NUS National University of Singapore | School of Computing

CS5340
Uncertainty Modeling in AI

Asst. Prof. Harold Soh

Dept of Computer Science

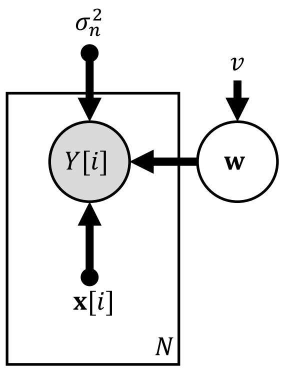
National University of Singapore

Preliminaries

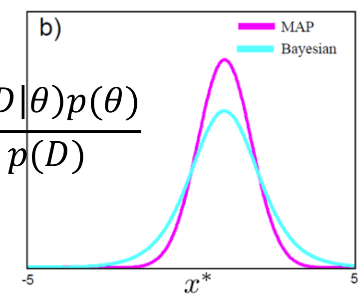
Introduction

CS5340 in a nutshell

CS5340 is about how to "**represent**" and "**reason**" with uncertainty in a computer.



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



Representation: The *language* is probability and probabilistic graphical models (PGM).

The language is used to model problems.

Reasoning: We use learning and inference algorithms to answer questions.

e.g., Belief-propagation/sum-product, MCMC, and variational Bayes

Probabilistic Graphical Modeling

Key Ideas:

- Represent the world as a collection of random variables $X_1, \dots, X_N$ with joint distribution $p(X_1, \dots, X_N)$.
- Learn the distribution from data.
- Perform "inference" (compute conditional distributions $p(X_i \mid X_1 = x_1, \dots, X_N = x_N)$).

What does all this mean precisely?

Adapted from: "Probabilistic Graphical Modeling" Lectures NYU, David Sontag

Learning Outcomes

Students should be able to:

1. Describe uncertain quantities with random variables and joint probabilities.

2. Explain the basic rules of probability – sum, product, Bayes', independence and expectation rules.

3. Use the common probabilities distributions – Bernoulli, categorical, univariate and multivariate normal distributions.

4. Explain the use of conjugate distributions.

Acknowledgements

- A lot of slides and content of this lecture are adopted from:

1. Simon Prince, “Computer Vision: Models, Learning, and Inference”, Chapter 1 and 2.
2. Daphne Koller and Nir Friedman, "Probabilistic graphical models”, Chapter 2.
3. Christopher Bishop, “Pattern Recognition and Machine Learning”, Chapter 2.
4. MIT Course: Mathematics for CS Readings, Chapter 14 and 18.
5. Dr. Lee Gim Hee’s CS5340 slides.

The Basics

Events, Outcomes, and Probability

Problem: You’re not feeling well…

- You’re not feeling well and go to the doctor.
- You take a blood test.
- Test comes back **positive** for *rare*, *fatal* disease.
- Should you:
  A. Skip CS5340 and start planning your funeral?
  B. Not worry. Be Happy.
  C. Take the test again (and again) until it comes back negative.
  D. Ask for more information.

Problem: You're not feeling well…

- You're not feeling well and go to the doctor.
- You take a blood test.
- Test comes back **<u>positive</u>** for *rare*, *fatal* disease.
  - Disease affects 0.1% of the population.
  - Test correctly identifies 99% of the people who have the disease.
  - If you do not have the disease, test may come back positive 2% of the time.

What to do now?

Depends how much you believe whether you have the disease.

What do we **mean** when we say:

- *“the probability that I have the rare fatal disease is 90%”*
- *“the probability of getting an even number when rolling a die is ½”*

Probability Space

- A probability space (Ω, E, P) models a process consisting of outcomes that occur randomly.
- Consists of three parts:
  - Outcome or sample space Ω
  - Event space E
  - Probability function $P: E \rightarrow [0,1]$

Outcome and Event Spaces

- Outcome space is an agreed upon space of possible outcomes, denoted by Ω.

Example: Outcomes of a dice roll, $\Omega = \{1,2,3,4,5,6\}$.

- Event space $E \subseteq 2^{\Omega}$ is a subset of the power set of Ω, it is the set of measurable events to which we assign probabilities.

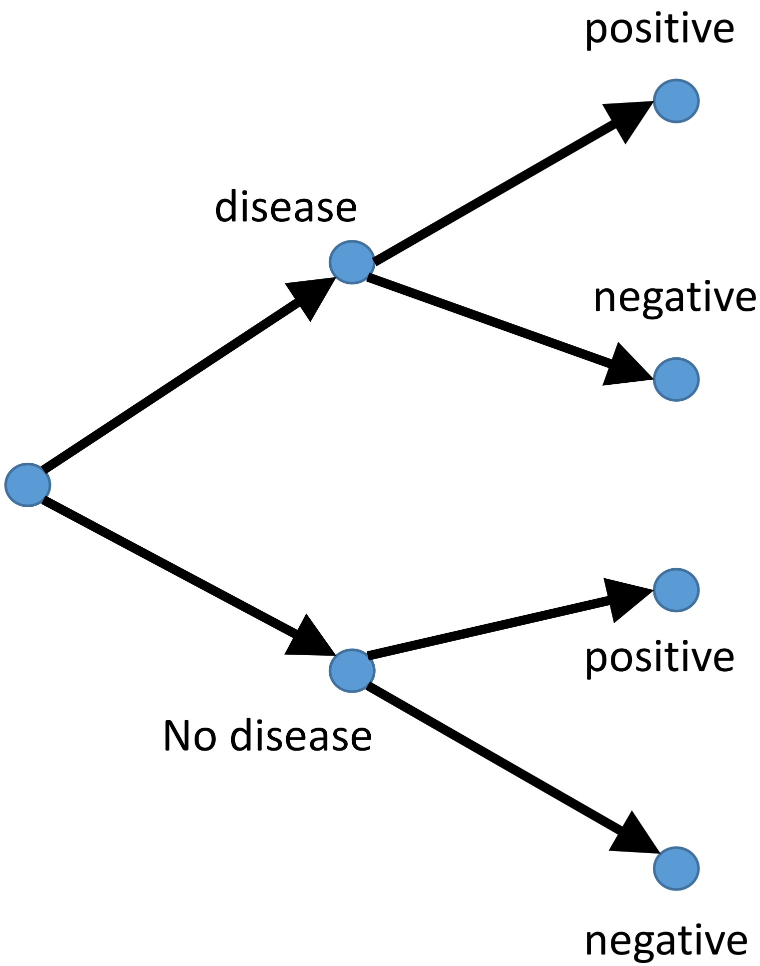
Example: The event space on whether a dice roll is odd or even, $E = \{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}$.

Outcome and Event Spaces

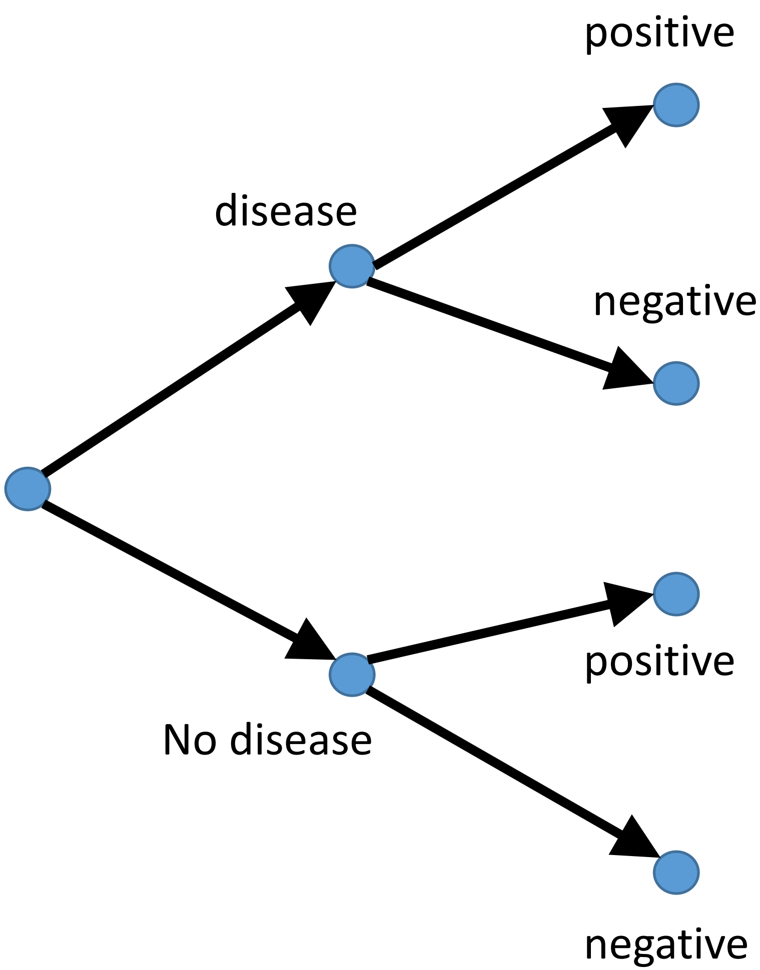
- Event space must satisfy three basic properties:

  1. It contains the empty event $\emptyset$, and the trivial event Ω.
  2. It is closed under union, i.e. if $\alpha, \beta \in E$, then so is $\alpha \cup \beta$.
  3. It is closed under complement, i.e. if $\alpha \in E$, then so is $\Omega - \alpha$.

Tree Diagram Example



Tree Diagram Example



Outcome Space Ω

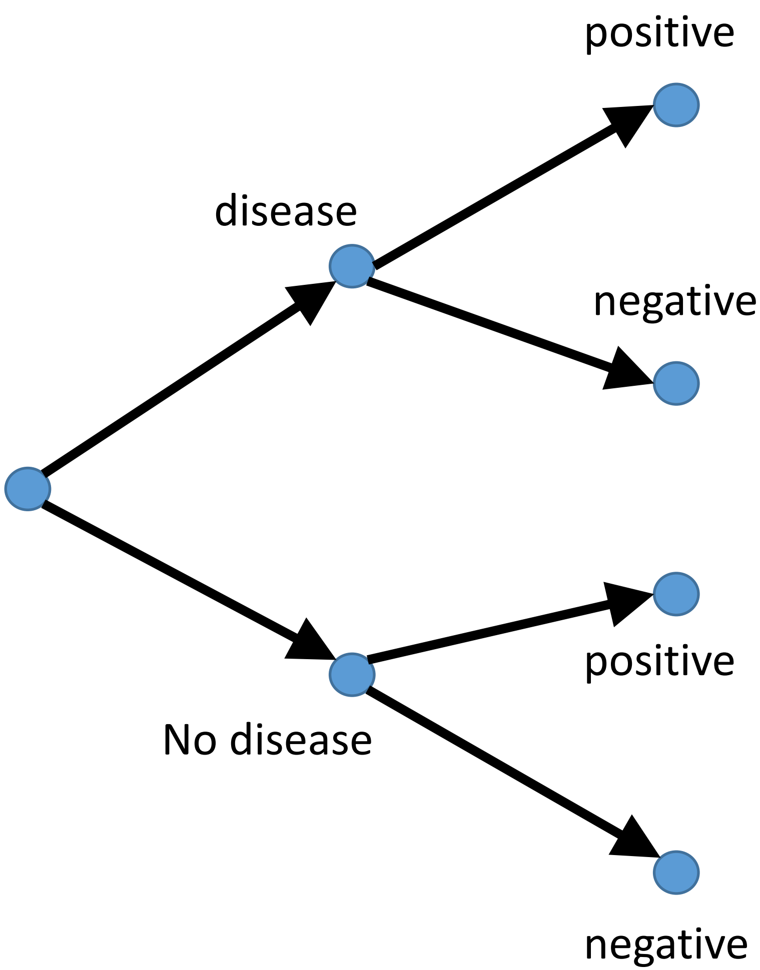
(disease, positive)

(disease, negative)

(no disease, positive)

(no disease, negative)

Tree Diagram Example



Outcome Space Ω

(disease, positive)

(disease, negative)

(no disease, positive)

(no disease, negative)

Example Event Spaces:

Test is positive or negative

Event Space E

$\emptyset$

{(disease, positive), (no disease, positive)}

{(disease, negative), (no disease, negative)}

Ω

Disease or no disease

Event Space E

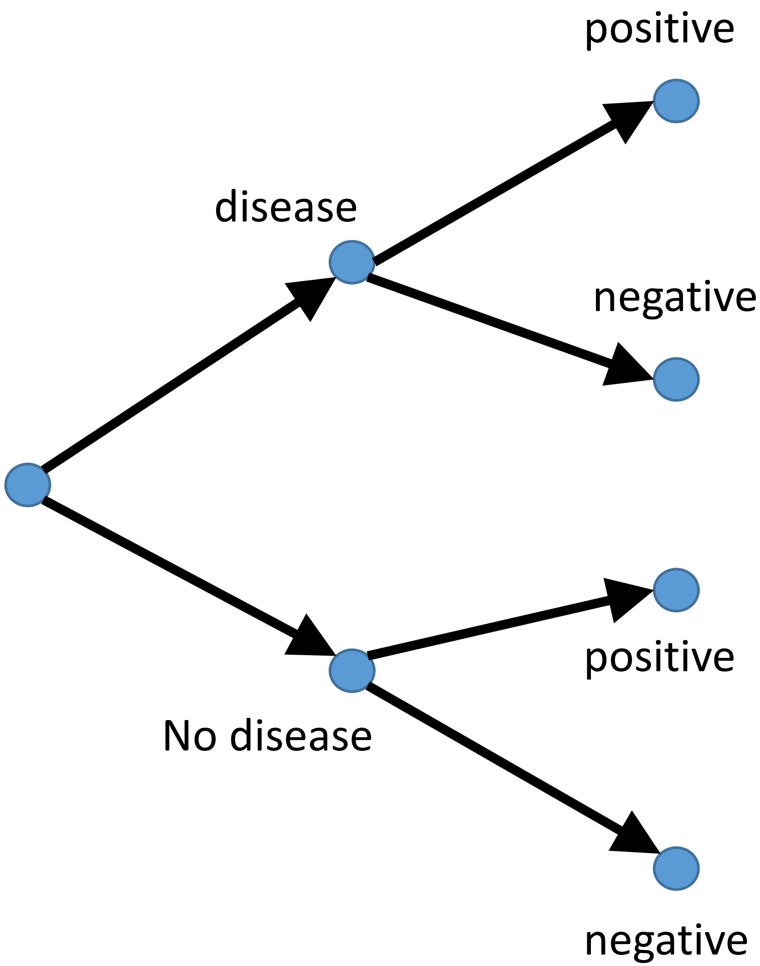
$\emptyset$

{(disease, positive), (disease, negative)}

{(no disease, positive), (no disease, negative)}

Ω

Tree Diagram Example



Outcome Space Ω

(disease, positive)

