Continuous, with prior.

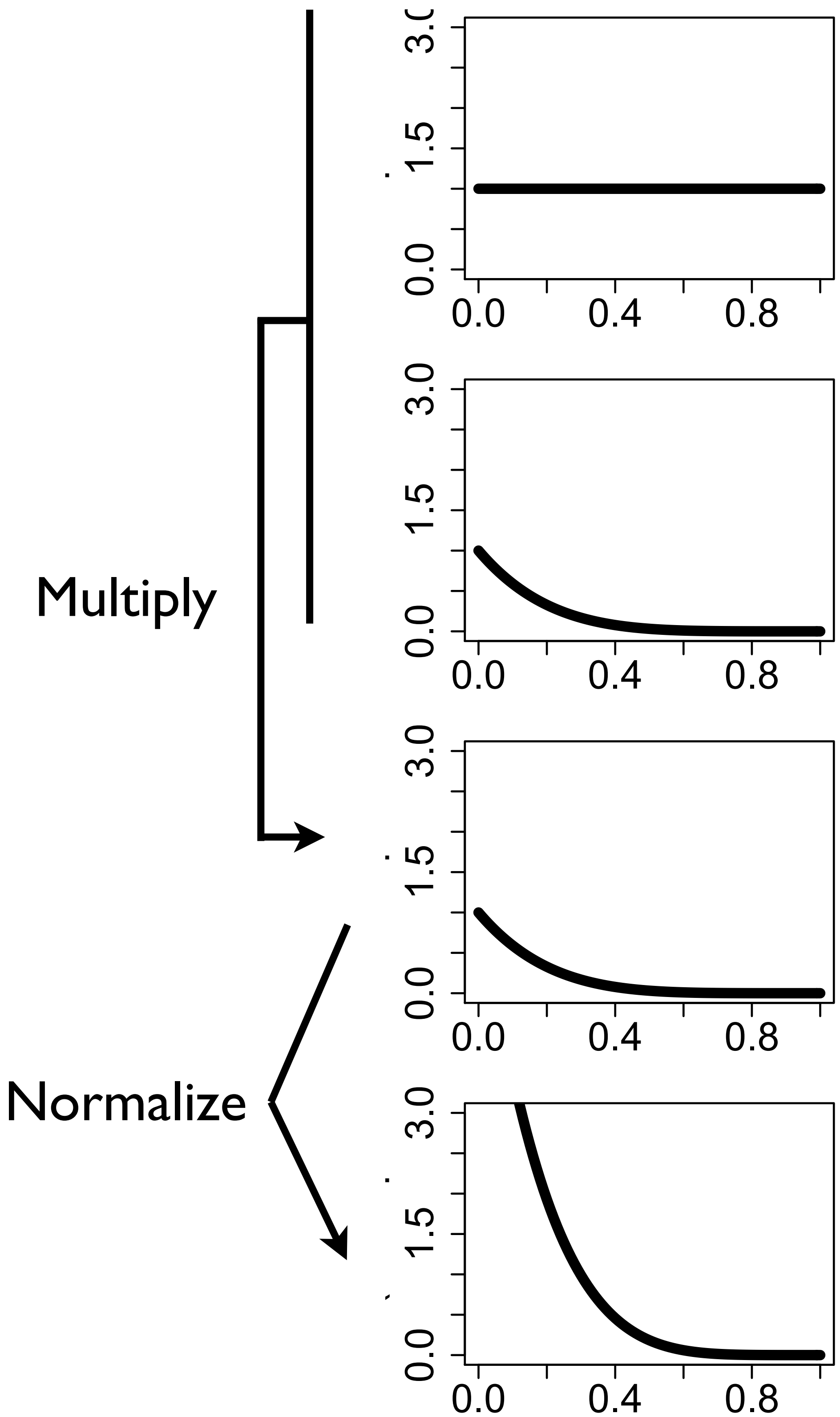


Beta(2,2)

L(theta)
0 heads
5 tails

		Sum to 1?
$P(H=h)$	Prior	Yes
$P(E\|H=h)$	Likelihood	No
$P(E\|H=h)P(H=h)$	Unnorm. Posterior	No
$\frac{1}{Z}P(E\|H=h)P(H=h)$	Posterior	Yes

Bayes Rule: Continuous, uniform prior



Uniform distribution:
"Uninformative prior"

