EE5907/EE5027 Lecture 2
MLE - MAP

Robby T. Tan
National University of Singapore

Probability Review

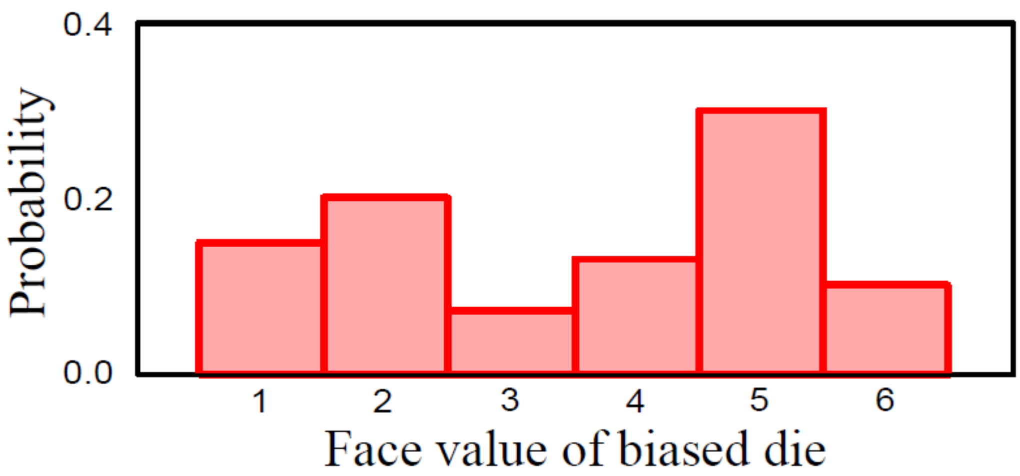
What is a random variable?

- A random variable x is a quantity that is uncertain
- May be result of experiment (e.g., flipping a coin) or real world measurement (e.g., measuring temperature)
- If observe x multiple times, we get different values
- Some values occur more than others; this information captured by probability distribution p(x)
- If x is discrete, then “p” is “probability mass function” (or pmf). If x is continuous, then “p” is “probability distribution function” (or pdf).

Adapted from S. Prince

Discrete Random Variable x

- Take on discrete values

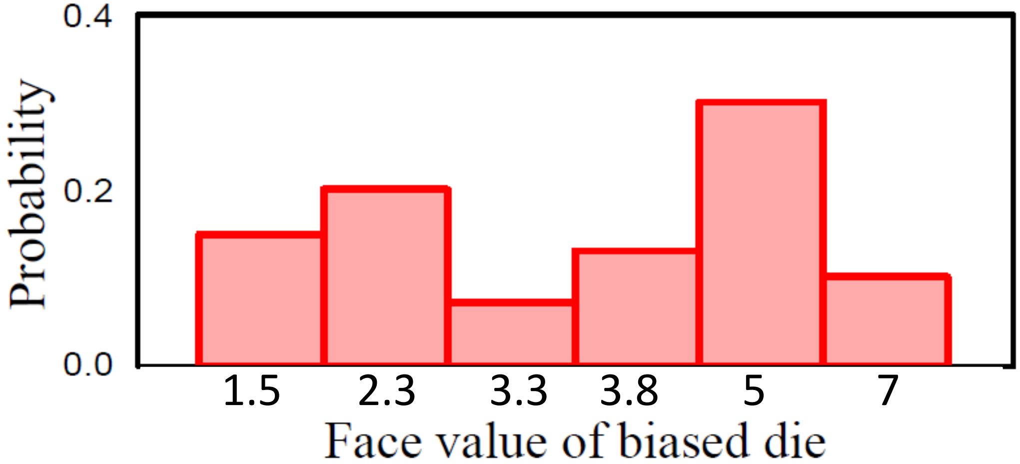


$$\sum_x p(x) = 1$$

Adapted from S. Prince

Discrete Random Variable x

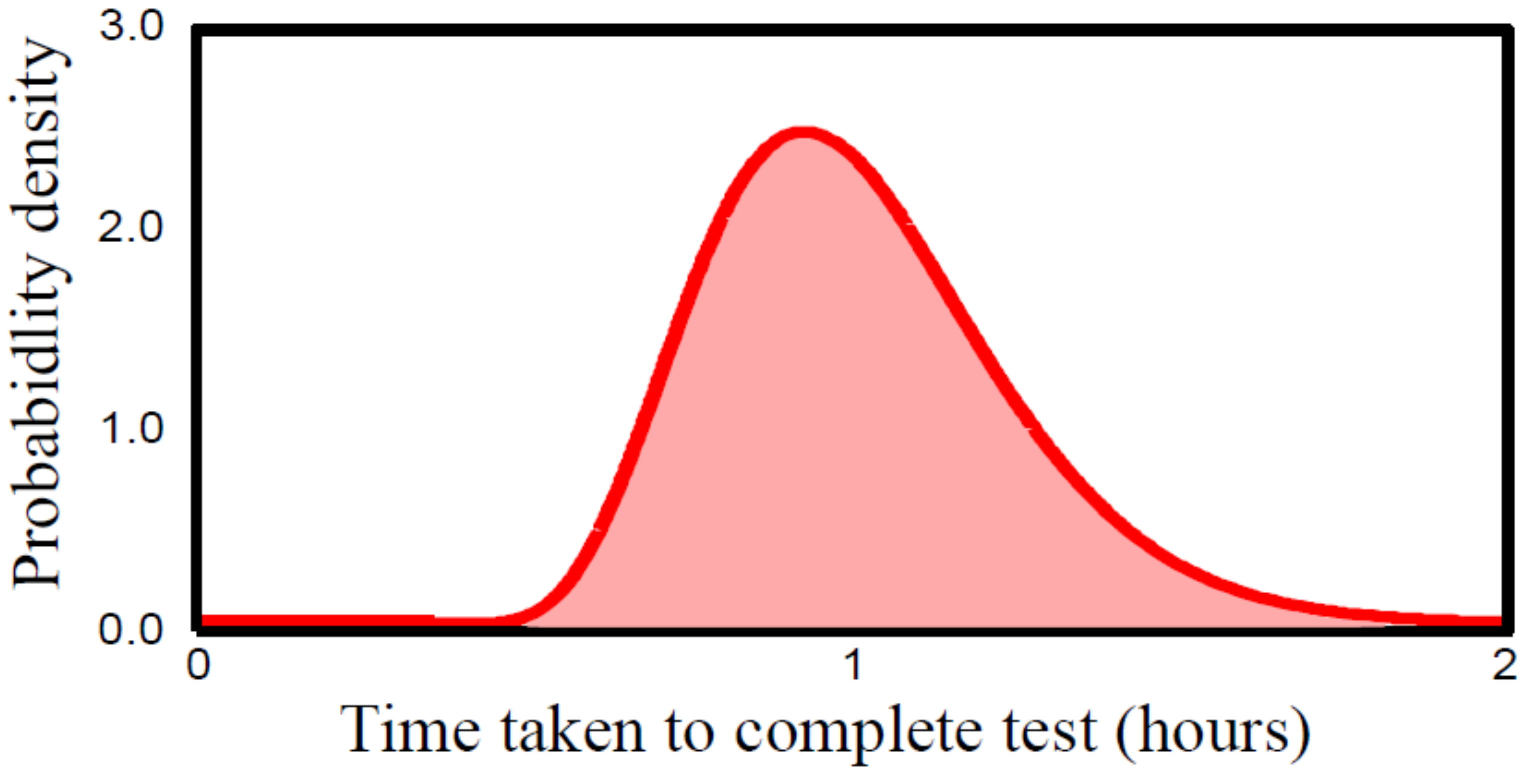
- Take on finite or countably infinite values



$$\sum_x p(x) = 1$$

Adapted from S. Prince

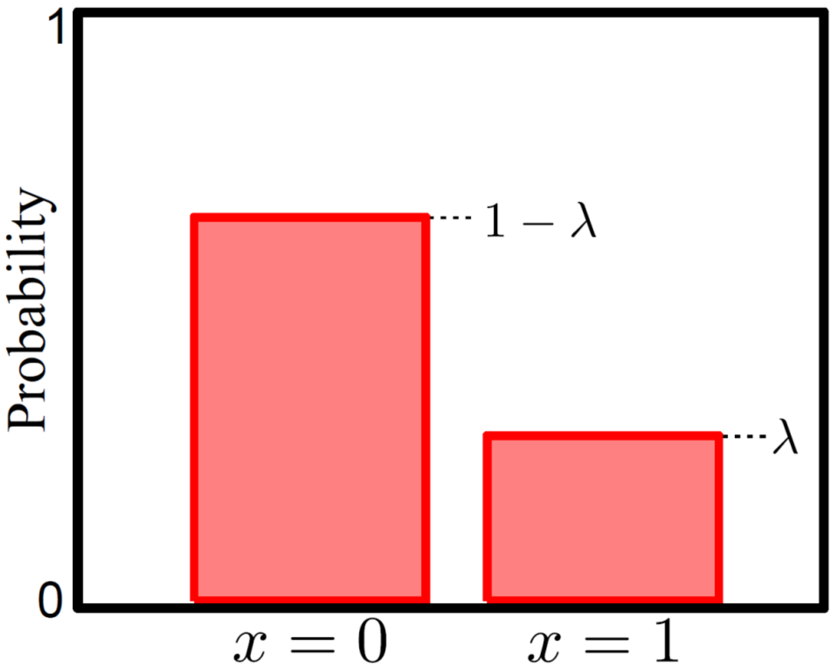
Continuous Random Variable



Famous Discrete Random Variables

- Bernoulli: http://en.wikipedia.org/wiki/Bernoulli_distribution
- Categorical: http://en.wikipedia.org/wiki/Categorical_distribution
- Binomial: http://en.wikipedia.org/wiki/Binomial_distribution
- Geometric: http://en.wikipedia.org/wiki/Geometric_distribution
- Poisson: http://en.wikipedia.org/wiki/Poisson_distribution
- …