(disease, negative)

(no disease, positive)

(no disease, negative)

Example Event Spaces:

Event Space E

$\emptyset$

{(disease, positive)}

{(disease, negative)}

{(no disease, positive)}

{(no disease, negative)}

Ω

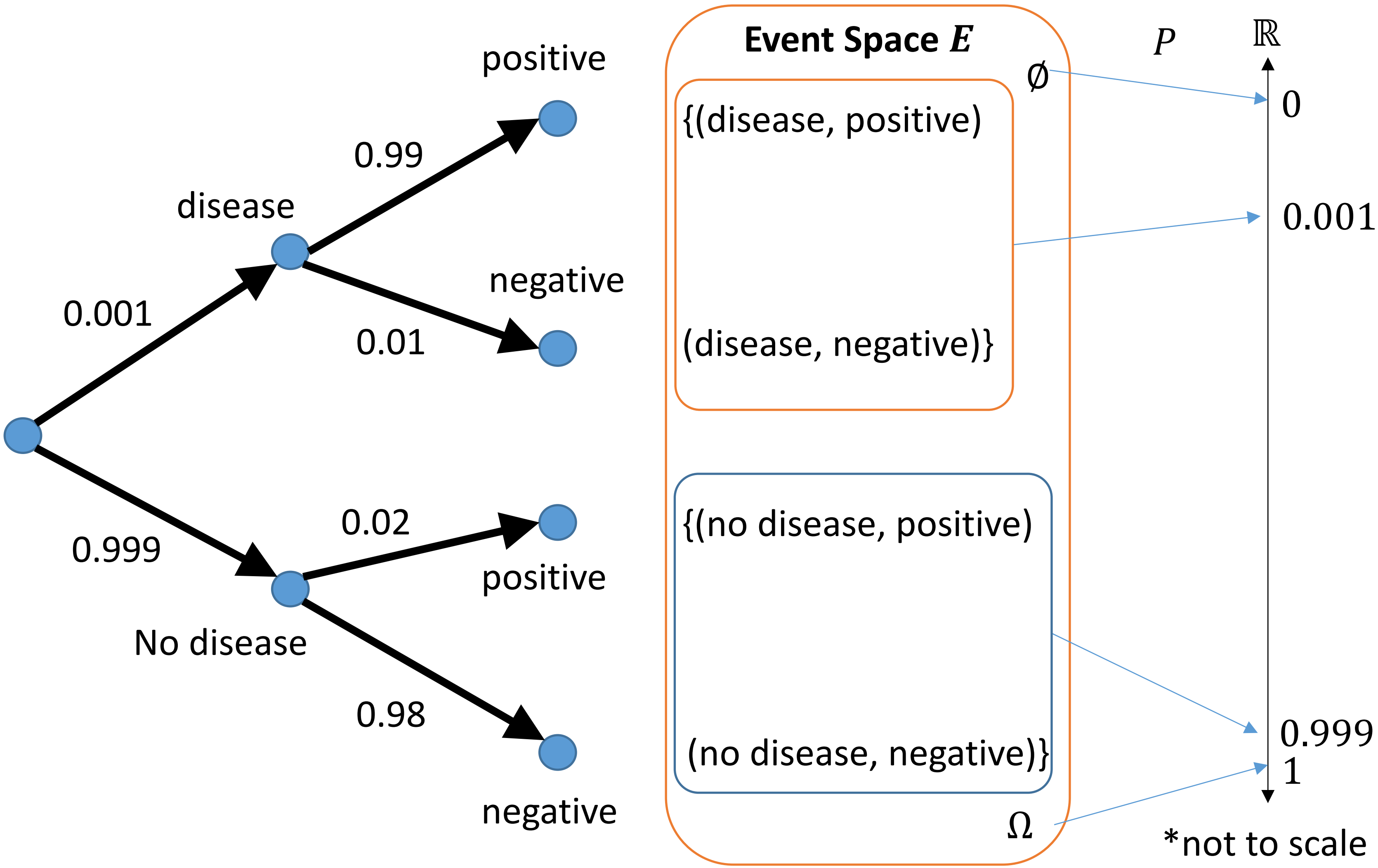
Question: This event space is incomplete. What other events are missing?

Probability Distributions

- A probability distribution P over (Ω, E) is a mapping from events in E to real values that satisfies the following conditions, i.e. axioms of probability:

1. Non-negativity, i.e. $P(\alpha) \geq 0,\ \forall\, \alpha \in E$.
2. Probability of all outcomes sums to 1, i.e. $P(\Omega) = 1$.
3. Mutually disjoint events: If $\alpha, \beta \in E$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$.

Tree Diagram Example



*not to scale

Random Variables

- A random variable, denoted as X (upper case), is the formal machinery for discussing attributes and their values in different outcomes.

- More formally, it is a function $X: \Omega \rightarrow S$ that maps a set of possible **outcomes** Ω to a space S
  - S usually a subset of $\mathbb{R}$, but can be other sets.

Random Variables: 2 Coin Flips

- Independent, Unbiased Coin Flips
- Possible outcomes $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$
- Random variable to indicate either both heads or both tails

$$X(\omega) = \begin{cases} 1 \text{ if HH or TT} \\ 0 \text{ otherwise} \end{cases}$$

Indicator Random Variables

- **Example:** 3 independent, unbiased coin flips.

$$\Omega = \{ \text{HHH}, \text{TTT}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH} \}$$

$$X(\omega) = \begin{cases} 1 \text{ if HHH or TTT} \\ 0 \text{ otherwise} \end{cases}$$

Exercise: Can you come up with a random variable which represents the *number* of heads? How does it relate to the event space?

Indicator Random Variables

- Indicator random variable maps every **outcome** to either 0 or 1.
- **For example:** whether you have the disease

$$\Omega = \{(d,\oplus),(\neg d,\oplus),(d,\ominus),(\neg d,\ominus)\}$$

$$X(\omega) = \begin{cases} 1 \text{ if } (d,\oplus) \\ 1 \text{ if } (d,\ominus) \\ 0 \text{ if } (\neg d,\oplus) \\ 0 \text{ if } (\neg d,\ominus) \end{cases}$$

Random Variables

- The set of values that a random variable X can take is denoted as $Val(X)$.

- A lower case letter, e.g. x, is a generic value of a random variable X, a.k.a. realization of the random variable.
  - E.g.: for an indicator random variable, $x = 1$ or $x = 0$

- The value of a random variable $Val(X)$ can be:
  - ➢Discrete, i.e. takes values from a predefined set, or
  - ➢Continuous, i.e. take values that are real numbers.

Random Variables

Examples:

Random variables with discrete values

- Rolling a six-faced die: $Val(X) = \{1,2, \ldots, 6\}$
- Weather conditions: $Val(X) = \{\text{"rain"}, \text{"cloud"}, \text{"snow"}, \text{"sun"}, \text{"wind"}\}$
- Number of people on the next train: $Val(X) = \mathbb{Z}_{\geq 0}$

Continuous random variables

- Time taken to finish an exam: $Val(X) = [1,2]$ hours
- Height of a tree: $Val(X) = \mathbb{R}_{>0}$
- Ambient Temperature: $Val(X) = \mathbb{R}$

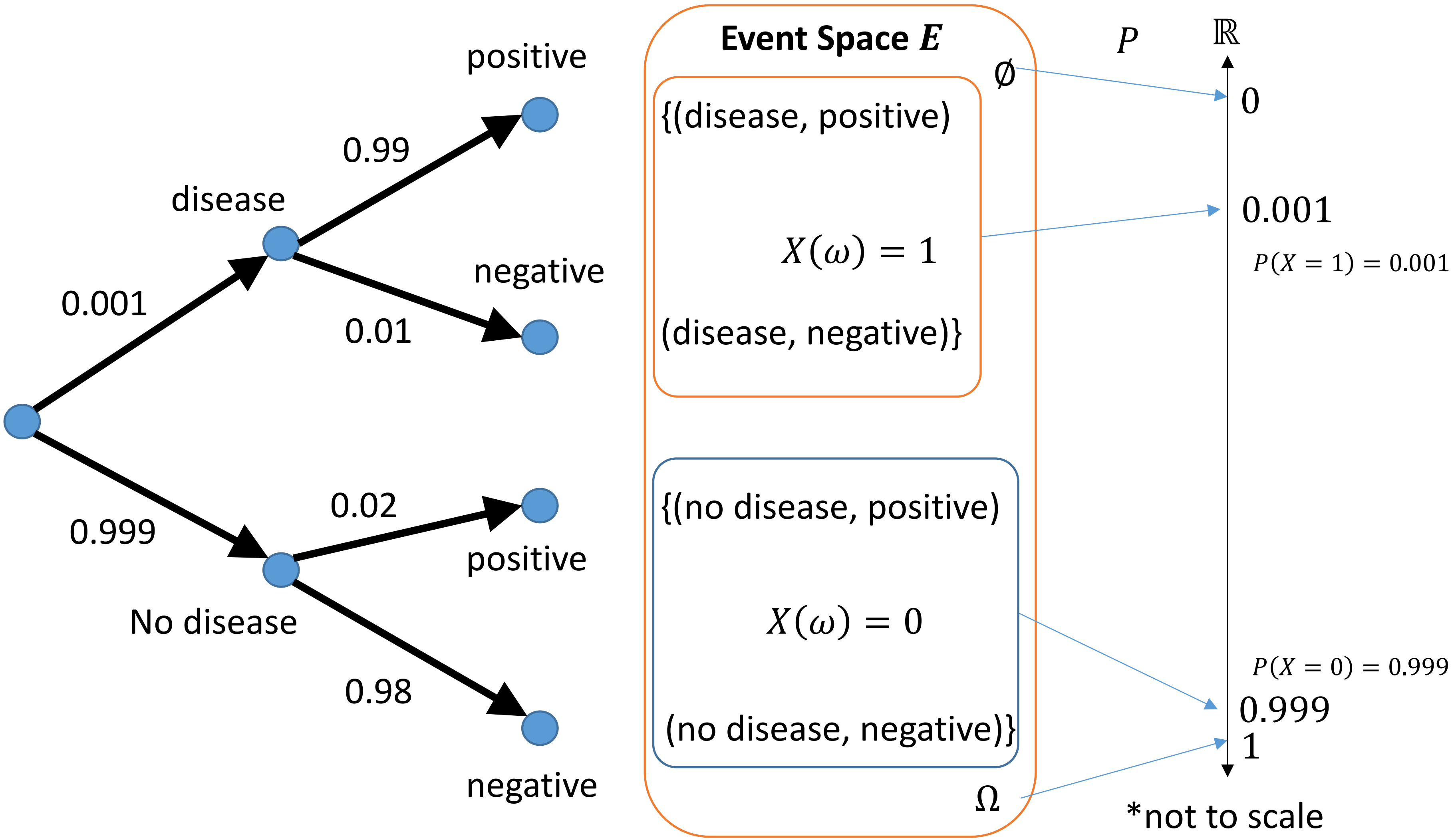
Probabilities & Random Variables

- Interested in the probability of random variables taking on certain values.
  - E.g: the probability of:
    - all heads/tails given 3 independent coin flips
    - the number new COVID-19 patients will be 0

- $P(x)$ is often used as a shorthand notation for $P(X = x)$.
  - Recall: x is a generic value of a random variable X

- We use the notation x^i to represent a specific value of X.

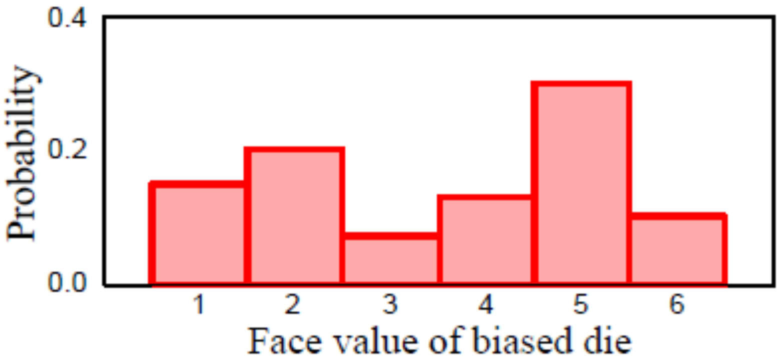
Tree Diagram Example



Probability Distributions: Discrete Vs Continuous

- Discrete: Probability mass function, $P(x)$

$P(x)$



$Val(X) = \{1,2,3,4,5,6\}$

$$\sum_{i=1}^{K} P\left(X = x^i\right) = 1$$

$$0 \leq P\left(X = x^i\right) \leq 1, \ \forall\, i = 1, \ldots K$$

$$K = |Val(X)|$$

Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability Distributions: Discrete Vs Continuous

- For continuous distributions, we have a problem:

$$P\left(X = x^i\right) = 0,\ \forall\ x^i \in Val(X)$$

circumference $= l$

Consider interval [a,b], what should be its probability?