<u>Sum to 1?</u>

$$P(H = h)$$

Prior — Yes

L(theta)
0 heads
5 tails

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

Likelihood — No

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

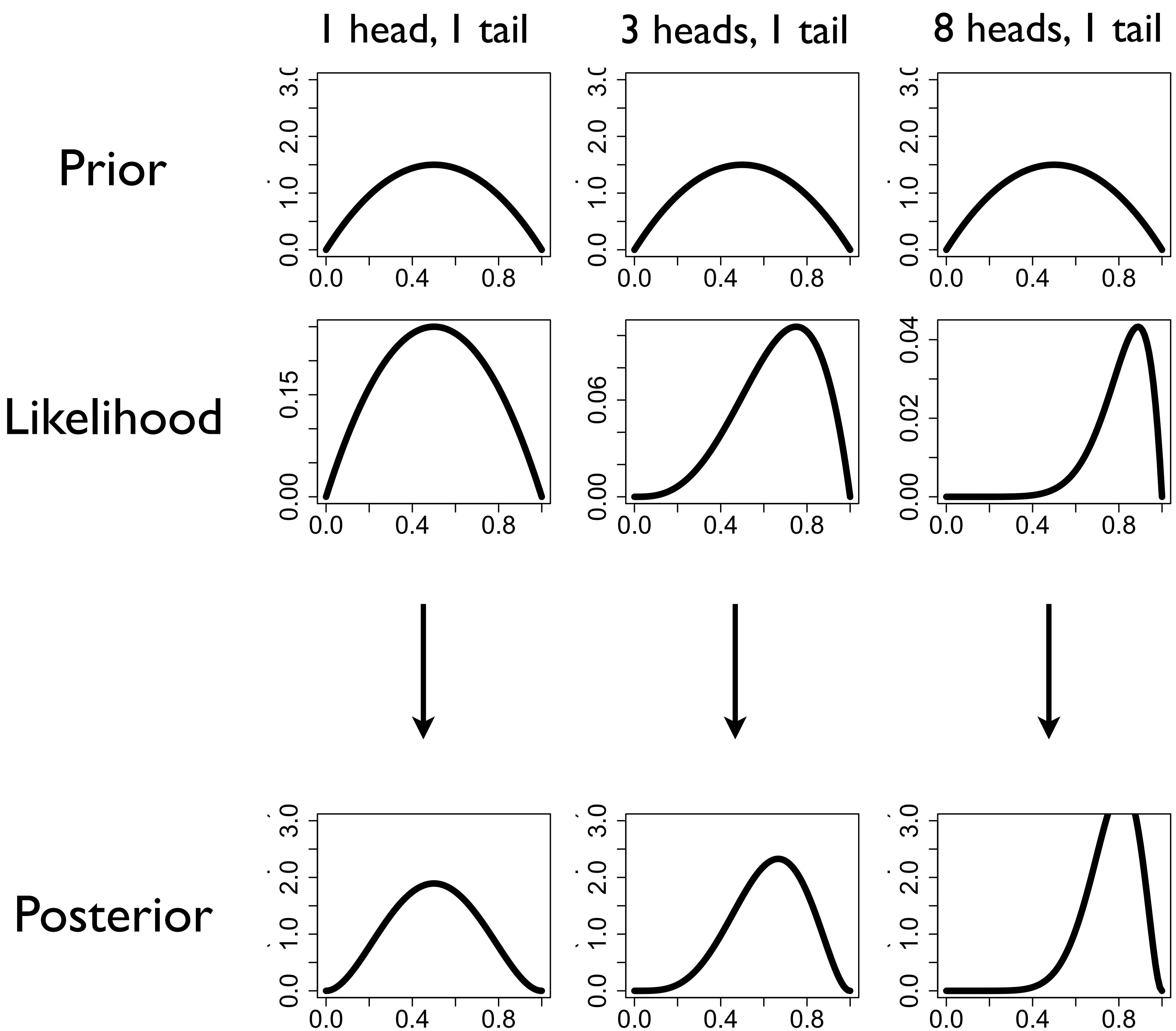
Unnorm. Posterior — No

Uniform prior implies that
posterior is just renormalized likelihood

$$\frac{1}{Z}P(E|H = h)P(H = h)$$

Posterior — Yes

Bayes Rule: More data swamps priors.



$$p_{Beta}(\theta; 2, 2)$$

$$\frac{1}{Z}\theta(1-\theta)$$

$$\theta^{\#H}(1-\theta)^{\#T}$$

$$\frac{1}{Z}\theta^{\#H+1}(1-\theta)^{\#T+1}$$

Bayes Rule: odds-ratio form

Posterior prob | Likelihood | Prior prob

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

[0, 1] | [0, 1] | [0, 1]

$$\frac{P(h_1|E)}{P(h_2|E)} = \frac{P(E|h_1)}{P(E|h_2)}\frac{P(h_1)}{P(h_2)}$$

Posterior Odds | Likelihood Ratio | Prior Odds

[0, inf) | [0, inf) | [0, inf)

Bayes Rule: odds-ratio form

Posterior prob | Likelihood | Prior prob

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

[0, 1] | [0, 1] | [0, 1]

$$\frac{P(h_1|E)}{P(h_2|E)} = \frac{P(E|h_1)}{P(E|h_2)} \frac{P(h_1)}{P(h_2)}$$

Posterior Odds | Likelihood Ratio | Prior Odds

[0, inf) | [0, inf) | [0, inf)

Binary prob rescalings

		Range	50-50 coin flip
prob	p	[0,1]	1/2
log-prob	log(p)	(-inf, 0]	log(1/2)
odds	p / (1-p)	[0, +inf)	1
log-odds "logit"	log[p/(1-p)]	(-inf, +inf)	0

i.i.d. likelihood

"Bernoulli" random process:
X_i comes out H at prob θ, else comes out T.