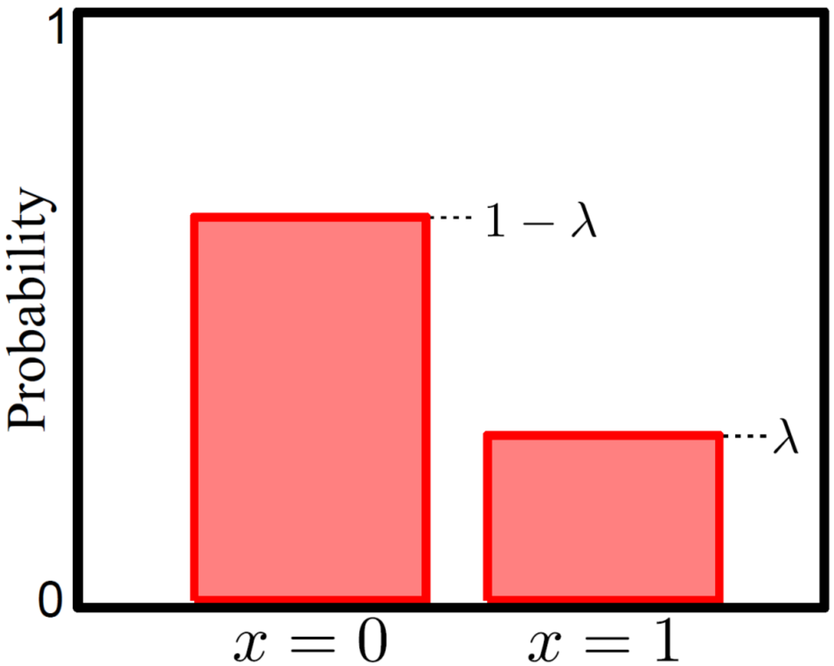
Bernoulli Distribution



Bernoulli distribution describes situation where only two possible outcomes x = 0 / x = 1 (e.g. failure/success)

Takes a single parameter $\lambda \in [0,1]$

Bernoulli Distribution

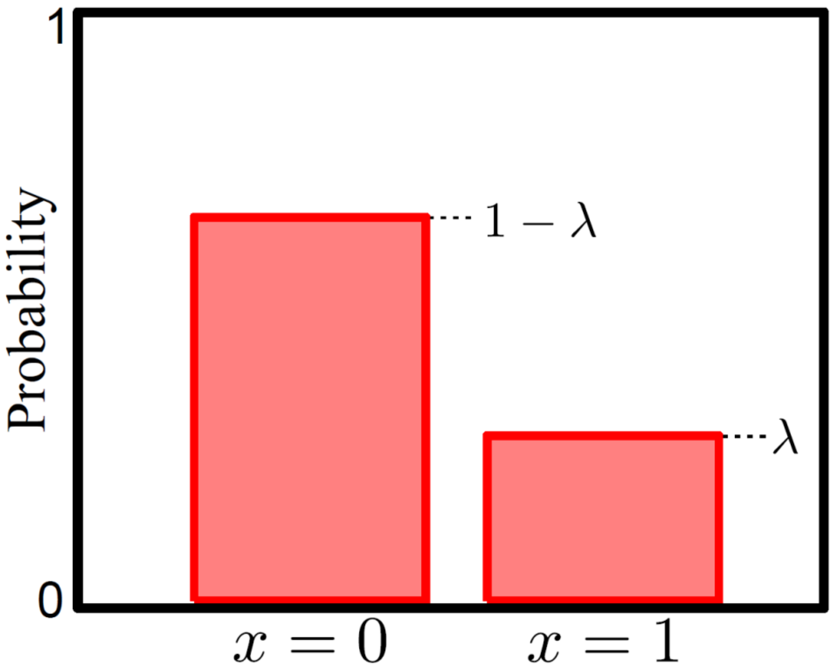


$$\begin{aligned} Pr(x=0) &= 1-\lambda \\ Pr(x=1) &= \lambda. \end{aligned}$$

Bernoulli distribution describes situation where only two possible outcomes x = 0 / x = 1 (e.g. failure/success)

Takes a single parameter $\lambda \in [0,1]$

Bernoulli Distribution



$$
\begin{aligned}
Pr(x=0) &= 1-\lambda \\
Pr(x=1) &= \lambda.
\end{aligned}
$$

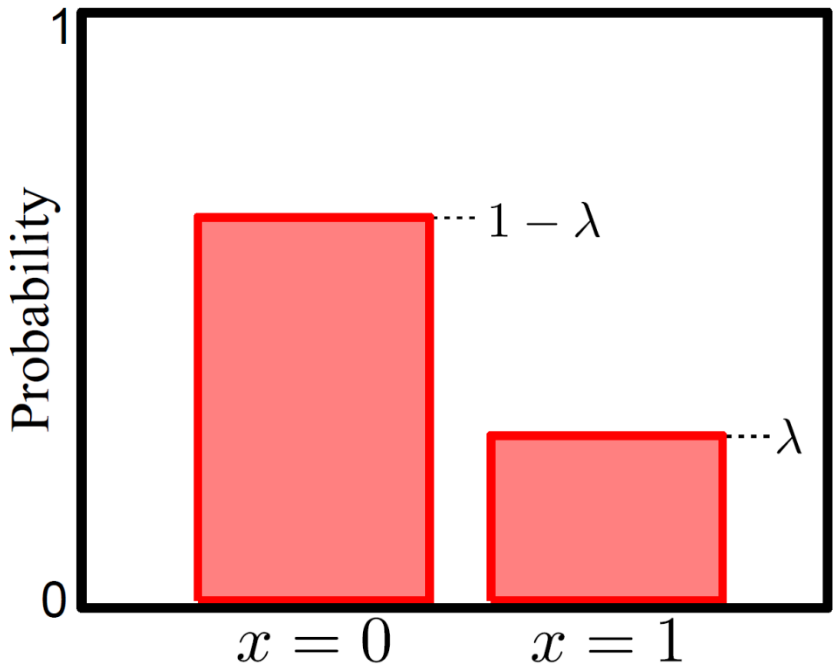
or

$$Pr(x) = \lambda^x(1-\lambda)^{1-x}$$

Bernoulli distribution describes situation where only two possible outcomes x = 0 / x = 1 (e.g. failure/success)

Takes a single parameter $\lambda \in [0, 1]$

Bernoulli Distribution



$$
\begin{aligned}
Pr(x=0) &= 1-\lambda \\
Pr(x=1) &= \lambda.
\end{aligned}
$$

or

$$Pr(x) = \lambda^x (1-\lambda)^{1-x}$$

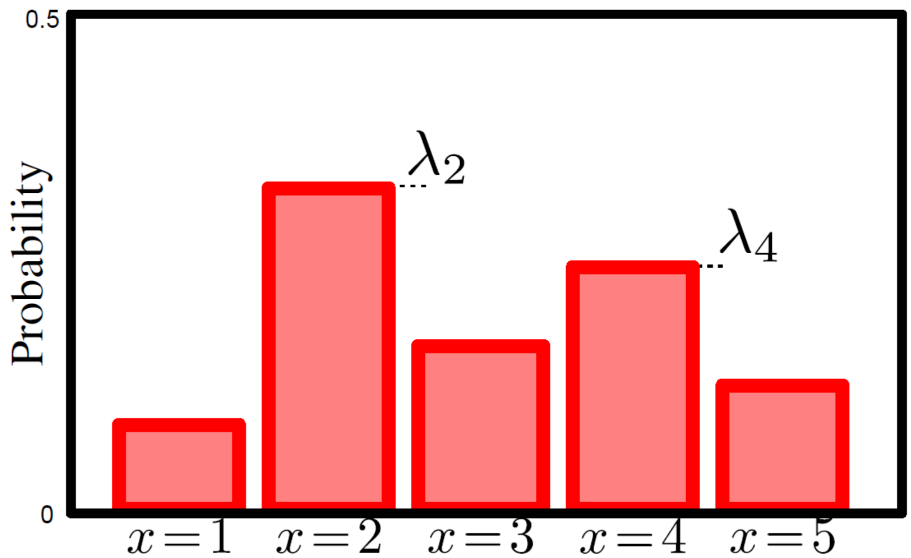
For short we write:

$$\boxed{p(x) = \text{Ber}(x|\lambda)}$$

Bernoulli distribution describes situation where only two possible outcomes x = 0 / x = 1 (e.g. failure/success)

Takes a single parameter $\lambda \in [0,1]$

Categorical Distribution



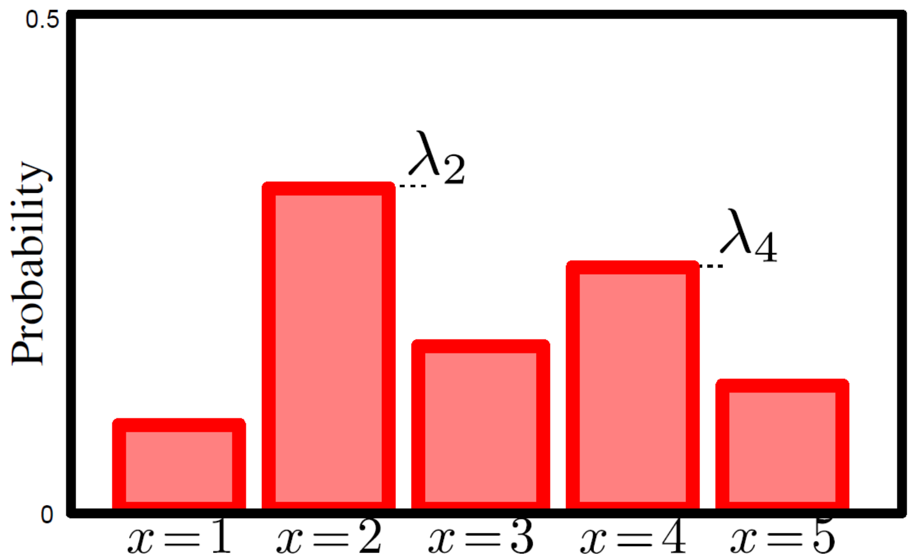
Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0, 1]$ where $\sum_{k=1}^{K} p(X = k) = \sum_{k=1}^{K} \lambda_k = 1$