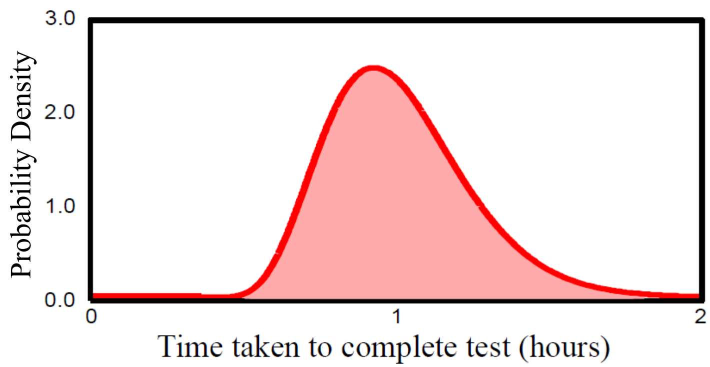
$$\mathrm{P}([a,b]) \propto length[a,b]$$
$$P([a,b]) = (b-a)/l$$

Probability Distributions: Discrete Vs Continuous

- Continuous: Probability density function is a function (denoted by a lower case p) $p(x): \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$.

$$\int_{Val(X)} p(x)dx = 1 \qquad p\left(X = x^i\right) \geq 0, \quad \forall \; x^i \in Val(X)$$

$p(x)$



$Val(X) = [0,2]$

$P(X)$ is the cumulative function of X:

$$P(X \leq a) = \int_{-\infty}^{a} p(x)\, dx$$

$$P(a \leq X \leq b) = \int_{a}^{b} p(x)\, dx$$

Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Back to our Wheel

- What is pdf $p(x)$ if we want uniform?
- Remember: the area under the curve must equal 1

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

$$\int_{-\infty}^{+\infty} p(x)\,dx = \int_{0}^{l} p(x)\,dx$$

$$= \int_{0}^{l} c\,dx$$

$$= cl = 1$$

So, we have $c = \frac{1}{l}$

circumference $= l$
What should the probability density function be?

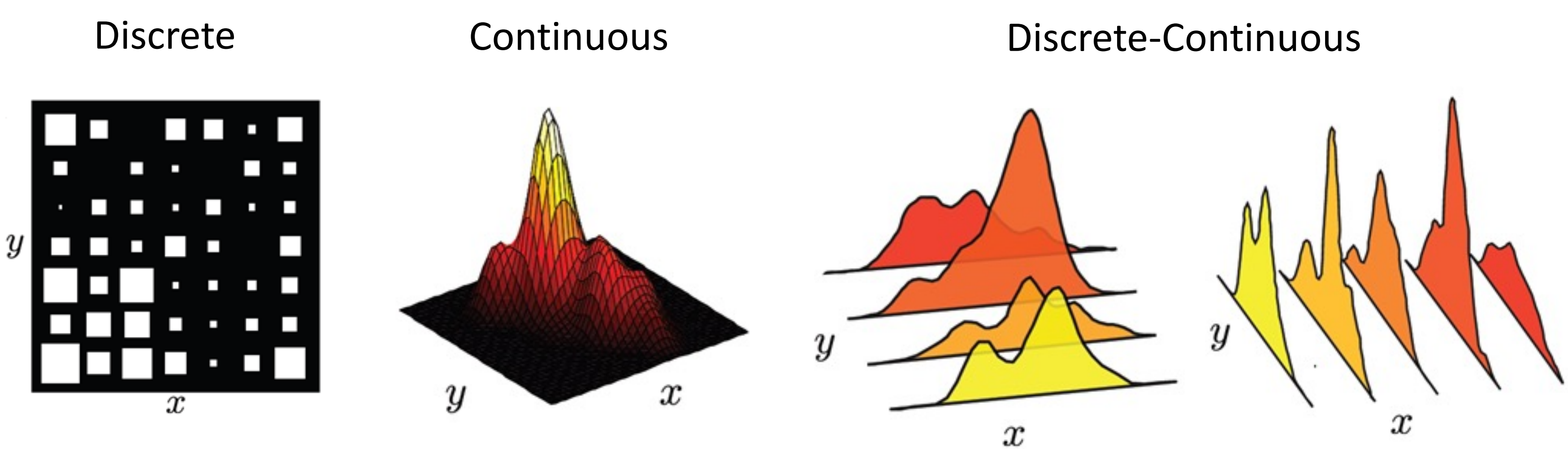
Probability Distributions: Discrete Vs Continuous

In this course, we abuse notation by denoting both the probability mass function and probability density function as the lower case $p(x)$

We silently note the property differences in $P(x)$ when X is discrete or continuous.

Probability: Joint Probability

- Consider all combination of events of two random variables X and Y.
- Some combinations of outcomes are more likely than others.



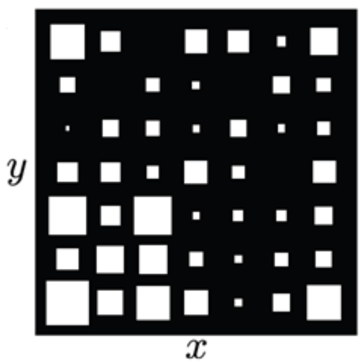
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Probability: Joint Probability

- This is captured in the joint probability distribution $p(x, y)$.
- Read as "probability of X and Y".
- Can be more than two random variables, i.e. $p(a, b, c, \ldots)$.

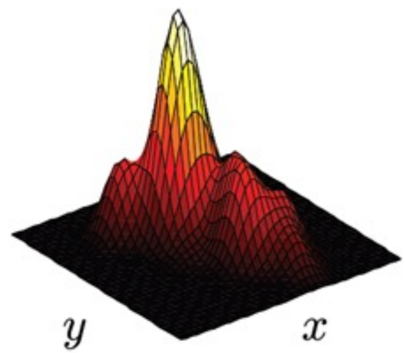
Discrete

$p(x, y)$



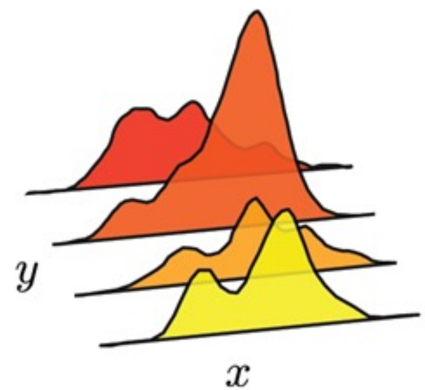
Continuous

$p(x, y)$

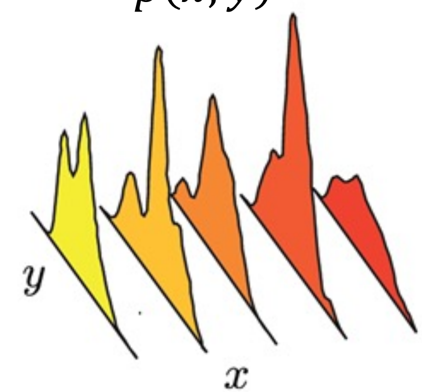


Discrete-Continuous

$p(x, y)$



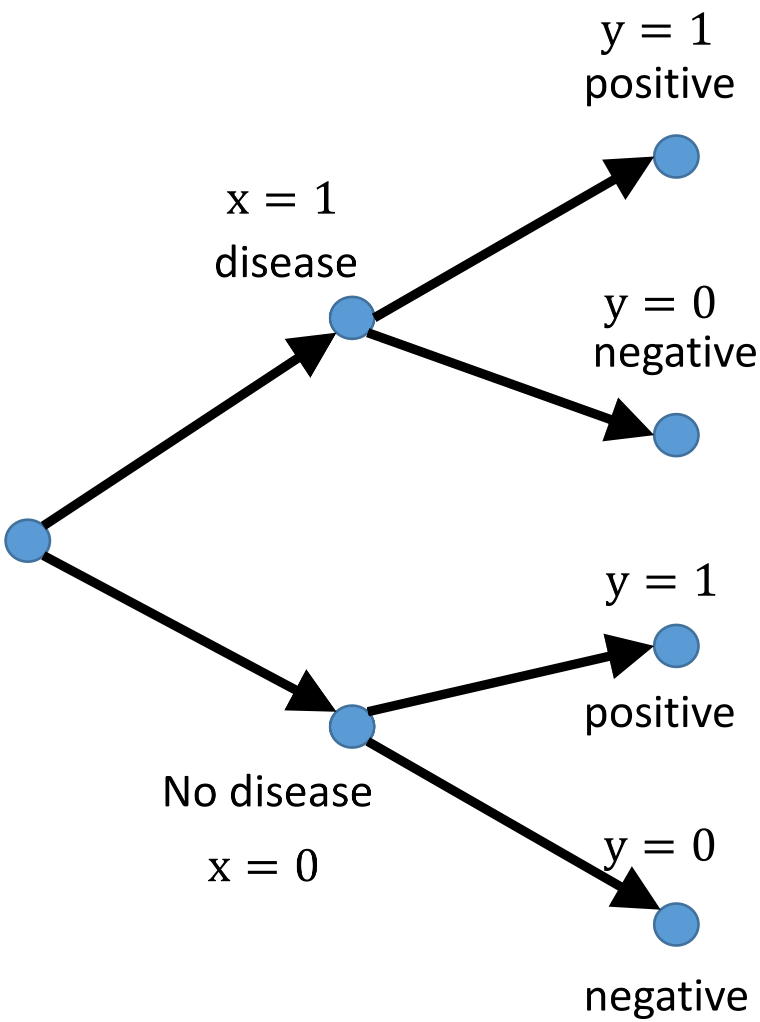
$p(x, y)$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

NUS National University of Singapore | School of Computing

Tree Diagram Example



We can now have:

$$p(x, y) = P(X = x \text{ and } Y = y)$$

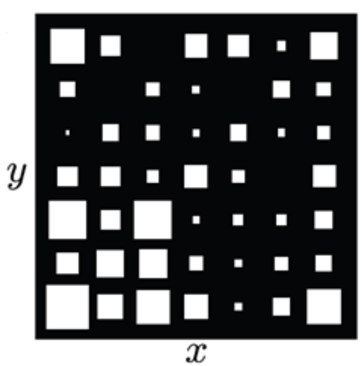
Basic Operations

Marginalization, Conditioning, Bayes Rule and Expectations

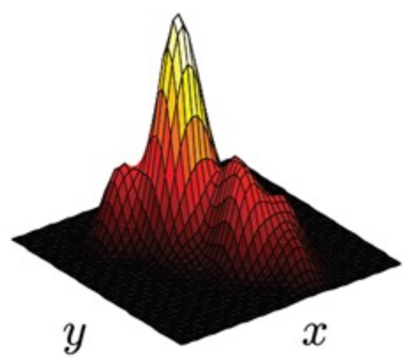
Probability: Joint Probability

- Consider all combination of events of two random variables X and Y.
- Some combinations of outcomes are more likely than others.

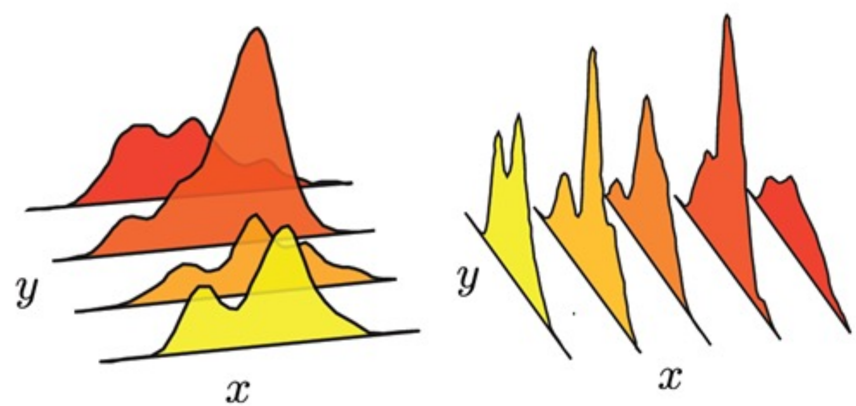
Discrete



Continuous



Discrete-Continuous



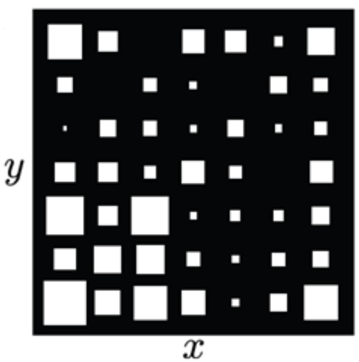
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Probability: Joint Probability

- This is captured in the joint probability distribution $p(x, y)$.
- Read as "probability of X and Y".
- Can be more than two random variables, i.e. $p(a, b, c, \ldots)$.

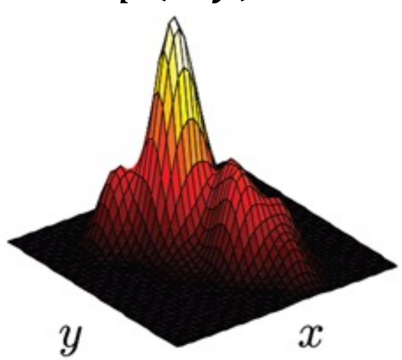
Discrete

$p(x, y)$



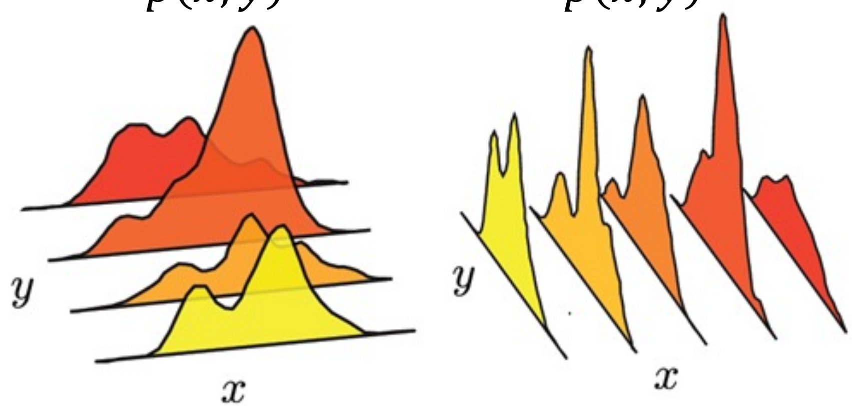
Continuous

$p(x, y)$



Discrete-Continuous

$p(x, y)$ $p(x, y)$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Summary: Sum and Product Rules

- Sum rule:

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

$$p(x) = \sum_{y} p(x,y)$$

- Product/Chain rule:

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

Probability: Marginalization

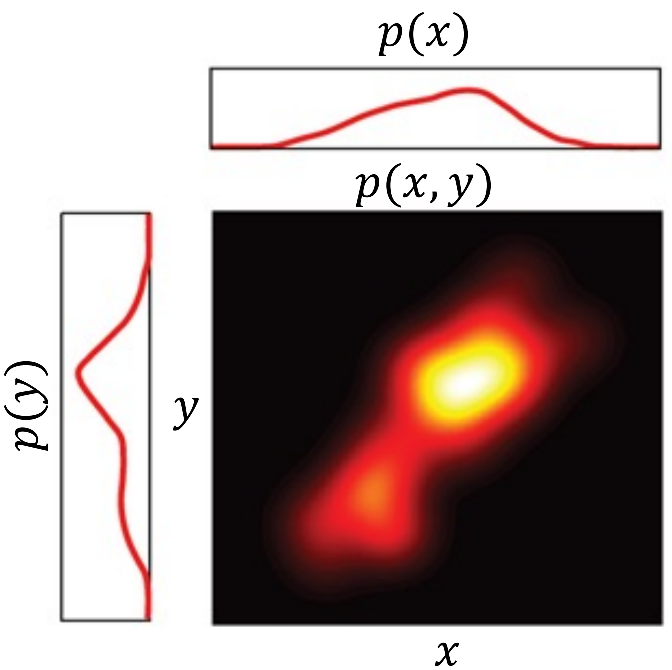
- Recover probability distribution of any variable in a joint distribution by integrating (or summing) over all other variables.