$$L(\theta) = P(X_1, X_2, X_3, ..., X_n \mid \theta) = ???$$

"Likelihood of θ": Prob of data under parameter θ

$$= \prod_i P(X_i \mid \theta)$$

Independent... and

$$= \prod_i \{\theta \text{ if } X_i = \text{H else } 1 - \theta\}$$

Identically Distributed!

$$= \prod_i \theta^{1\{X_i = H\}} (1-\theta)^{1\{X_i = T\}}$$

$$= \theta^{\#\{X_i = H\}} (1-\theta)^{\#\{X_i = T\}}$$

Maximum Likelihood

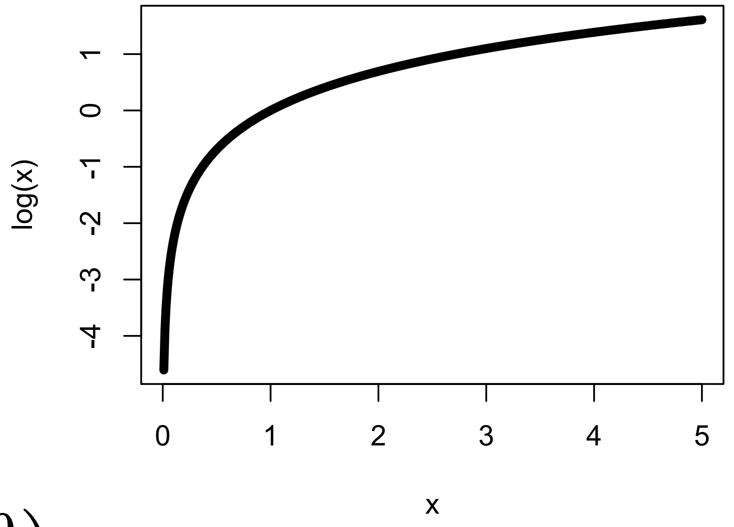
- If we want to believe a single point value of theta, what should we believe?
- How about the *maximum likelihood* value of theta: that makes the data most probable

$$L(\theta) = \theta^{\#\{X_i=H\}}(1-\theta)^{\#\{X_i=T\}}$$

$$\arg\max_{\theta} L(\theta)$$

$$\arg\max_{\theta} \log L(\theta)$$

Most important trick:
Take the log!
Same argmax since monotonic.

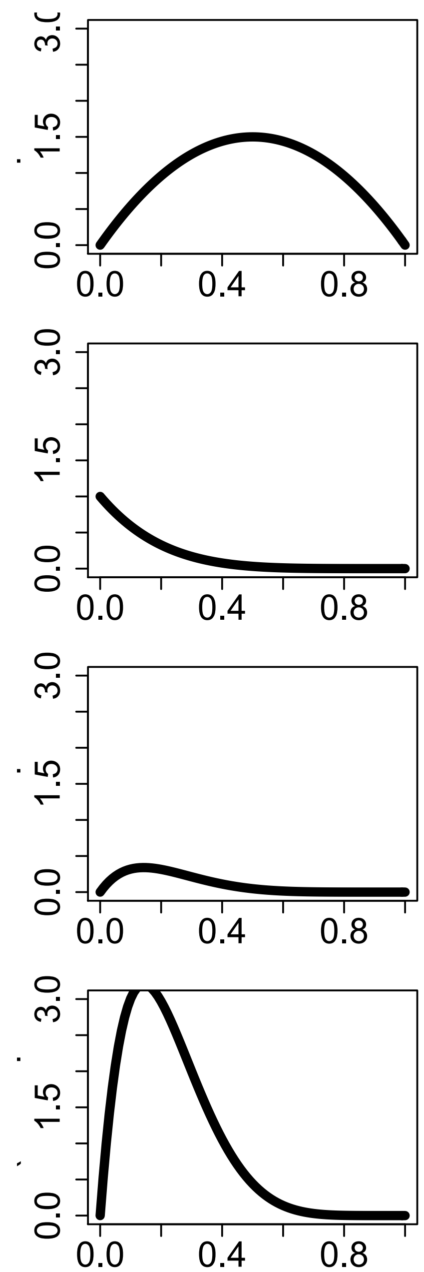


$$\#\{X_i = H\}\log\theta + \#\{X_i = T\}\log(1-\theta)$$

Maximum Posterior (MAP)

$$p(\theta|D) = P(D|\theta)p(\theta)/P(D)$$
$$p(\theta|D) \propto P(D|\theta)p(\theta)$$
$$\arg\max_\theta p(\theta|D) = \arg\max_\theta P(D|\theta)p(\theta)$$

Another trick: unnormalized is good enough for MAP



$P(H = h)$

Prior

$P(E|H = h)$

Likelihood

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

Unnorm. Posterior

$\frac{1}{Z}P(E|H = h)P(H = h)$

Posterior

Entropy

$$H(X) = -\sum_x p(x) \log p(x)$$

$$-H(X) = E[\log p(x)]$$