$\lambda = \{\lambda_1, \cdots, \lambda_K\}$

Categorical Distribution

$$Pr(x = k) = \lambda_k$$

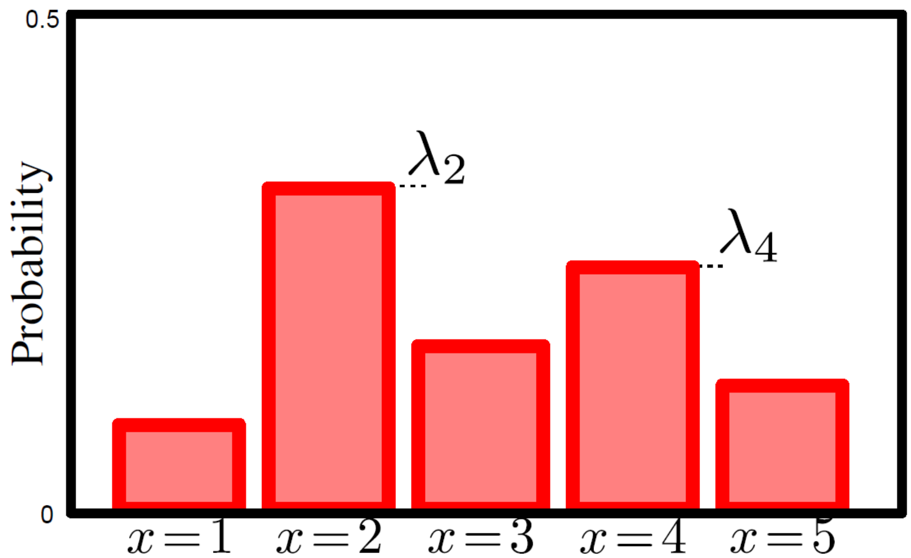


Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0,1]$ where $\sum_{k=1}^{K} p(X=k) = \sum_{k=1}^{K} \lambda_k = 1$

$\lambda = \{\lambda_1, \cdots, \lambda_K\}$

Categorical Distribution



$$Pr(x = k) = \lambda_k$$

or can think of data as vector with all elements zero except k^th e.g. $\mathbf{e}_4$ = [0,0,0,1,0]

$$Pr(x = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{\mathbf{e}_{kj}} = \lambda_k$$

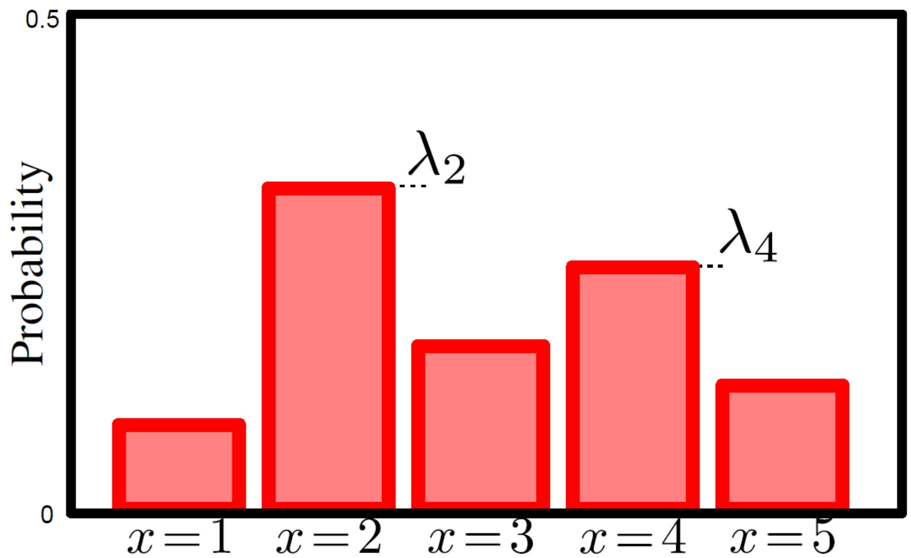
where $\mathbf{e}_{kj}$ is the j-th element of $\mathbf{e}_k$

Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0, 1]$ where $\sum_{k=1}^{K} p(X = k) = \sum_{k=1}^{K} \lambda_k = 1$

$\lambda = \{\lambda_1, \cdots, \lambda_K\}$

Categorical Distribution



$$Pr(x = k) = \lambda_k$$

or can think of data as vector with all elements zero except k[th] e.g. $\mathbf{e}_4$ = [0,0,0,1,0]

$$Pr(x = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{\mathbf{e}_{kj}} = \lambda_k$$

where $\mathbf{e}_{kj}$ is the j-th element of $\mathbf{e}_k$

For short we write: $\boxed{p(x) = \text{Cat}(x|\lambda)}$

Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0, 1]$ where $\sum_{k=1}^{K} p(X = k) = \sum_{k=1}^{K} \lambda_k = 1$

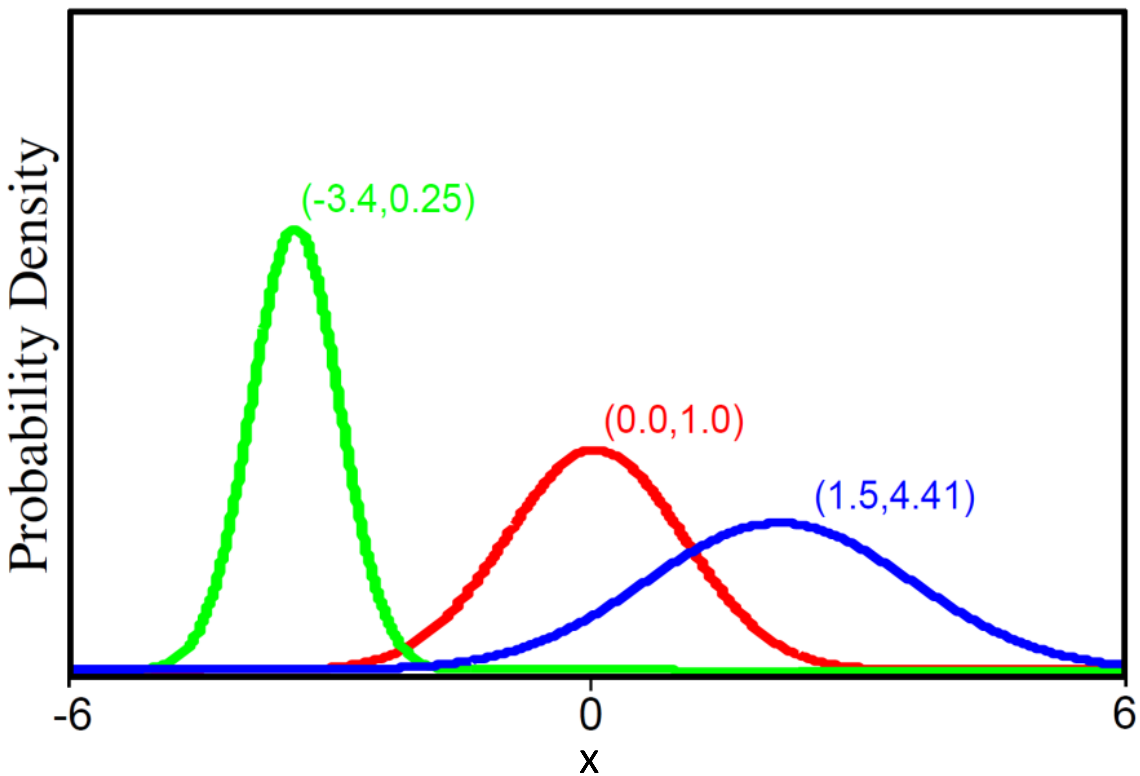
$\lambda = \{\lambda_1, \cdots, \lambda_K\}$

Famous Continuous Random Variables

- Gaussian: http://en.wikipedia.org/wiki/Normal_distribution
- Uniform: http://en.wikipedia.org/wiki/Uniform_distribution_(continuous)
- Exponential: http://en.wikipedia.org/wiki/Exponential_distribution
- Beta: http://en.wikipedia.org/wiki/Beta_distribution
- …

Gaussian/Normal Distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



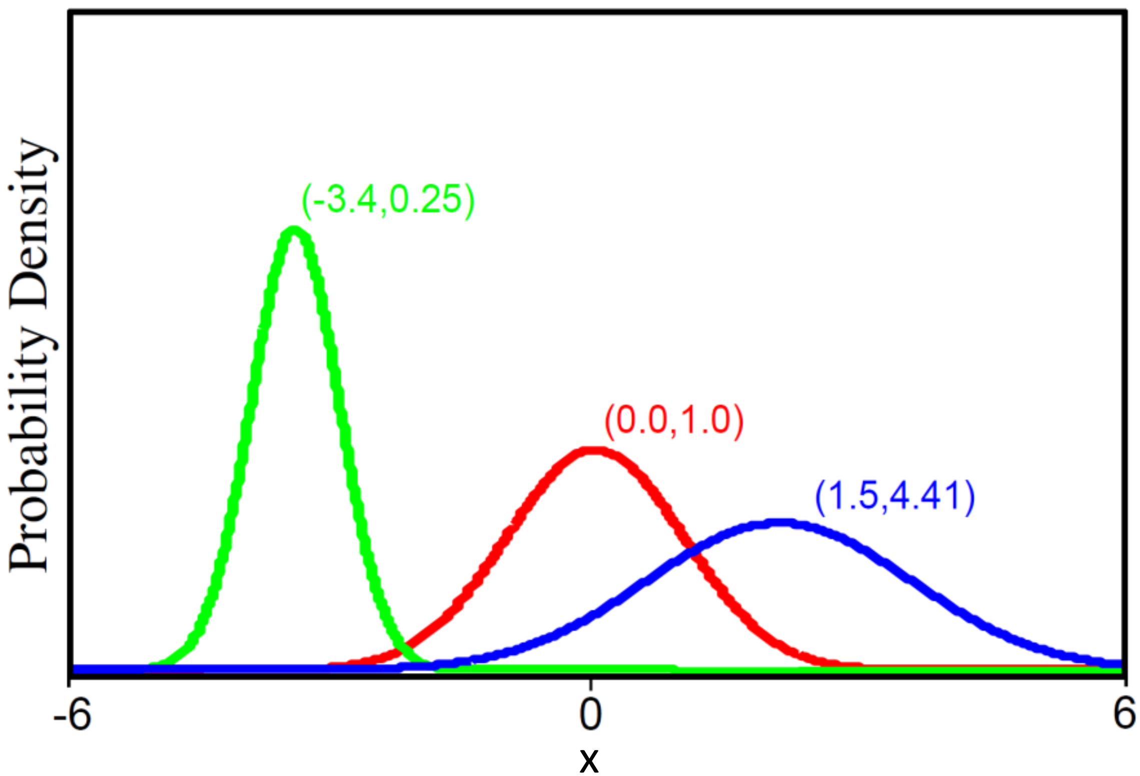
2 parameters mean μ and variance $\sigma^2>0$

Adapted from S. Prince

Gaussian/Normal Distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

≤ 0



2 parameters mean μ and variance $\sigma^2>0$

Adapted from S. Prince

Questions?