- Also known as the "sum rule" of probability.

Continuous:

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

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



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Marginalization

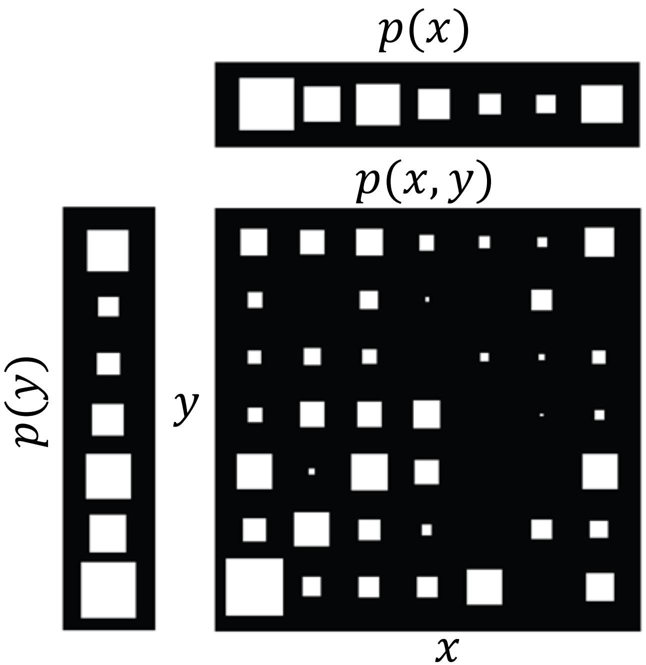
- Recover probability distribution of any variable in a joint distribution by integrating (or summing) over all other variables.

- Also known as the "sum rule" of probability.

Discrete:

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

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



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Marginalization

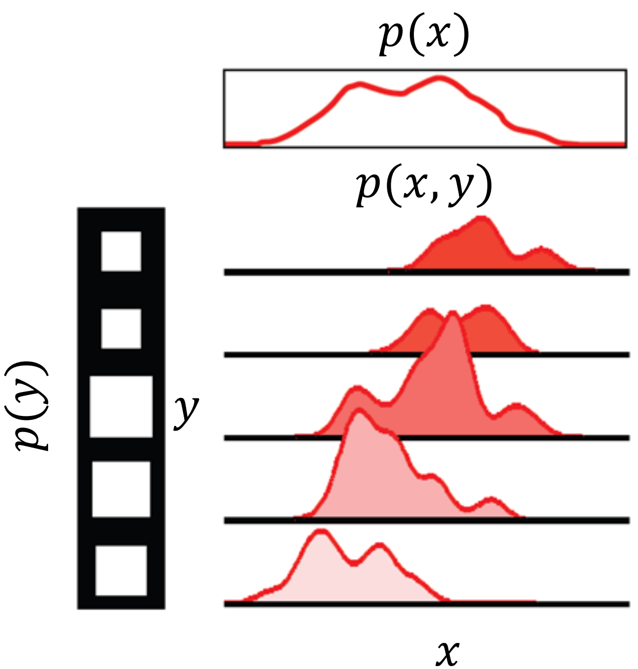
- Recover probability distribution of any variable in a joint distribution by integrating (or summing) over all other variables.

- Also known as the "sum rule" of probability.

Discrete-continuous:

$$p(x) = \sum_{y} p(x, y)$$

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



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Marginalization

- Works in higher dimensions too!

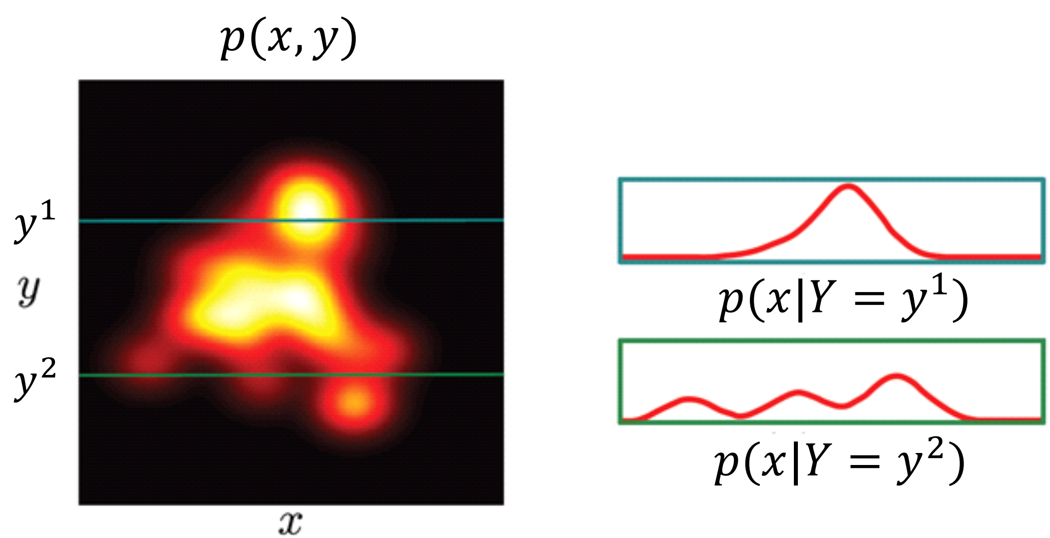
Example:

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

Probability: Conditional Probability

- $p(x|Y = y^*)$: "probability of X given $Y = y^*$".

- Relative propensity of the random variable X to take different outcomes given that the random variable Y is fixed to value y^*.



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Conditional Probability

- Conditional probability can be extracted from joint probability.
- Extract appropriate slice and normalize (so that the area is 1):

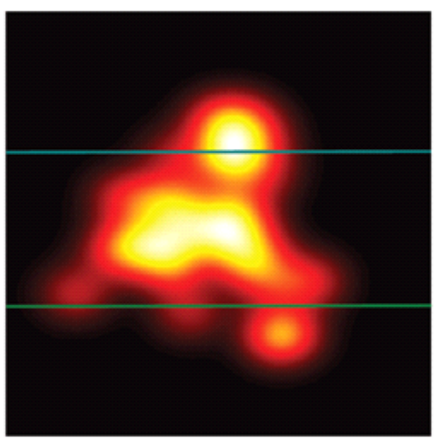
$$P(x|Y=y^*) = \frac{p(x,Y=y^*)}{\int p(x,Y=y^*)dx} = \frac{p(x,Y=y^*)}{p(Y=y^*)}$$

$p(x,y)$

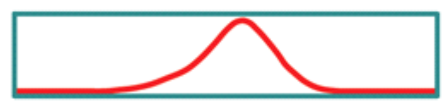
y^1

y

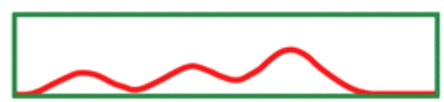
y^2



x



$p(x|Y=y^1)$



$p(x|Y=y^2)$

Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Conditional Probability

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

- Usually written in compact form:

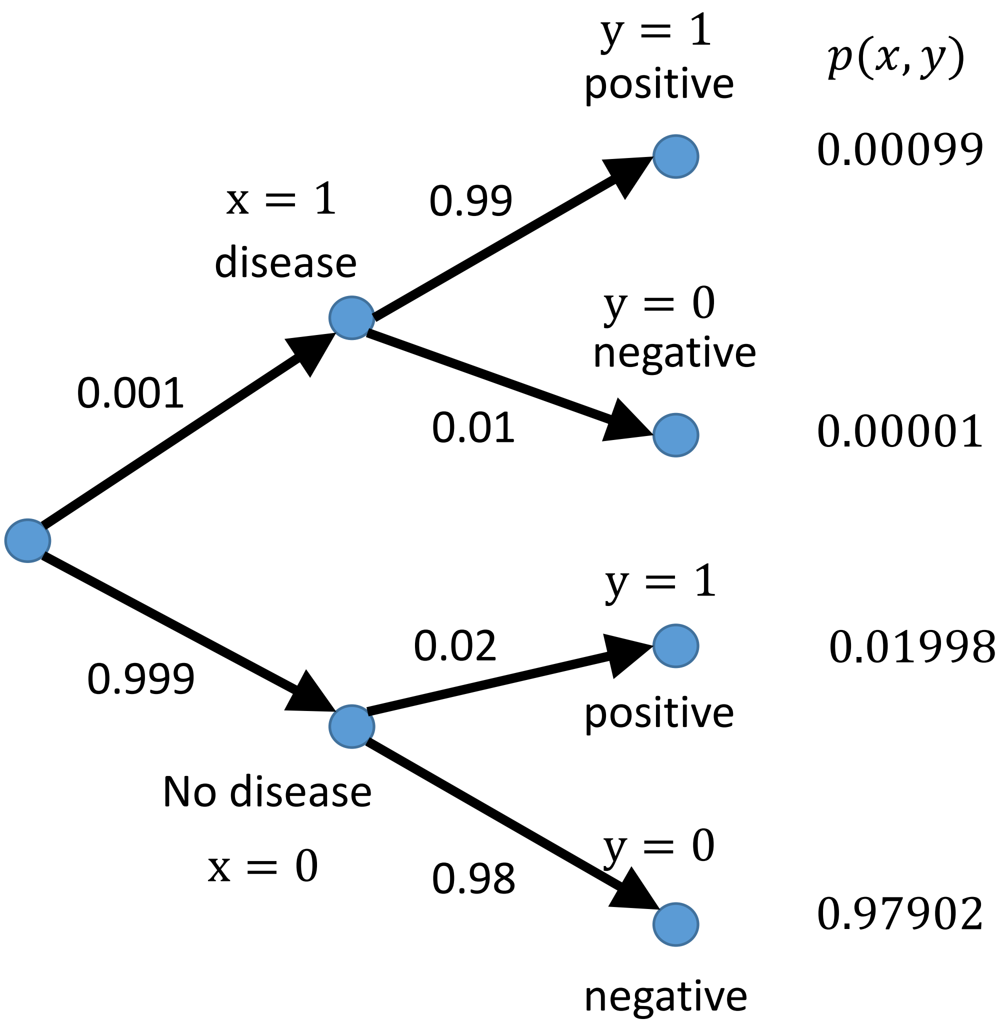
$$p(x|y) = \frac{p(x, y)}{p(y)}$$

- Which can be re-arranged to give:

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

known as "chain rule" or "product rule" of probability.

What is the probability I have the disease given the test is positive?



We want: $p(X = 1 \mid Y = 1)$

$$p(x|Y=1) = \frac{p(x, Y=1)}{p(Y=1)}$$
$$= \frac{p(x, Y=1)}{\sum_x p(x, Y=1)}$$

$$p(Y=1)$$
$$= 0.00099 + 0.01998$$
$$= 0.02097$$

$$p(X=1|Y=1)$$
$$= 0.00099/0.02097$$
$$= 0.0472 < 5\%$$

Probability: Conditional Probability

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

- Works for higher dimensions too!

Example:

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

Summary: Sum and Product Rules

- Sum rule:

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

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

- Product/Chain rule:

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

Probability: Bayes' Rule

- Formulated by Reverend Thomas Bayes in 1763 in "An Essay towards solving a Problem in the Doctrine of Chances"
- Further developed by Laplace and Jeffreys

Thomas Bayes
1701–1761

- *"[Bayes Theorem] is to the theory of probability what the Pythagorean theorem is to geometry"*

  – Sir Harold Jeffreys

Image source: "Pattern Recognition and Machine Learning", Christopher Bishop

Probability: Bayes' Rule

- Recall:

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

Thomas Bayes
1701–1761

- Eliminating $p(x, y)$, we get:

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

- Rearranging:

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

Image source: "Pattern Recognition and Machine Learning", Christopher Bishop

Probability: Bayes' Rule

Terminology:

Likelihood – propensity for observing a certain value of X given a certain value of Y

Prior – what we know about Y before seeing X

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

Posterior – what we know about Y after observing X

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

Problem: You're sick!

- You're not feeling well and go to the doctor.
- You take a blood test.
- Test comes back **positive** for *rare*, *fatal* disease.
  - Disease affects 0.1% of the population.
  - Test correctly identifies 99% of the people who have the disease.
  - If you do not have the disease, test may come back positive 2% of the time.

$$p(d|\oplus) = \frac{p(\oplus|d)p(d)}{p(\oplus)}$$

$$= \frac{p(\oplus|d)p(d)}{p(\oplus|d)p(d) + p(\oplus|\neg d)p(\neg d)}$$

Problem: You're sick!

- Test comes back **<u>positive</u>** for *rare*, *fatal* disease.
  - Disease affects 0.1% of the population.
  - Test correctly identifies 99% of the people who have the disease.
  - If you do not have the disease, test may come back positive 2% of the time.

$$p(d|\oplus) = \frac{p(\oplus|d)p(d)}{p(\oplus)}$$

$$= \frac{p(\oplus|d)p(d)}{p(\oplus|d)p(d) + p(\oplus|\neg d)p(\neg d)}$$

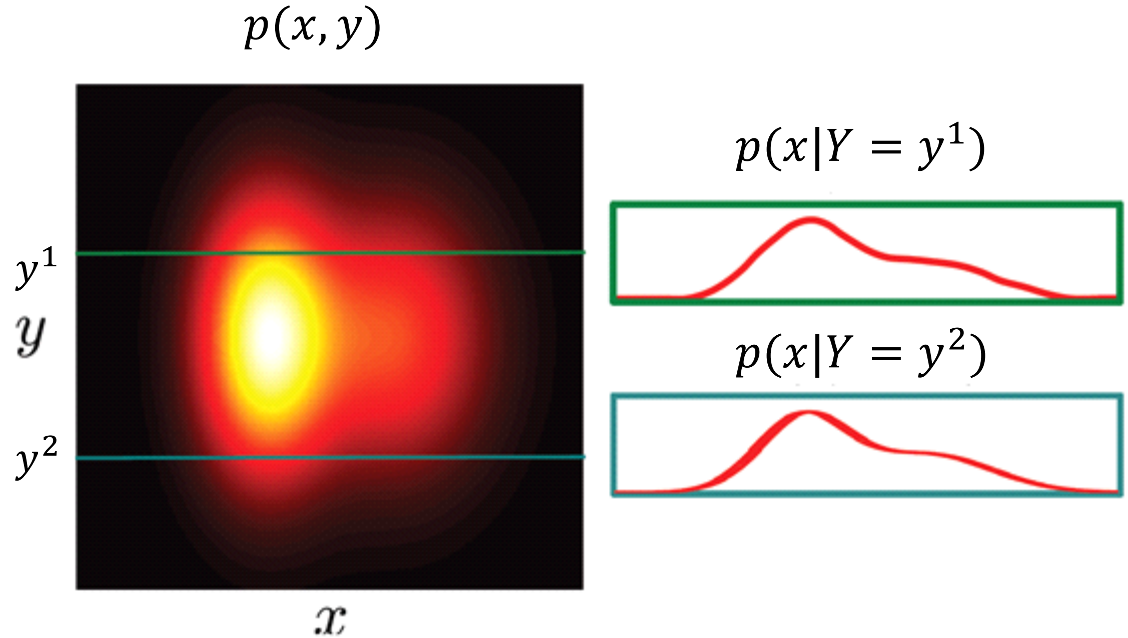
$$= \frac{0.99\times0.001}{0.99\times0.001 + 0.02\times0.999} = 0.047 < 5\%$$

Probability: Independence

- The independence of X and Y means that every conditional distribution is the same.
- The value of Y tells us nothing about X and vice-versa.

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

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



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

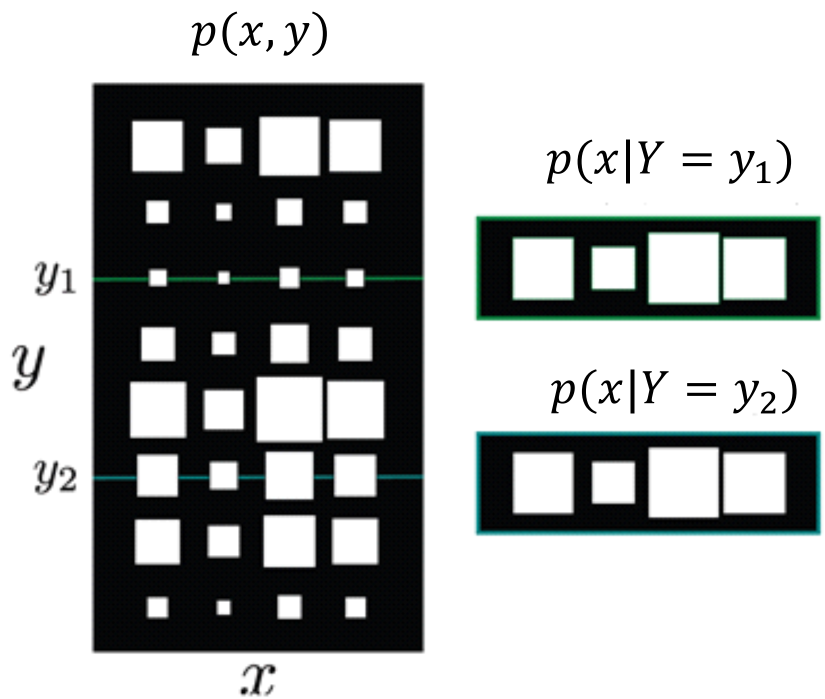
NUS National University of Singapore | School of Computing

Probability: Independence

- The independence of X and Y means that every conditional distribution is the same.
- The value of Y tells us nothing about X and vice-versa.

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

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



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Independence

- When variables are independent, the joint factorizes into a product of the marginals:

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

Probability: Expectation

- The expected or average value of some function $f[x]$ taking into account the distribution of X.

Definition:

$$E\big[f[x]\big] = \sum_x f[x]p(x)$$

$$E\big[f[x]\big] = \int f[x]p(x)dx$$

Probability: Rules of Expectation

- **Rule 1**: Expected value of a constant is the constant.

$$E[\kappa] = \kappa$$

- **Rule 2**: Expected value of constant times function is constant times expected value of function.

$$E\big[\kappa f[x]\big] = \kappa E[f[x]]$$

NUS National University of Singapore | School of Computing

Probability: Rules of Expectation

- **Rule 3**: Expectation of sum of functions is sum of expectation of functions.

$$E\big[f[x] + g[x]\big] = E\big[f[x]\big] + E\big[g[x]\big]$$

- **Rule 4**: Expectation of product of functions in variables X and Y is product of expectations of functions if X and Y are independent.

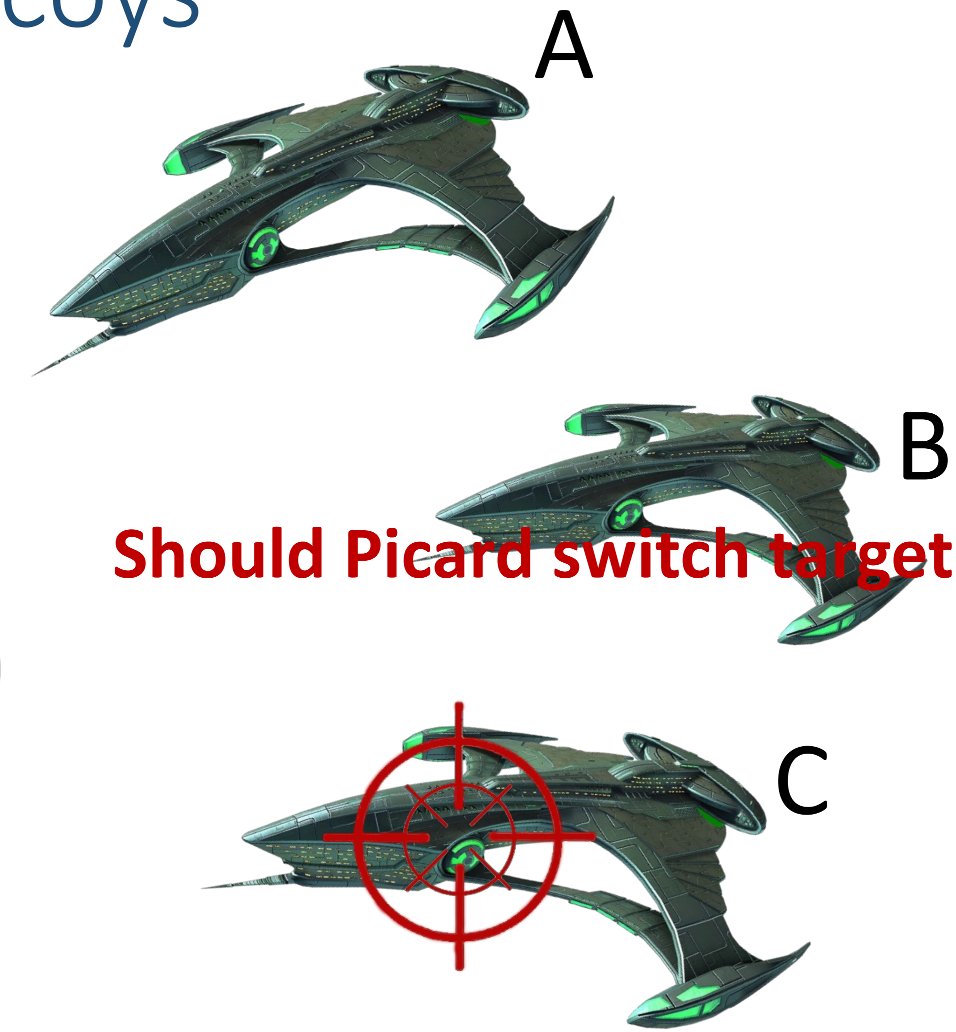
$$E\big[f[x]g[y]\big] = E\big[f[x]\big]E\big[g[y]\big],$$

if X and Y are independent

Probability Problems

Dastardly Decoys and Two Envelopes

Problem 1: Dastardly Decoys



Should Picard switch targets?

Problem 1: Solution

True Ship	Picard's Guess	Decoy Revealed	Outcome	Switch Wins	Probability
A 1/3	A 1/3	B 1/2	(A,A,B)		1/18
		C 1/2	(A,A,C)		1/18
	B 1/3	C 1	(A,B,C)	✓	1/9
	C 1/3	B 1	(A,C,B)	✓	1/9
B 1/3	A 1/3	C 1	(B,A,C)	✓	1/9
	B 1/3	A 1/2	(B,B,A)		1/18
		C 1/2	(B,B,C)		1/18
	C 1/3	A 1	(B,C,A)	✓	1/9
C 1/3	A 1/3	B 1	(C,A,B)	✓	1/9
	B 1/3	A 1	(C,B,A)	✓	1/9
	C 1/3	A 1/2	(C,C,A)		1/18
		B 1/2	(C,C,B)		1/18

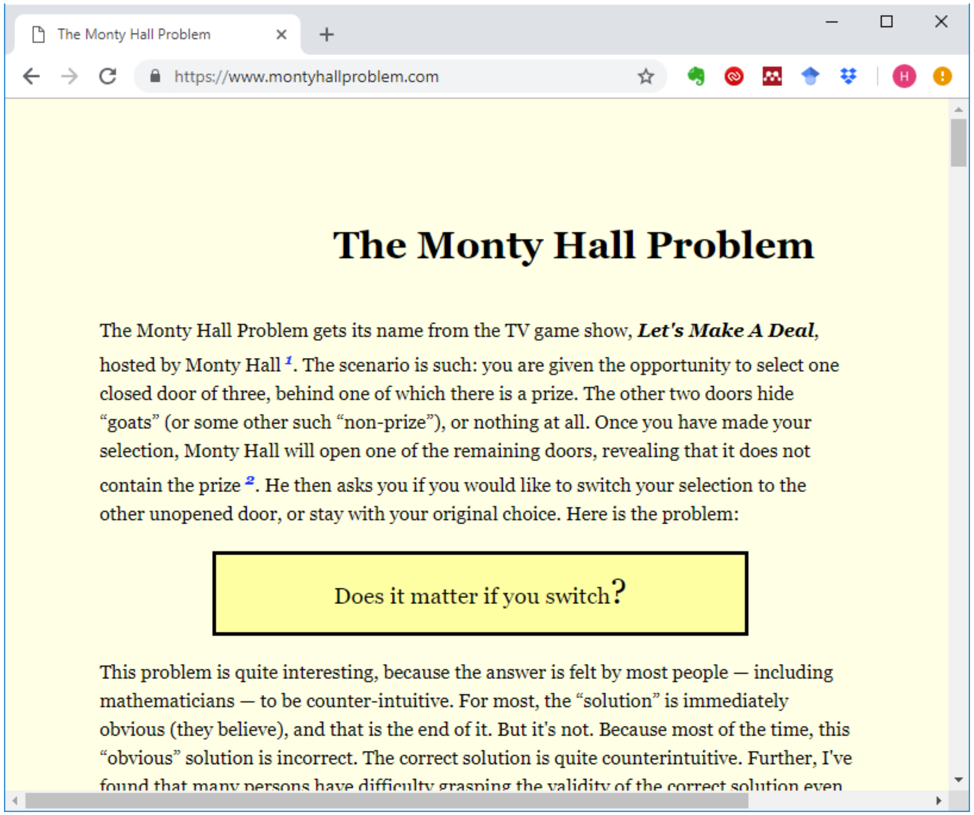
$$p(\text{switch wins}) = 6 \times \frac{1}{9} = \frac{2}{3}$$

Should switch!

Adapted from "Math for CS", Lehman, Leighton and Meyer


CHECKMATE
memegenerator.net

Monty-Hall Problem in Disguise



The Monty Hall Problem

https://www.montyhallproblem.com

The Monty Hall Problem

The Monty Hall Problem gets its name from the TV game show, ***Let's Make A Deal***, hosted by Monty Hall [1]. The scenario is such: you are given the opportunity to select one closed door of three, behind one of which there is a prize. The other two doors hide "goats" (or some other such "non-prize"), or nothing at all. Once you have made your selection, Monty Hall will open one of the remaining doors, revealing that it does not contain the prize [2]. He then asks you if you would like to switch your selection to the other unopened door, or stay with your original choice. Here is the problem:

Does it matter if you switch?