Joint Probability

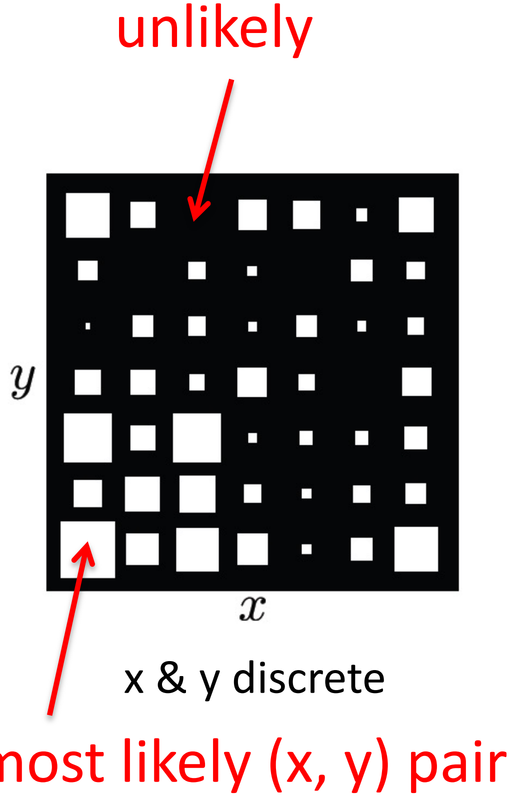
- If we observe two random variables x & y multiple times, then some combinations of outcomes more likely than others

- This information captured by joint probability distribution

- Written as p(x, y), which is read as "joint probability distribution of x and y"

Adapted from S. Prince

Joint Probability p(x, y)



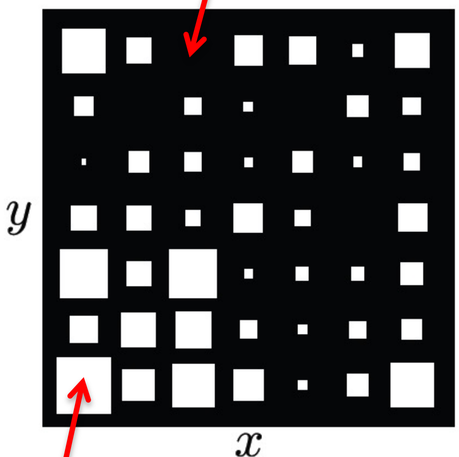
x & y discrete

$$\sum_y \sum_x p(x, y) = 1$$

Adapted from S. Prince

Joint Probability p(x, y)

unlikely

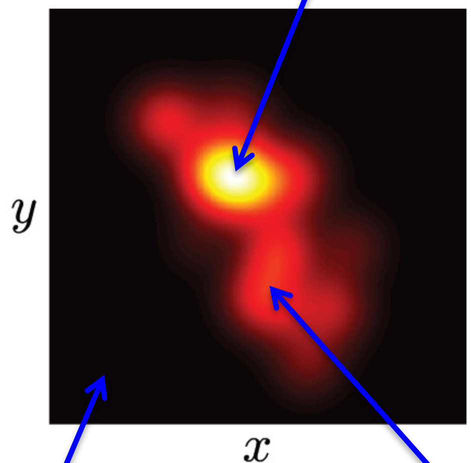


x & y discrete

most likely (x, y) pair

$$\sum_y \sum_x p(x, y) = 1$$

most likely (x, y) pair



x & y continuous

unlikely

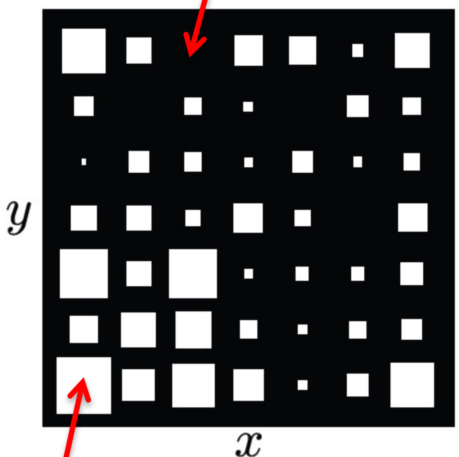
somewhat likely

$$\int_y \int_x p(x, y) dx dy = 1$$

Adapted from S. Prince

Joint Probability p(x, y)

unlikely

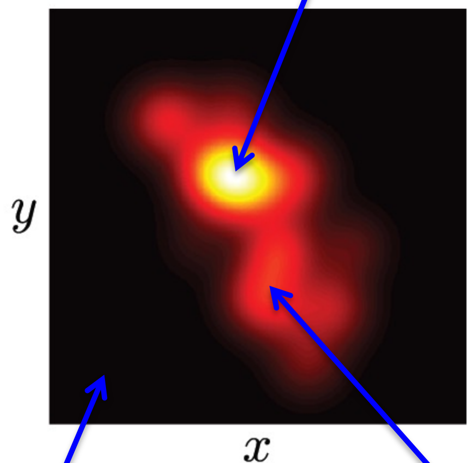


x & y discrete

most likely (x, y) pair

$$\sum_y \sum_x p(x,y) = 1$$

most likely (x, y) pair

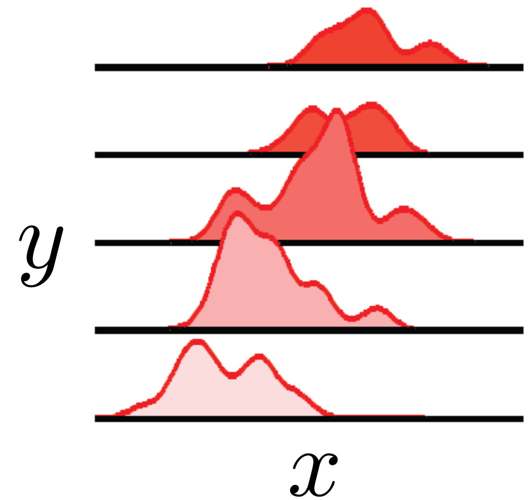


x & y continuous

unlikely

somewhat likely

$$\int_y \int_x p(x,y)dxdy = 1$$



x is continuous,
y is discrete

$$\sum_y \int_x p(x,y)dx = 1$$

Adapted from S. Prince

Questions?

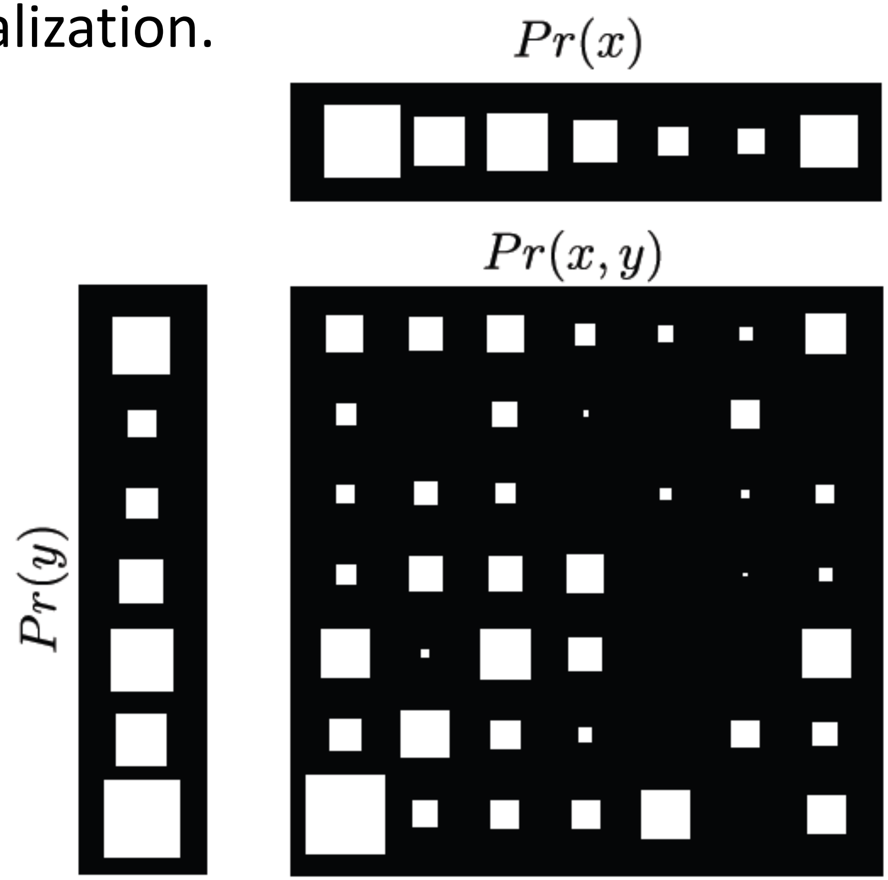
Marginalization / Law of Total Probability / Sum Rule

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variable(s). This is called marginalization.