This problem is quite interesting, because the answer is felt by most people — including mathematicians — to be counter-intuitive. For most, the "solution" is immediately obvious (they believe), and that is the end of it. But it's not. Because most of the time, this "obvious" solution is incorrect. The correct solution is quite counterintuitive. Further, I've found that many persons have difficulty grasping the validity of the correct solution even

Monty Hall problem - Wikipedia

https://en.wikipedia.org/wiki/Monty_Hall_problem

Not logged in Talk Contributions Create account Log in

WIKIPEDIA
The Free Encyclopedia

Article Talk Read Edit View history Search Wikipedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Monty Hall problem

From Wikipedia, the free encyclopedia

The **Monty Hall problem** is a brain teaser, in the form of a probability puzzle, loosely based on the American television game show *Let's Make a Deal* and named after its original host, Monty Hall. The problem was originally posed (and solved) in a letter by Steve Selvin to the *American Statistician* in 1975 (Selvin 1975a), (Selvin 1975b). It became famous as a question from a reader's letter quoted in Marilyn vos Savant's "Ask Marilyn" column in *Parade* magazine in 1990 (vos Savant 1990a):

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

1 ? 2 ?

In search of a new car, the player picks a door, say *1*. The game host then opens one of the other doors, say *3*, to reveal a goat and offers to let the player switch from door *1* to door *2*.

Vos Savant's response was that the contestant should switch to the other door (vos Savant 1990a). Under the standard assumptions, contestants who switch have a $\frac{2}{3}$ chance of winning the car, while contestants who stick to their initial choice have only a $\frac{1}{3}$ chance.

2 Envelopes Game

- Team 1:
  - Pick 2 **different** numbers between 0 and 10.
  - Write each number on a piece of paper each.
  - Turn the papers face down.
- Team 2:
  - Objective is to pick the **larger number**.
  - Pick one of the pieces of paper.
  - Have a peek at the number.
  - **Decide:** do you keep this number or *switch*?
- **Question:** Can Team 2 win more than 50% of the time?

2 Envelopes Game

- Team 1:
  - Pick 2 **different** numbers between 0 and 10.
  - Write each number on a piece of paper each.
  - Turn the papers face down.
- Team 2:
  - Objective is to pick the **larger number**.
  - Pick one of the pieces of paper.
  - Have a peek at the number.
  - **Decide:** do you keep this number or *switch*?
- **Question:** Can Team 2 win more than 50% of the time? **Yes! But How and Why? Tutorial Next Week!**

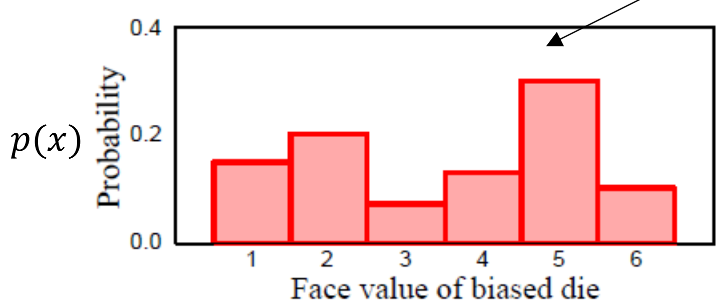
Probability Distributions

Basic distributions and Conjugacy

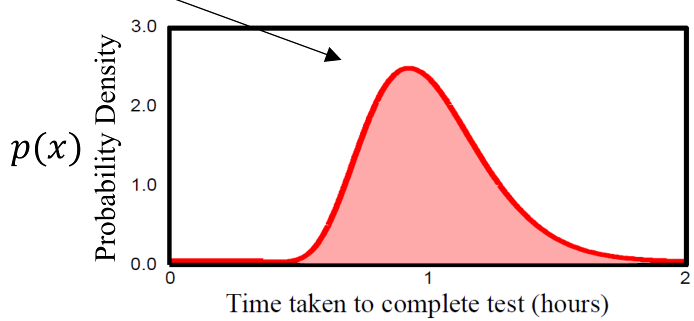
Probability Distributions

- We have seen the definitions of random variables, probability, and rules for manipulating probabilities.
- Question: "How do we assign the values of $p(X = x^i)$?"

How to get $p(X = x^i)$?



$val(X) = \{1,2,3,4,5,6\}$



$val(x) = [0,2]$

Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

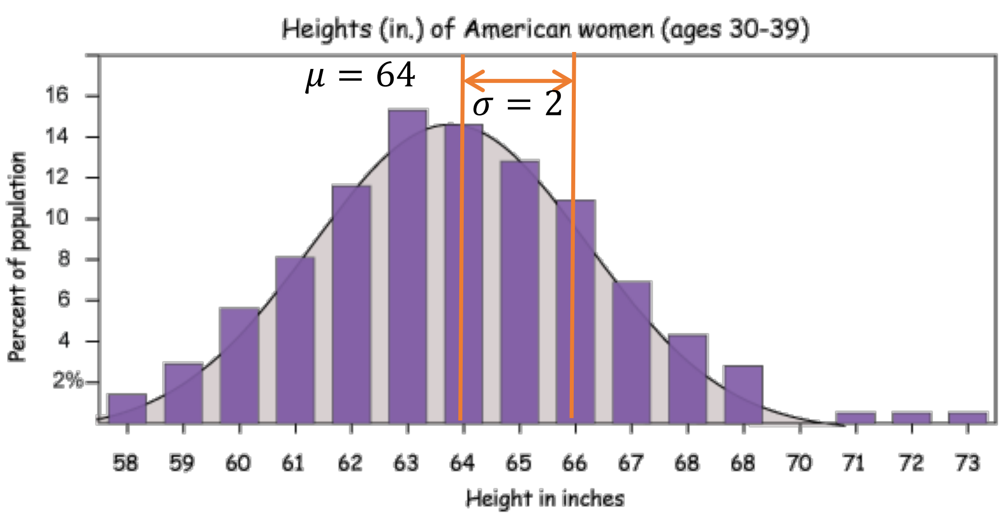
“Parametric” Probability Distributions

Q: “How do we assign the probability values?”

A: Use parametric probability distributions defined over some fixed set of parameters.

Example:

Fitting a Normal distribution to the heights of a population:



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

Parameters: mean $\mu = 64$, variance $\sigma^2 = 4$ are learned from data.

Image source: http://www.drcruzan.com/ProbStat_Distributions.html

Common Probability Distributions

- The choice of distribution depends on the type/domain of data to be modeled.

Data Type	**Domain**	**Distribution**
univariate, discrete, binary	$x \in \{0, 1\}$	Bernoulli
univariate, discrete, multi-valued	$x \in \{1, 2, \dots, K\}$	categorical
univariate, continuous, unbounded	$x \in \mathbb{R}$	univariate normal
univariate, continuous, bounded	$x \in [0, 1]$	beta
multivariate, continuous, unbounded	$\mathbf{x} \in \mathbb{R}^K$	multivariate normal
multivariate, continuous, bounded, sums to one	$\mathbf{x} = [x_1, x_2, \dots, x_K]^T$ $x_k \in [0, 1], \sum_{k=1}^K x_k = 1$	Dirichlet
bivariate, continuous, x_1 unbounded, x_2 bounded below	$\mathbf{x} = [x_1, x_2]$ $x_1 \in \mathbb{R}$ $x_2 \in \mathbb{R}^+$	normal-scaled inverse gamma
multivariate vector $\mathbf{x}$ and matrix $\mathbf{X}$, $\mathbf{x}$ unbounded, $\mathbf{X}$ square, positive definite	$\mathbf{x} \in \mathbb{R}^K$ $\mathbf{X} \in \mathbb{R}^{K \times K}$ $\mathbf{z}^T \mathbf{X} \mathbf{z} > 0 \quad \forall\, \mathbf{z} \in \mathbb{R}^K$	normal inverse Wishart

Problem: Infectious Agent

- Patient Zero is loose!
- When he comes into contact with someone, the chance he infects them is:
  $p(\text{Infected upon Contact}) = \lambda = 1/5$
- You know he came into contact with $n = 20$ people.
- You want to model the number of people Patient Zero infected.
- Which probability distribution applies to this scenario? Assume each contact is *independent*.
  - Common probability distributions and their applications
    - https://en.wikipedia.org/wiki/Probability_distribution#Common_probability_distributions_and_their_applications

Bernoulli Distribution

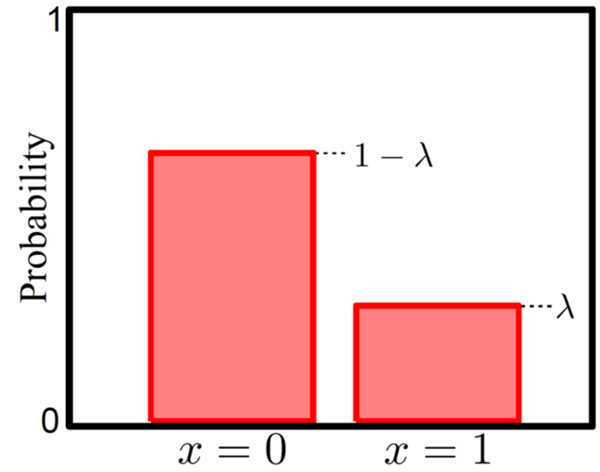
- Binary random variable X, i.e. $x \in \{0,1\}$
- A single parameter $\lambda \in [0,1]$.

$$\begin{aligned} p(X=0 \mid \lambda) &= 1-\lambda \\ p(X=1 \mid \lambda) &= \lambda \end{aligned}$$

Jacob Bernoulli
1654–1705

Or

$$\boxed{\begin{aligned} p(x) &= \lambda^x (1-\lambda)^{1-x}, \\ p(x) &= \mathrm{Bern}_x[\lambda] \end{aligned}}$$



Example:
X is the outcome of flipping a coin, $X=1$ represents 'heads', and $X=0$ represents 'tails'.

Images source: "Pattern Recognition and Machine Learning", Christopher Bishop
"Computer Vision: Models, Learning, and Inference", Simon Prince

NUS National University of Singapore | School of Computing

Binomial Distribution

- Discrete random variable X, i.e. $x \in \{0,1,2,\ldots,n\}$
- Two parameters $n \in \{0,1,2,\ldots\}, \lambda \in [0,1]$.

Jacob Bernoulli

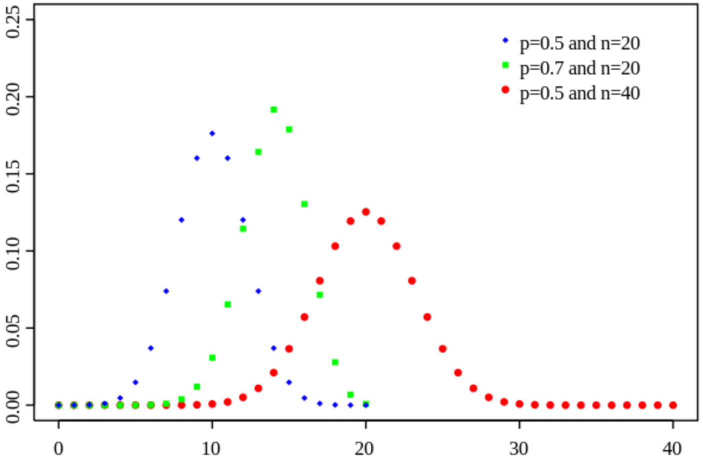
1654–1705

$$p(x) = \binom{n}{x}\lambda^x(1-\lambda)^{n-x},$$

$$p(x) = \mathrm{Bin}_x[n, \lambda]$$

Example:

X is the number of heads when flipping a coin 10 times.



Images source: “https://en.wikipedia.org/wiki/Binomial_distribution”, Wikipedia

NUS National University of Singapore | School of Computing

Categorical Distribution

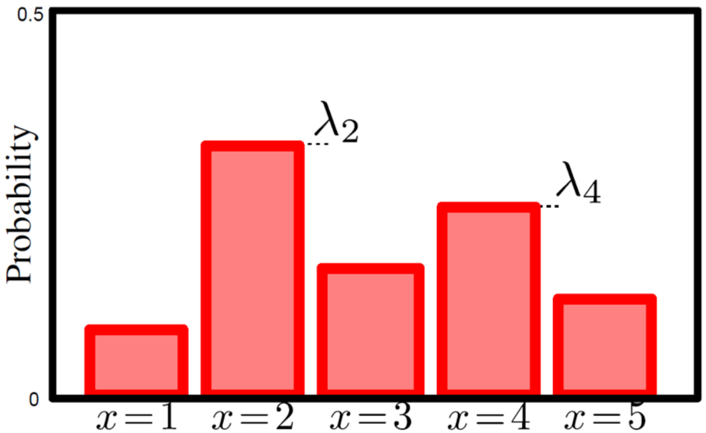
- Discrete variables $\boldsymbol{X}$ that take on 1-of-*K* possible mutually exclusive states, e.g. a *K*-faced die.
- $\boldsymbol{x}$ is represented by a *K*-dimensional vector $\mathbf{e}_k$ in which one of the elements $x_k = 1$, and $\sum_{k=1}^{K} x_k = 1$.
- e.g. $K = 5$, and $\boldsymbol{x} = \mathbf{e}_3 = [0,0,1,0,0]^T$.
- *K* parameters $\lambda = [\lambda_1, \dots, \lambda_K]^T$, where $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$.

$$p(\boldsymbol{X} = \mathbf{e}_k \mid \lambda) = \lambda_k$$

Or

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} \lambda_k^{x_k} = \lambda_k,$$

$$p(\boldsymbol{x}) = \mathrm{Cat}_x[\lambda]$$



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Univariate Normal Distribution

- Also known as the Gaussian distribution.
- Univariate normal distribution describes single continuous variable X, i.e. $x \in \mathbb{R}$.
- Two parameters $\mu \in \mathbb{R}$ (mean) and $\sigma^2 > 0$ (variance).