$$p(x) = \sum_y p(x, y)$$

$$p(y) = \sum_x p(x, y)$$



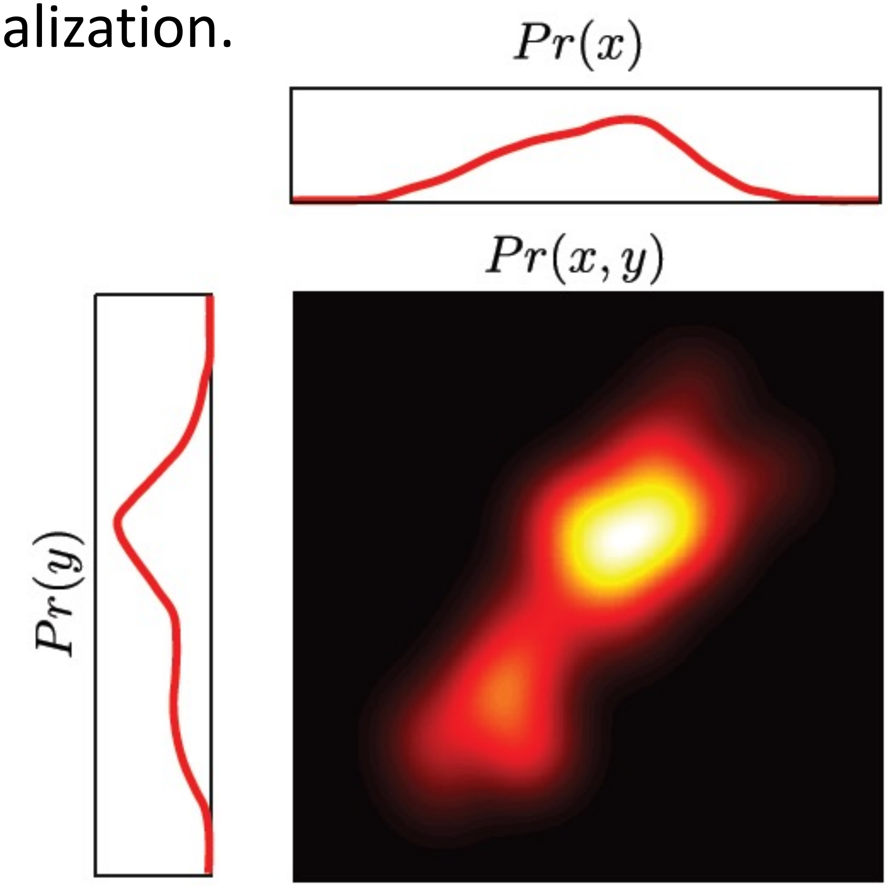
Adapted from S. Prince

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variable(s). This is called marginalization.

$$p(x) = \int_y p(x,y)dy$$

$$p(y) = \int_x p(x,y)dx$$



Adapted from S. Prince

Marginalization Example

p(x, y)

		x	
		0	2.5
y	-3	0	1/2
	-1	1/8	1/4
	2	1/8	0

Marginalization Example

$$p(y) = \sum_x p(x, y)$$

p(x, y)

		x	
		0	2.5
y	-3	0	1/2
	-1	1/8	1/4
	2	1/8	0

p(y)?

p(x)?

$$p(x) = \sum_y p(x, y)$$

Marginalization Example

$$p(y) = \sum_x p(x, y)$$

p(x, y)

		x		
		0	2.5	p(y)?
y	-3	0	1/2	
	-1	1/8	1/4	
	2	1/8	0	
	p(x)?	1/4	3/4	

$$p(x) = \sum_y p(x, y)$$

Marginalization Example

$$p(y) = \sum_x p(x, y)$$

p(x, y)

		x		
		0	2.5	p(y)?
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)?	1/4	3/4	

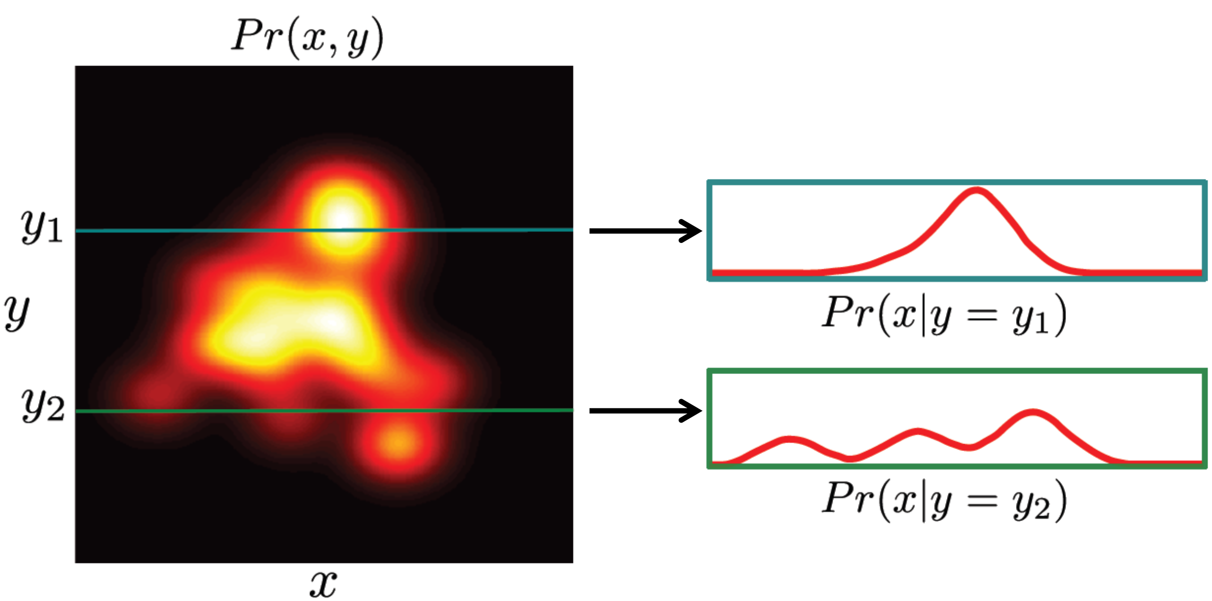
$$p(x) = \sum_y p(x, y)$$

Questions?

Conditional Probability

Conditional Probability

- Suppose we observe y to be y_1, then $p(x \mid y = y_1)$ is how likely x will take on various values given this observation
- $p(x \mid y = y_1)$ read as "conditional probability of X given Y is equal to y_1"



$$\int_x p(x|y=y_1)dx = 1$$

$$\int_x p(x|y=y_2)dx = 1$$

Adapted from S. Prince

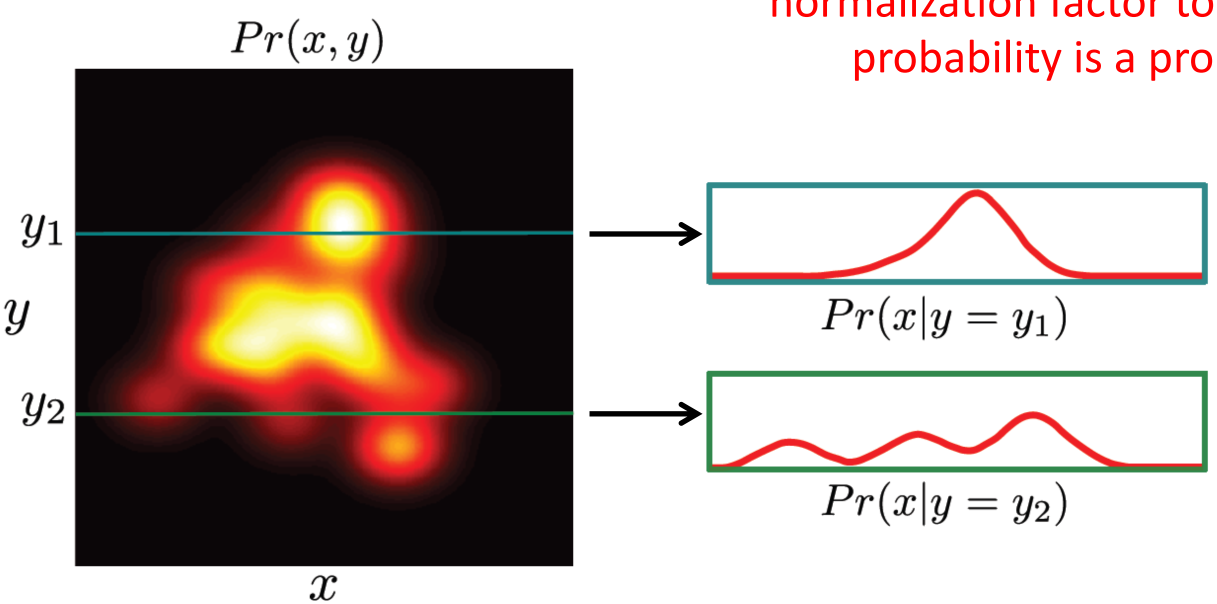
Conditional Probability

- Conditional probability can be computed from joint probability

slice of joint distribution

$$p(x|y = y^*) = \frac{p(x, y = y^*)}{p(y = y^*)}$$

normalization factor to ensure conditional probability is a proper distribution



$$\int_x p(x|y = y_1)dx = 1$$

$$\int_x p(x|y = y_2)dx = 1$$

Adapted from S. Prince

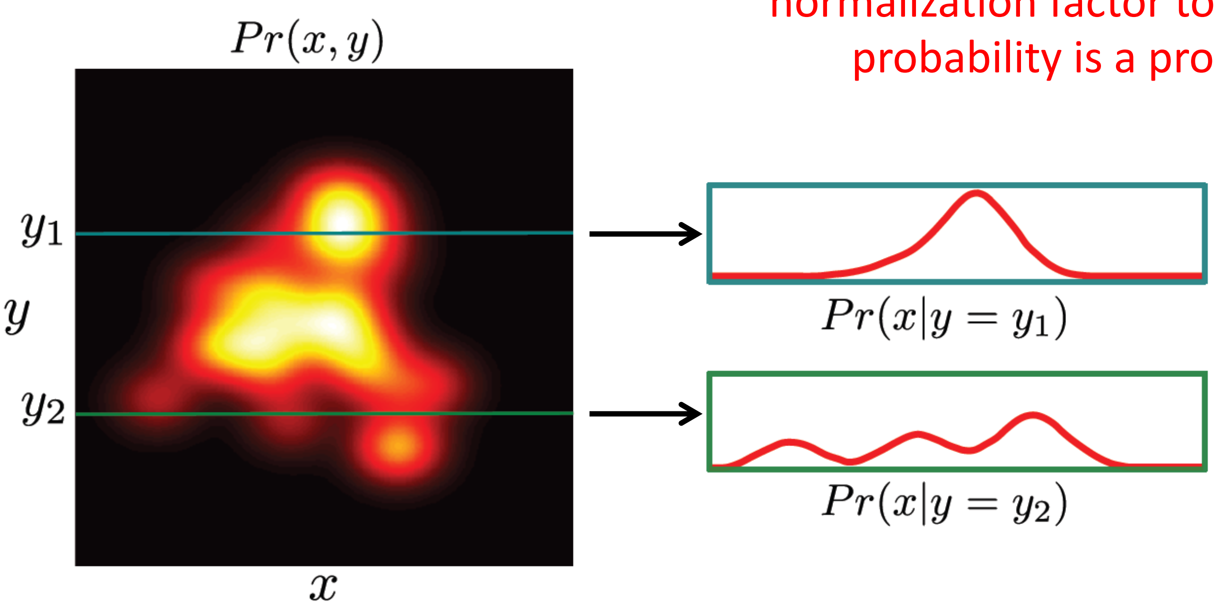
Conditional Probability

- Conditional probability can be computed from joint probability

slice of joint distribution

$$p(x|y=y^*) = \frac{p(x, y=y^*)}{p(y=y^*)} = \frac{p(x, y=y^*)}{\int p(x, y=y^*)dx}$$

normalization factor to ensure conditional probability is a proper distribution



$$\int_x p(x|y=y_1)dx = 1$$

$$\int_x p(x|y=y_2)dx = 1$$

Adapted from S. Prince

Conditional Probability

$$p(x|y=y^*) = \frac{p(x,y=y^*)}{p(y=y^*)} = \frac{p(x,y=y^*)}{\int p(x,y=y^*)dx}$$

- More usually written in compact form

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

Adapted from S. Prince

Conditional Probability

$$p(x|y = y^*) = \frac{p(x, y = y^*)}{p(y = y^*)} = \frac{p(x, y = y^*)}{\int p(x, y = y^*)dx}$$

- More usually written in compact form

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

- Can be re-arranged to give

$$p(x, y) = p(y)p(x|y)$$
$$p(x, y) = p(x)p(y|x)$$

Adapted from S. Prince

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

Conditional Probability Example

		x		
		0	2.5	p(y)
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x,y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) =$$

$$p(x=2.5|y=-1) =$$

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) = \frac{p(x=0, y=-1)}{p(y=-1)}$$

$$p(x=2.5|y=-1) =$$

Conditional Probability Example

		x		
		0	2.5	p(y)
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) = \frac{p(x=0, y=-1)}{p(y=-1)} = \frac{1/8}{3/8} = \frac{1}{3}$$

$$p(x=2.5|y=-1) =$$

Conditional Probability Example

		x		
		0	2.5	p(y)
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) = \frac{p(x=0, y=-1)}{p(y=-1)} = \frac{1/8}{3/8} = \frac{1}{3}$$

$$p(x=2.5|y=-1) = \frac{p(x=2.5, y=-1)}{p(y=-1)} = \frac{1/4}{3/8} = \frac{2}{3}$$

Questions?

Bayes' Rule

Deriving Bayes' Rule (y continuous)

From before:

$$p(x,y) = p(y)p(x|y)$$

$$p(x,y) = p(x)p(y|x)$$

Equate RHS

Adapted from S. Prince

Deriving Bayes' Rule (y continuous)

From before:

$$p(x, y) = p(y)p(x|y)$$

$$p(x, y) = p(x)p(y|x)$$

Equate RHS

Combining:

$$p(y)p(x|y) = p(x)p(y|x)$$

Adapted from S. Prince

Deriving Bayes' Rule (y continuous)

From before:

$$p(x,y) = p(y)p(x|y)$$

$$p(x,y) = p(x)p(y|x)$$

Equate RHS

Combining:

$$p(y)p(x|y) = p(x)p(y|x)$$

Re-arranging:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

Adapted from S. Prince

Deriving Bayes' Rule (y continuous)

From before:

$$p(x,y) = p(y)p(x|y)$$

$$p(x,y) = p(x)p(y|x)$$

Equate RHS

Combining:

$$p(y)p(x|y) = p(x)p(y|x)$$

Re-arranging:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

$$= \frac{p(y)p(x|y)}{\int p(x,y)dy}$$

$$p(x) = \int p(x,y)dy$$

Adapted from S. Prince

Deriving Bayes' Rule (y continuous)

From before:

$$p(x,y) = p(y)p(x|y)$$

$$p(x,y) = p(x)p(y|x)$$

Equate RHS

Combining:

$$p(y)p(x|y) = p(x)p(y|x)$$

Re-arranging:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

$$p(x) = \int p(x,y)dy$$

$$= \frac{p(y)p(x|y)}{\int p(x,y)dy}$$

$$p(x,y) = p(y)p(x|y)$$

$$= \frac{p(y)p(x|y)}{\int p(y)p(x|y)dy}$$

Adapted from S. Prince

Deriving Bayes' Rule (y discrete)

From before:

$$p(x, y) = p(y)p(x|y)$$

$$p(x, y) = p(x)p(y|x)$$

Equate RHS

Combining:

$$p(y)p(x|y) = p(x)p(y|x)$$

Re-arranging:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

$$p(x) = \sum_y p(x, y)$$

$$= \frac{p(y)p(x|y)}{\sum_y p(x, y)}$$

$$p(x, y) = p(y)p(x|y)$$

$$= \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

Adapted from S. Prince

Bayes' Rule

Prior – what we know about y BEFORE seeing x

Likelihood – propensity for observing a certain value of x given a certain value of y

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

Posterior – what we know about y AFTER seeing x

Evidence – a constant to ensure that the left hand side is a valid distribution

Adapted from S. Prince

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

$$p(y = 1|x = 1) = \frac{p(y = 1)p(x = 1|y = 1)}{p(y = 1)p(x = 1|y = 1) + p(y = 0)p(x = 1|y = 0)}$$

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

$$
\begin{aligned}
p(y = 1|x = 1) &= \frac{p(y = 1)p(x = 1|y = 1)}{p(y = 1)p(x = 1|y = 1) + p(y = 0)p(x = 1|y = 0)} \\
&= \frac{0.004 \times 0.8}{0.004 \times 0.8 + 0.996 \times 0.1} = 0.031
\end{aligned}
$$

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

$$
\begin{aligned}
p(y=1|x=1) &= \frac{p(y=1)p(x=1|y=1)}{p(y=1)p(x=1|y=1) + p(y=0)p(x=1|y=0)} \\
&= \frac{0.004 \times 0.8}{0.004 \times 0.8 + 0.996 \times 0.1} = 0.031
\end{aligned}
$$

- US government no longer recommend mammogram for women in 40s

Questions?

Independence

Independence

- If x & y are independent, then knowing x tells us nothing about y (and vice versa):

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

Independence

- If x & y are independent, then knowing x tells us nothing about y (and vice versa):

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

- If x & y are independent, then joint distribution factorizes into product of marginal distributions:

$$\begin{aligned} p(x, y) &= p(x)p(y|x) \\ &= p(x)p(y) \end{aligned}$$

Independence

- If x & y are independent, then knowing x tells us nothing about y (and vice versa):

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

- If x & y are independent, then joint distribution factorizes into product of marginal distributions:

$$\begin{aligned} p(x, y) &= p(x)p(y|x) \\ &= p(x)p(y) \end{aligned}$$

- Conversely, if joint distribution can be factorized into product of marginal distributions, then x & y are independent

Questions?

N Random Variables

N random variables (aka random vector)

- Have focused on 2 random variables x and y

N random variables (aka random vector)

- Have focused on 2 random variables x and y
- In real applications, usually more than 2 variables (e.g., photo has > 1M pixels)

N random variables (aka random vector)

- Have focused on 2 random variables x and y
- In real applications, usually more than 2 variables (e.g., photo has > 1M pixels)
- If we observe x_1, x_2, …, x_N multiple times, some combinations of outcomes more likely than others

N random variables (aka random vector)

- Have focused on 2 random variables x and y
- In real applications, usually more than 2 variables (e.g., photo has > 1M pixels)
- If we observe x_1, x_2, …, x_N multiple times, some combinations of outcomes more likely than others
- This information captured by joint probability distribution function

N random variables (aka random vector)

- Have focused on 2 random variables x and y
- In real applications, usually more than 2 variables (e.g., photo has > 1M pixels)
- If we observe $x_1, x_2, \ldots, x_N$ multiple times, some combinations of outcomes more likely than others
- This information captured by joint probability distribution function
- Written as $p(x_1, x_2, \ldots, x_N)$, read as probability distribution of x_1 to x_N
- If $x_1, x_2, \ldots, x_N$ continuous, then p refers to joint probability distribution function (pdf). If discrete, then refers to joint probability mass function (pmf)

N random variables (aka random vector)

- Have focused on 2 random variables x and y
- In real applications, usually more than 2 variables (e.g., photo has > 1M pixels)
- If we observe $x_1, x_2, \ldots, x_N$ multiple times, some combinations of outcomes more likely than others
- This information captured by joint probability distribution function
- Written as $p(x_1, x_2, \ldots, x_N)$, read as probability distribution of x_1 to x_N
- If $x_1, x_2, \ldots, x_N$ continuous, then p refers to joint probability distribution function (pdf). If discrete, then refers to joint probability mass function (pmf)
- Many properties for two random variables generalize naturally to more variables