Carl Friedrich Gauss

1777–1855

$$p(X = a \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(a-\mu)^2}{2\sigma^2}, \quad a \in \mathbb{R}$$

Or

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

$$p(x) = \mathrm{Norm}_x[\mu, \sigma^2]$$

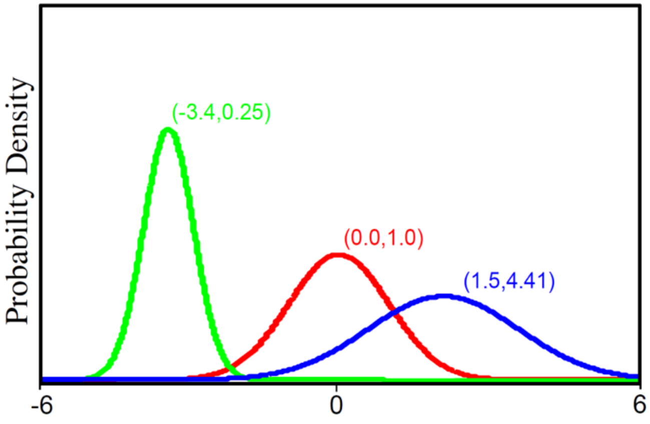


Image sources: "Pattern Recognition and Machine Learning", Christopher Bishop
"Computer Vision: Models, Learning, and Inference", Simon Prince

Multivariate Normal Distribution

- Multivariate normal distribution describes a *D*-dimensional continuous variable $\boldsymbol{X}$, i.e. $\boldsymbol{x} \in \mathbb{R}^D$.
- *D*-dimensional mean $\boldsymbol{\mu} \in \mathbb{R}^D$, and $D \times D$ symmetrical positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}_{+}^{D \times D}$.

$$p(\boldsymbol{X} = \boldsymbol{a} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{ -0.5(\boldsymbol{a} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{a} - \boldsymbol{\mu}) \}, \quad \boldsymbol{a} \in \mathbb{R}^D$$

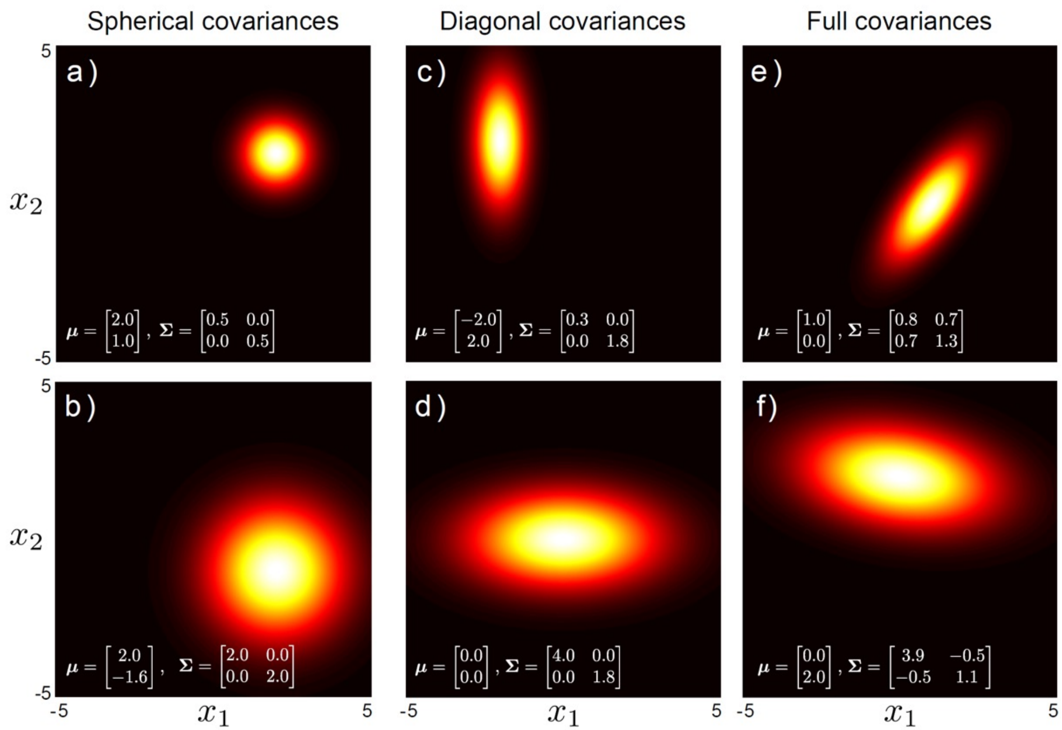
Or

$$\boxed{\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{ -0.5(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \} \\ p(\boldsymbol{x}) &= \mathrm{Norm}_{\boldsymbol{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}] \end{aligned}}$$

Types of Covariance

- Covariance matrix has three forms: spherical, diagonal and full.

$$\boldsymbol{\Sigma}_{spher} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \quad \boldsymbol{\Sigma}_{diag} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \boldsymbol{\Sigma}_{full} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$$



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Problem: Infectious Agent

- Patient Zero is loose!
- When he comes into contact with someone, the chance he infects them is:
  $p(\text{Infected upon Contact}) = \lambda = \ 1/5$

  Where did this number come from?
- You know he came into contact with $n = 20$ people.
- You want to model the number of people Patient Zero infected.
- Which probability distribution applies to this scenario? Assume each contact is *independent*.
  - Common probability distributions and their applications
    - https://en.wikipedia.org/wiki/Probability_distribution#Common_probability_distributions_and_their_applications

Learning with Bayes' Rule

Likelihood – propensity for observing the data given a certain parameter θ

Prior – what we know about θ before seeing D

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

Posterior – what we know about θ after observing the data D

Conjugate Distributions

- Conjugate distributions can be used to model the parameters of probability distributions.
- Product of a probability distribution and it's conjugate has the same form as the conjugate times a constant.
- Parameters of conjugate distributions are known as hyperparameters because they control the parameter distributions.
- List of conjugate distributions: https://en.wikipedia.org/wiki/Conjugate_prior

Distribution	**Domain**	**Parameters modeled by**
Bernoulli	$x \in \{0, 1\}$	beta
categorical	$x \in \{1, 2, \ldots, K\}$	Dirichlet
univariate normal	$x \in \mathbb{R}$	normal inverse gamma
multivariate normal	$\mathbf{x} \in \mathbb{R}^k$	normal inverse Wishart

Importance of Conjugate Distributions

Learning the parameters θ of a parametric probability distribution:

Recall the Bayes' Rule:

1. Choose prior that is conjugate to likelihood

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

2. The posterior will have same form as conjugate prior distribution, i.e. closed-form.

Problem: Infectious Agent

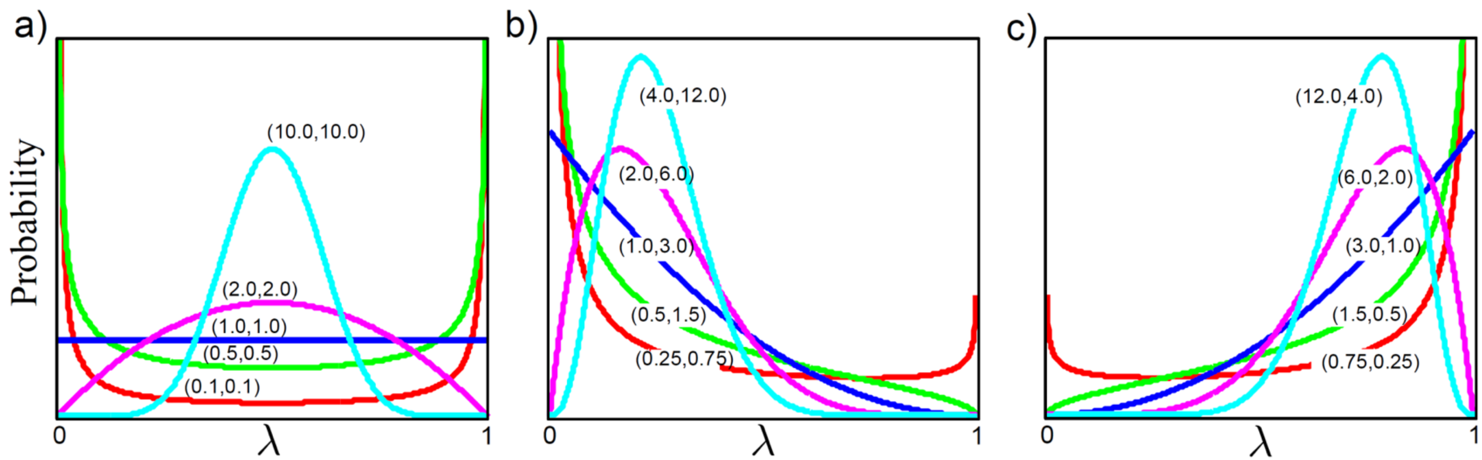
- Patient Zero is loose!
- You want to model the uncertainty over the binomial/Bernoulli parameter $\lambda = p(\text{Infected upon Contact})$
- What is the conjugate distribution for λ? Assume each contact is *independent*.
  - Conjugate Distributions: https://en.wikipedia.org/wiki/Conjugate_prior

Conjugate Distribution: Beta Distribution

- Conjugate distribution of Bernoulli distribution.
- Defined over parameter of the Bernoulli distribution $\lambda \in [0,1]$.

$$p(\lambda) = \frac{\Gamma[\alpha + \beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

$$p(\lambda) = \mathrm{Beta}_{\lambda}[\alpha, \beta]$$



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Conjugate Distribution: Beta Distribution

$$p(\lambda) = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}\lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

$$p(\lambda) = \mathrm{Beta}_{\lambda}[\alpha,\beta]$$

Gamma Function:

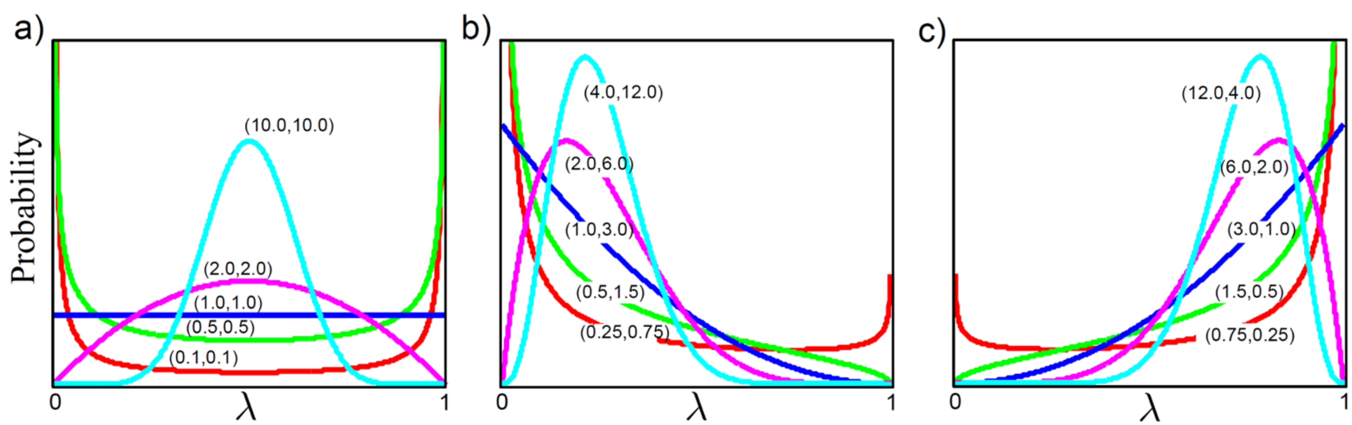
$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt, \qquad z \in \mathbb{C}$$

$$\Gamma(n) = (n-1)!\,, \qquad n \in \mathbb{Z}_{>0}$$

$$B(\alpha,\beta) = \frac{\Gamma[\alpha]\Gamma[\beta]}{\Gamma[\alpha+\beta]}$$

$$= \int_{t=0}^{1} t^{\alpha-1}(1-t)^{\beta-1}$$

- Two hyperparameters $\alpha, \beta > 0$.



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Tutorial: Beta-Binomial

- Show that the Beta distribution is conjugate to the Binomial distribution.

Bernoulli

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

$$p(x) = \mathrm{Bern}_x[\lambda]$$

Binomial

$$p(x) = \binom{n}{x} \lambda^x (1-\lambda)^{n-x}\,,$$

$$p(x) = \mathrm{Bin}_x[n, \lambda]$$

Beta

$$p(\lambda) = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)} \lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

$$p(\lambda) = \mathrm{Beta}_\lambda[\alpha, \beta]$$

Also:

$$B(\alpha,\beta) = \frac{\Gamma[\alpha]\Gamma[\beta]}{\Gamma[\alpha+\beta]}$$

$$= \int_{t=0}^{1} t^{\alpha-1}(1-t)^{\beta-1}$$

Learning Outcomes

Students should be able to:

1. Describe uncertain quantities with random variables and joint probabilities.

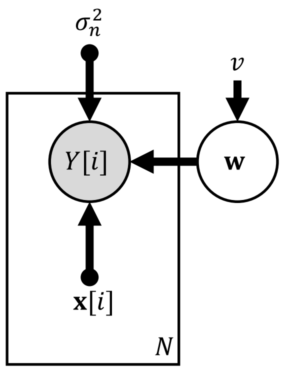
2. Explain the basic rules of probability – sum, product, Bayes', independence and expectation rules.

3. Use the common probabilities distributions – Bernoulli, categorical, univariate and multivariate normal distributions.

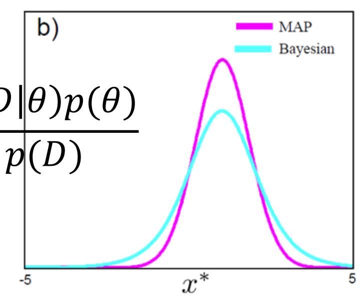
4. Explain the use of conjugate distributions.

CS5340 in a nutshell

CS5340 is about how to “**represent**” and “**reason**” with uncertainty in a computer.



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



Representation: The *language* is probability and probabilistic graphical models (PGM).

The language is used to model problems.