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables

$$Pr(x) = \int Pr(x,y)\,dy$$

$$Pr(y) = \int Pr(x,y)\,dx$$

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables

$$
\begin{aligned}
Pr(x) &= \int Pr(x,y)\, dy \\
Pr(y) &= \int Pr(x,y)\, dx
\end{aligned}
$$

Works in higher dimensions as well – leaves joint distribution between whatever variables are left

$$Pr(x,y) = \sum_{w} \int Pr(w,x,y,z)\, dz$$

Conditional Probability

- Two variables

$$p(x, y) = p(x)p(y|x)$$

Conditional Probability

- Two variables

$$p(x, y) = p(x)p(y|x)$$

- Three variables

$$p(a, b, c) = p(a)p(b, c|a) = p(a)p(b|a)p(c|a, b)$$

Conditional Probability

- Two variables

$$p(x, y) = p(x)p(y|x)$$

- Three variables

$$p(a, b, c) = p(a)p(b, c|a) = p(a)p(b|a)p(c|a, b)$$

- N variables

$$\begin{aligned} p(x_1, \cdots, x_N) &= p(x_1)p(x_2, \cdots, x_N|x_1) \\ &= p(x_1)p(x_2|x_1)p(x_3, \cdots, x_N|x_1, x_2) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_N|x_1, \cdots, x_{N-1}) \end{aligned}$$

Independence

- If $x_1, \cdots, x_N$ are independent, then knowing any subset of x's tells us nothing about the remaining x's

Independence

- If $x_1, \cdots, x_N$ are independent, then knowing any subset of x's tells us nothing about the remaining x's

- If $x_1, \cdots, x_N$ are independent if and only if the joint distribution factorizes into product of marginal distributions:

$$p(x_1, \cdots, x_N) = p(x_1)p(x_2)\cdots p(x_N) \triangleq \prod_{n=1}^{N} p(x_n)$$

Independence

- If $x_1, \cdots, x_N$ are independent, then knowing any subset of x's tells us nothing about the remaining x's

- If $x_1, \cdots, x_N$ are independent <u>if and only if</u> the joint distribution factorizes into product of marginal distributions:

$$p(x_1, \cdots, x_N) = p(x_1)p(x_2)\cdots p(x_N) \triangleq \prod_{n=1}^{N} p(x_n)$$

- $x_1, \cdots, x_N$ are <u>independently and identically distributed</u> (i.i.d.) if they are independent and $p(x_1) = p(x_2) = \cdots = p(x_N)$

Conditional Independence

- x_1 and x_2 are conditionally independent given x_3 if and only if

$$p(x_1, x_2|x_3) = p(x_1|x_3)p(x_2|x_3)$$

Knowing x_2 tells us nothing about x_1 (and vice versa) if we already know x_3

Maximum-A-Posterior (MAP) and Maximum Likelihood (ML) Estimation

Digression: argmax and argmin

- $\operatorname{argmax}_x f(x)$ is value of x where $f(x)$ is biggest

Digression: argmax and argmin

- $\text{argmax}_x f(x)$ is value of x where $f(x)$ is biggest
- $\text{argmin}_x f(x)$ is value of x where $f(x)$ is smallest

Digression: argmax and argmin

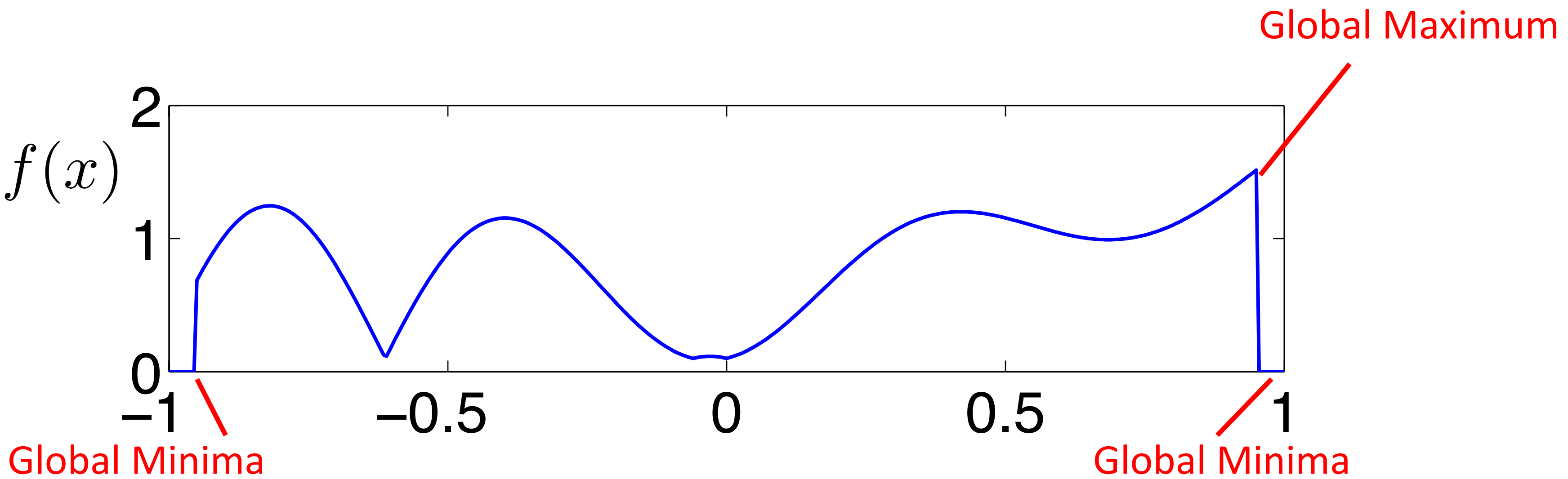
- $\text{argmax}_x\, f(x)$ is value of x where $f(x)$ is biggest
- $\text{argmin}_x\, f(x)$ is value of x where $f(x)$ is smallest
- $f(x) = \begin{cases} |\sin(4x)^2 + x|\exp(-x) + x^2 + 0.1, & -0.95 \leq x \leq 0.95 \\ 0 & \text{otherwise} \end{cases}$

Digression: argmax and argmin

- $\operatorname{argmax}_x f(x)$ is value of x where $f(x)$ is biggest
- $\operatorname{argmin}_x f(x)$ is value of x where $f(x)$ is smallest
- $f(x) = \begin{cases} |\sin(4x)^2 + x| \exp(-x) + x^2 + 0.1, & -0.95 \leq x \leq 0.95 \\ 0 & \text{otherwise} \end{cases}$
  - Generally easier to evaluate $f(x)$ than find maximum or minimum

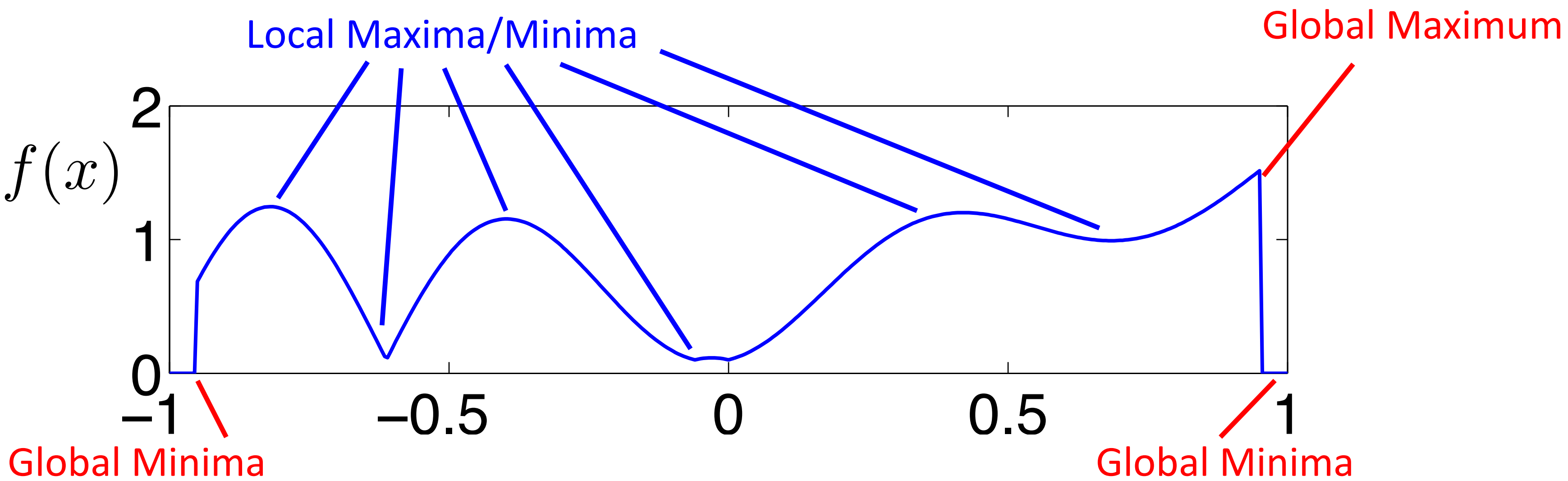
Digression: argmax and argmin

- $\text{argmax}_x\, f(x)$ is value of x where $f(x)$ is biggest
- $\text{argmin}_x\, f(x)$ is value of x where $f(x)$ is smallest
- $f(x) = \begin{cases} |\sin(4x)^2 + x|\exp(-x) + x^2 + 0.1, & -0.95 \leq x \leq 0.95 \\ 0 & \text{otherwise} \end{cases}$
  - Generally easier to evaluate $f(x)$ than find maximum or minimum



Digression: argmax and argmin

- $\mathrm{argmax}_x\, f(x)$ is value of x where $f(x)$ is biggest
- $\mathrm{argmin}_x\, f(x)$ is value of x where $f(x)$ is smallest
- $f(x) = \begin{cases} |\sin(4x)^2 + x|\exp(-x) + x^2 + 0.1, & -0.95 \leq x \leq 0.95 \\ 0 & \text{otherwise} \end{cases}$
  - Generally easier to evaluate $f(x)$ than find maximum or minimum
  - Real problems: may have to live with local maximum or minimum



MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$y_{MAP} \triangleq \operatorname*{argmax}_{y} p(y|x)$$

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$y_{MAP} \stackrel{\triangle}{=} \underset{y}{\operatorname{argmax}}\, p(y|x)$$

Bayes' rule

$$= \underset{y}{\operatorname{argmax}}\, \frac{p(y)p(x|y)}{p(x)}$$

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$
\begin{aligned}
y_{MAP} &\stackrel{\triangle}{=} \operatorname*{argmax}_y p(y|x) \\
&= \operatorname*{argmax}_y \frac{p(y)p(x|y)}{p(x)} \\
&= \operatorname*{argmax}_y p(y)p(x|y)
\end{aligned}
$$

Bayes' rule

p(x) not function of y

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$\begin{aligned} y_{MAP} &\stackrel{\triangle}{=} \underset{y}{\operatorname{argmax}}\, p(y|x) \\ &= \underset{y}{\operatorname{argmax}}\, \frac{p(y)p(x|y)}{p(x)} \\ &= \underset{y}{\operatorname{argmax}}\, p(y)p(x|y) \end{aligned}$$

Bayes' rule

p(x) not function of y

- Example: $p(y = \text{chair}|x = \text{photo}) = 0.6$, $p(y = \text{human}|x = \text{photo}) = 0.3$, $p(y = \text{cat}|x = \text{photo}) = 0.1 \implies y_{MAP} = \text{chair}$

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$\begin{aligned} y_{MAP} &\triangleq \operatorname*{argmax}_y p(y|x) \\ &= \operatorname*{argmax}_y \frac{p(y)p(x|y)}{p(x)} \\ &= \operatorname*{argmax}_y p(y)p(x|y) \end{aligned}$$

Bayes' rule

p(x) not function of y

- Example: $p(y = \text{chair}|x = \text{photo}) = 0.6$, $p(y = \text{human}|x = \text{photo}) = 0.3$, $p(y = \text{cat}|x = \text{photo}) = 0.1 \implies y_{MAP} = \text{chair}$
- Example: $p(y = \text{chair}|x = \text{photo}) \propto 1.2$, $p(y = \text{human}|x = \text{photo}) \propto 0.6$, $p(y = \text{cat}|x = \text{photo}) \propto 0.2$

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$y_{MAP} \stackrel{\triangle}{=} \operatorname*{argmax}_y p(y|x)$$

Bayes' rule

$$= \operatorname*{argmax}_y \frac{p(y)p(x|y)}{p(x)}$$

p(x) not function of y

$$= \operatorname*{argmax}_y p(y)p(x|y)$$

- Example: $p(y = \text{chair}|x = \text{photo}) = 0.6$, $p(y = \text{human}|x = \text{photo}) = 0.3$, $p(y = \text{cat}|x = \text{photo}) = 0.1 \implies y_{MAP} = \text{chair}$

- Example: $p(y = \text{chair}|x = \text{photo}) \propto 1.2$, $p(y = \text{human}|x = \text{photo}) \propto 0.6$, $p(y = \text{cat}|x = \text{photo}) \propto 0.2$
  - Notice $1.2 + 0.6 + 0.2 = 2$, so not a valid probability distribution, which is why we use "$\propto$" rather than "$=$" but y_{MAP} still the same (chair)

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$
\begin{aligned}
y_{MAP} &\triangleq \underset{y}{\operatorname{argmax}}\, p(y|x) \\
&= \underset{y}{\operatorname{argmax}}\, \frac{p(y)p(x|y)}{p(x)} \\
&= \underset{y}{\operatorname{argmax}}\, p(y)p(x|y)
\end{aligned}
$$

Bayes' rule

p(x) not function of y

- Example: $p(y = \text{chair}|x = \text{photo}) = 0.6$, $p(y = \text{human}|x = \text{photo}) = 0.3$, $p(y = \text{cat}|x = \text{photo}) = 0.1 \implies y_{MAP} = \text{chair}$

- Example: $p(y = \text{chair}|x = \text{photo}) \propto 1.2$, $p(y = \text{human}|x = \text{photo}) \propto 0.6$, $p(y = \text{cat}|x = \text{photo}) \propto 0.2$

  - Notice $1.2 + 0.6 + 0.2 = 2$, so not a valid probability distribution, which is why we use "$\propto$" rather than "$=$" but y_{MAP} still the same (chair)
  - If we want proper distribution, $c = 1.2+0.6+0.2 = 2$, so we can normalize to become proper distribution: $p(y = \text{chair}|x = \text{photo}) = 1.2/c = 0.6$, $p(y = \text{human}|x = \text{photo}) = 0.6/c = 0.3$, $p(y = \text{cat}|x = \text{photo}) = 0.2/c = 0.1$

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$
\begin{aligned}
y_{MAP} &\stackrel{\triangle}{=} \operatorname*{argmax}_y p(y|x) \\
&= \operatorname*{argmax}_y \frac{p(y)p(x|y)}{p(x)} \\
&= \operatorname*{argmax}_y p(y)p(x|y)
\end{aligned}
$$

Bayes' rule

p(x) not function of y

- Prior $p(y)$ is constant (uniform) $\implies$ maximum likelihood (ML) estimate

$$
y_{MAP} = \operatorname*{argmax}_y p(x|y) \stackrel{\triangle}{=} y_{ML}
$$

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$
\begin{aligned}
y_{MAP} &\triangleq \operatorname*{argmax}_y p(y|x) \\
&= \operatorname*{argmax}_y \frac{p(y)p(x|y)}{p(x)} \\
&= \operatorname*{argmax}_y p(y)p(x|y)
\end{aligned}
$$

Bayes' rule

p(x) not function of y

- Prior $p(y)$ is constant (uniform) $\implies$ maximum likelihood (ML) estimate

$$
y_{MAP} = \operatorname*{argmax}_y p(x|y) \triangleq y_{ML}
$$

- ML often easier to compute; often use when prior is unknown (or do not want to assume priors)

MAP and ML Estimation

- Maximum-A-Posteriori (MAP) estimation:

$$
\begin{aligned}
y_{MAP} &\triangleq \underset{y}{\operatorname{argmax}}\, p(y|x) \\
&= \underset{y}{\operatorname{argmax}} \frac{p(y)p(x|y)}{p(x)} \\
&= \underset{y}{\operatorname{argmax}}\, p(y)p(x|y)
\end{aligned}
$$

Bayes' rule

p(x) not function of y

- Prior $p(y)$ is constant (uniform) $\implies$ maximum likelihood (ML) estimate

$$y_{MAP} = \underset{y}{\operatorname{argmax}}\, p(x|y) \triangleq y_{ML}$$

- ML often easier to compute; often use when prior is unknown (or do not want to assume priors)

- If # samples goes to infinity (infinite amount of data), then $\lim_{N\to\infty} y_{MAP} = y_{ML}$

What is hard here?

- ML is a special case of MAP, so let's focus on MAP (for now)
- In previous example of chair, human & cat, I gave you $p(y \mid x = \text{photo})$, but how to get $p(y \mid x = \text{photo})$ in the first place?
- Much of machine learning is about how to choose a model for $p(y \mid x)$ and how to evaluate/optimize model parameters

Questions?

Probabilistic Estimation of Model Parameters

Probabilistic Estimation of Model Parameters

- Example:
  - p(y = chair | x = photo) = 0.6
  - p(y = human | x = photo) = 0.3
  - p(y = cat | x = photo) = 0.1
  - How can 0.6, 0.3 and 0.1 appear on right side, but not left side of "=" sign?
- We should technically include $\boldsymbol{\theta} = \begin{bmatrix} 0.6 \\ 0.3 \\ 0.1 \end{bmatrix}$ as variable on left hand side
  - p(y = chair | x = photo, $\boldsymbol{\theta}$) = $\boldsymbol{\theta}_1$ = 0.6
  - p(y = human | x = photo, $\boldsymbol{\theta}$) = $\boldsymbol{\theta}_2$ = 0.3
  - p(y = cat | x = photo, $\boldsymbol{\theta}$) = $\boldsymbol{\theta}_3$ = 0.1
  - In this example, posterior distribution is categorical distribution with parameter $\boldsymbol{\theta}$
- In general, $\boldsymbol{\theta}$ needs to be learned from the training set for both generative models p(x, y | $\boldsymbol{\theta}$) & discriminative models p(y | x, $\boldsymbol{\theta}$)
- In first bullet point on this slide: y & x are concrete things (photo, chair, human, cat), but ML/MAP can also be used to estimate "abstract" quantities like $\boldsymbol{\theta}$. In other words, we can also treat parameters of probability distribution as random variables and estimate them using ML/MAP

Parameters of Probability Distribution can themselves be treated as random variables

- Given training set D = $\{x_i, y_i\}_{i=1:N}$, where x = feature, y = target label
- Goal: learn parameters $\boldsymbol{\theta}$ of generative model p(x, y | $\boldsymbol{\theta}$) from D, so that given new test data x, can predict y using MAP estimate of posterior by plugging in estimate of $\boldsymbol{\theta}$: $p(y \mid x, \boldsymbol{\theta}) \propto p(x, y \mid \boldsymbol{\theta}) = p(x \mid y, \boldsymbol{\theta}) p(y \mid \boldsymbol{\theta})$
- Strategy 1 (Maximum likelihood)
  - Step 1: Estimate $\boldsymbol{\theta}_{ML} = \text{argmax}_{\boldsymbol{\theta}}\, p(D \mid \boldsymbol{\theta})$
  - Step 2: Plug in $\boldsymbol{\theta}_{ML}$ into p(x, y | $\boldsymbol{\theta}_{ML}$) and find MAP estimate of y
- Strategy 2 (Maximum-A-Posteriori)
  - Step 1: Estimate $\boldsymbol{\theta}_{MAP} = \text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta} \mid D)$
  - Step 2: Plug in $\boldsymbol{\theta}_{MAP}$ into p(x, y | $\boldsymbol{\theta}_{MAP}$) and find MAP estimate of y

- Since parameters $\boldsymbol{\theta}$ of probability distributions are treated as random variables that can be estimated, let's see how this is done for various distributions

Questions?