Reasoning: We use learning and inference algorithms to answer questions.

e.g., Belief-propagation/sum-product, MCMC, and variational Bayes

Appendix

More Conjugate Distributions

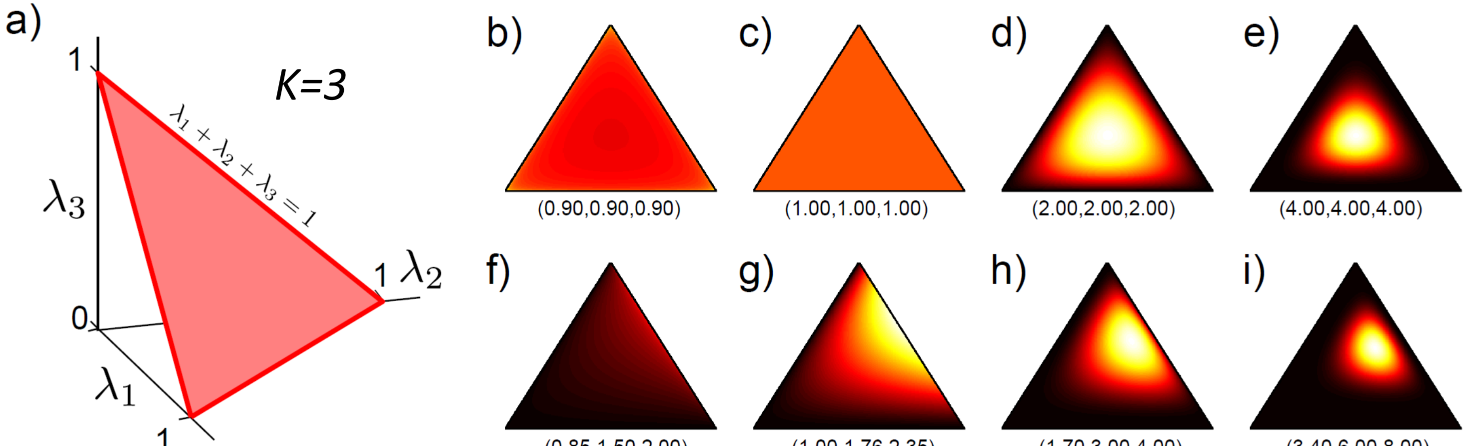
Dirichlet Distribution

- Conjugate distribution of categorical distribution.
- Defined over K parameters of Categorical distribution, $\lambda_k \in [0,1]$, where $\sum_k \lambda_k = 1$.

$$p(\lambda_1, \ldots, \lambda_K) = \frac{\Gamma[\sum_{k=1}^{K} \alpha_k]}{\prod_{k=1}^{K} \Gamma[\alpha_k]} \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1},$$

$$p(\lambda_1, \ldots, \lambda_K) = \mathrm{Dir}_{\lambda_{1\ldots K}}[\alpha_1, \ldots \alpha_K]$$

Peter Gustav Lejeune Dirichlet (1805-1859)



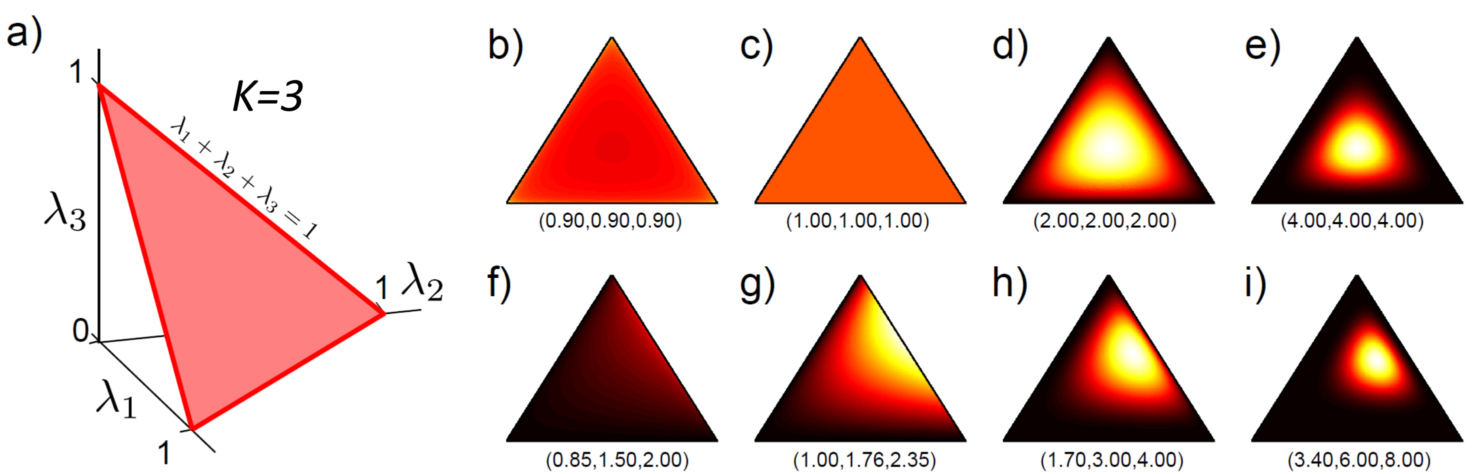
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince
http://www.amt.edu.au/biogdirichlet.html

Dirichlet Distribution

$$p(\lambda_1, \dots, \lambda_K) = \frac{\Gamma[\sum_{k=1}^{K} \alpha_k]}{\prod_{k=1}^{K} \Gamma[\alpha_k]} \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1},$$

$$p(\lambda_1, \dots, \lambda_K) = \mathrm{Dir}_{\lambda_{1\dots K}}[\alpha_1, \dots \alpha_K]$$

- *K* hyperparameters $\alpha_k > 0$.



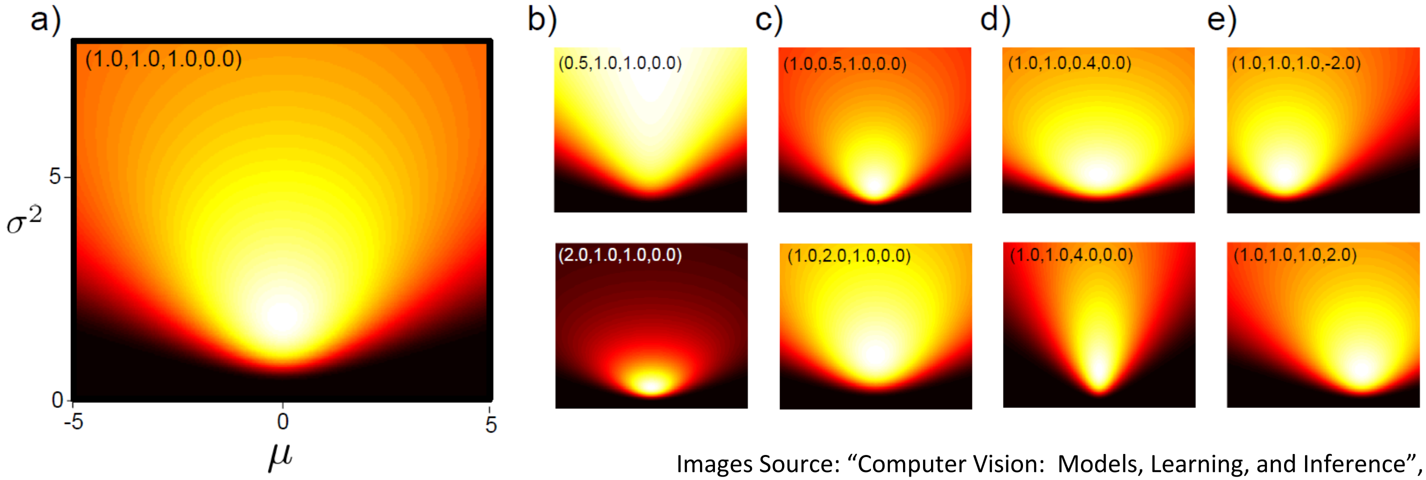
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Normal Inverse Gamma Distribution

- Conjugate distribution of univariate normal distribution.
- Defined on parameters $\mu, \sigma^2 > 0$ of univariate normal distribution.

$$p(\mu,\sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma[\alpha]}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left[-\frac{2\beta+\gamma(\delta-\mu)^2}{2\sigma^2}\right]$$

$$p(\mu,\sigma^2) = \text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]$$



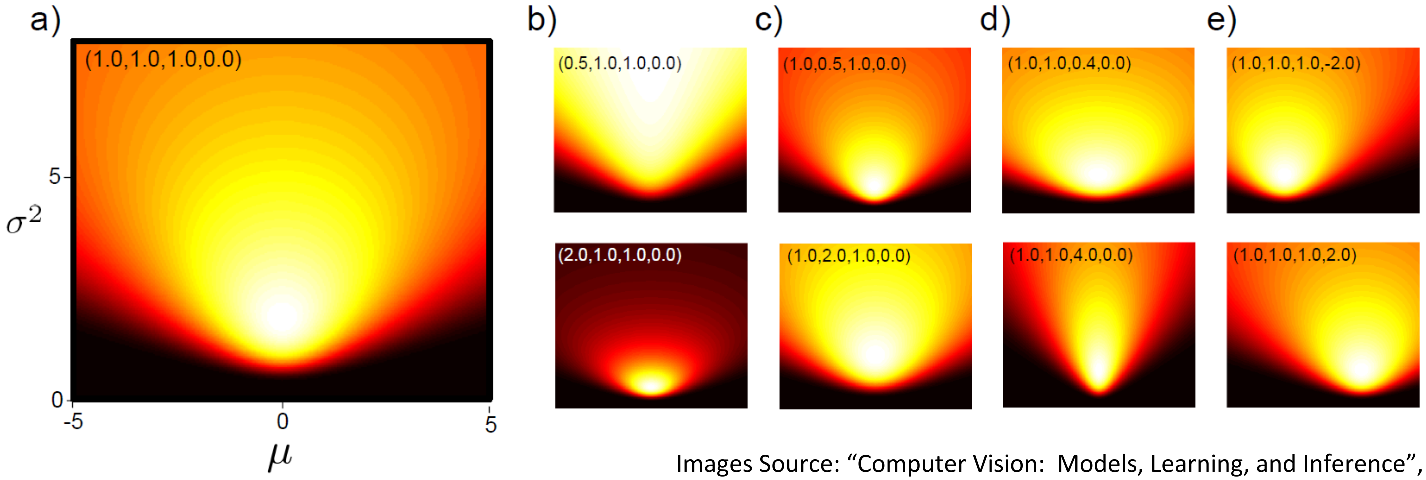
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Normal Inverse Gamma Distribution

$$p(\mu,\sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma[\alpha]}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left[-\frac{2\beta+\gamma(\delta-\mu)^2}{2\sigma^2}\right]$$

$$p(\mu,\sigma^2) = \text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]$$

- Four hyperparameters $\alpha, \beta, \gamma > 0$ and $\delta \in \mathbb{R}$.



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Normal Inverse Wishart

John Wishart
(1898-1956)

- Conjugate distribution of multivariate normal distribution.
- Defined on parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ of multivariate normal distribution.

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\gamma^{D/2}|\boldsymbol{\Psi}|^{\alpha/2}\exp[-0.5\left(\mathrm{Tr}\left[\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right] + \gamma(\boldsymbol{\mu}-\boldsymbol{\delta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\delta})\right)]}{2^{\alpha D/2}(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{(\alpha+D+2)/2}\Gamma_D[\alpha/2]}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{NorIWis}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}[\alpha, \boldsymbol{\Psi}, \gamma, \boldsymbol{\delta}]$$

- Four hyperparameters: a positive scalar α, a positive definite matrix $\boldsymbol{\Psi} \in \mathbb{R}_+^{D\times D}$, a positive scalar γ, and a vector $\boldsymbol{\delta} \in \mathbb{R}^D$.

Multivariate gamma function:

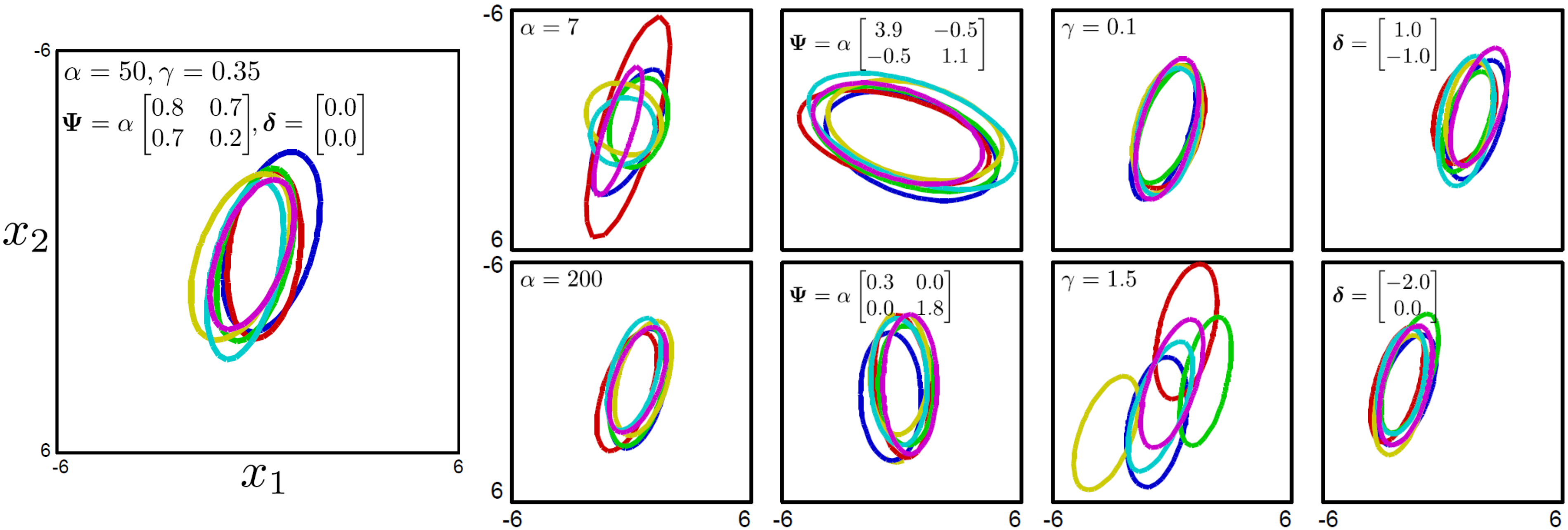
$$\Gamma_D[a] = \pi^{a(a-1)/4}\prod_{j=1}^{a}\Gamma[a+(1-j)/2]$$

NUS National University of Singapore | School of Computing

Normal Inverse Wishart

- Samples from Normal Inverse Wishart:

Modifies: Dispersion of covariances | Average covariance | Dispersion of mean vectors | Average value of mean vectors



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince