NUS National University of Singapore | School of Computing

CS5340
Uncertainty Modeling in AI

Asst. Prof. Lee Gim Hee

AY 2018/19

Semester 1

Course Information

Lecturer:

Dr. Lee Gim Hee
Department of Computer Science
Office: COM2-03-54
Email: gimhee.lee@comp.nus.edu.sg

Class:

Time: Every Wednesday, 1830hrs – 2130hrs
Venue: LT18

Mode of Assessment:

40% CA (coding assignment, max 2 students) **Due: 16 Nov, 2359 hrs**
60% Final Exam (one A4 cheat sheet is allowed) **05 December, Afternoon**

Teaching Assistants

Xie Yaqi
Department of Computer Science
Email: e0205023@u.nus.edu
Lab: COM1-01-09

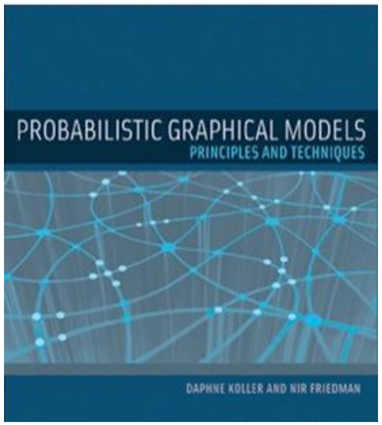
Zhao Na
Department of Computer Science
Email: e0147044@u.nus.edu
Lab: AS6-05-02

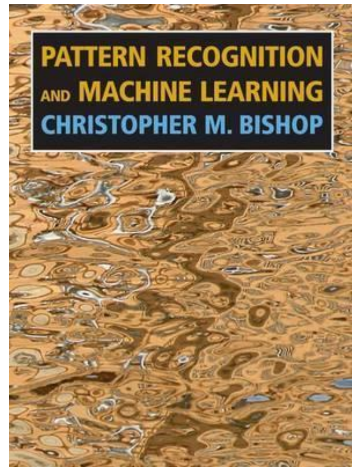
Course Schedule

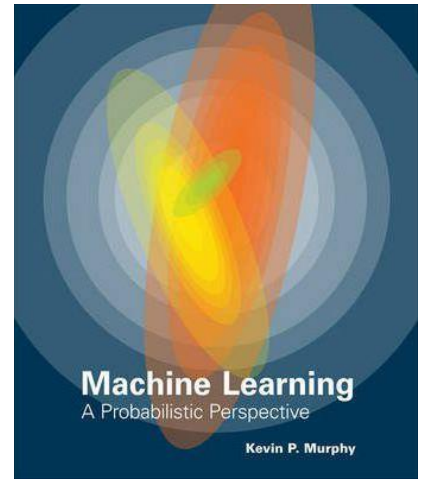
Week	Date	Topic	Remarks
1	15 Aug	Introduction to probabilities and probability distributions	
2	22 Aug	Fitting probability models	Hari Raya Haji*
3	29 Aug	Bayesian networks (Directed graphical models)	
4	05 Sep	Markov random Fields (Undirected graphical models)	
5	12 Sep	I will be traveling	No Lecture
6	19 Sep	Variable elimination and belief propagation	
-	26 Sep	Recess week	No lecture
7	03 Oct	Factor graph and the junction tree algorithm	
8	10 Oct	Parameter learning with complete data	
9	17 Oct	Mixture models and the EM algorithm	
10	24 Oct	Hidden Markov Models (HMM)	
11	31 Oct	Monte Carlo inference (Sampling)	
12	07 Nov	Variational inference	
13	14 Nov	Graph-cut and alpha expansion	

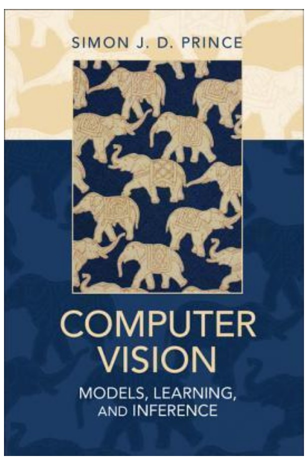
* **Make-up lecture**: 25 Aug (Sat), 9.30am-12.30pm, **LT 15**

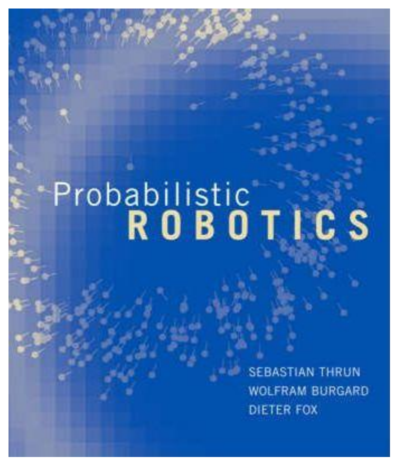
Recommended Readings (Not Compulsory)

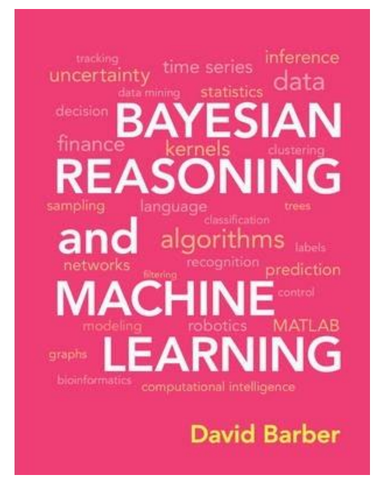












Probabilistic Graphical Modeling

One of the most exciting advances in machine learning (AI, signal processing, coding, control, robotics, computer vision . . .) in the last decades.

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

Probabilistic Graphical Modeling

How can we gain global insight based on local observations?

Adapted from: “Probabilistic Graphical Modeling” Lectures NYU, David Sontag

Probabilistic Graphical Modeling

Key Ideas:

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

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

Reasoning Under Uncertainty

- As humans, we are continuously making predictions under uncertainty.
- Classical AI and ML research ignored this phenomena.
- Many of the most recent advances in technology are possible because of this probabilistic approach.

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

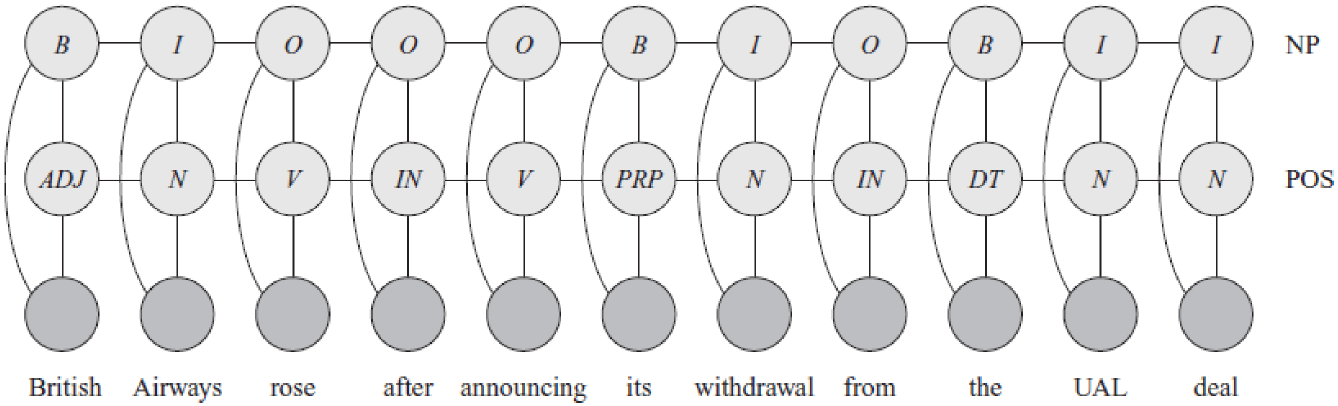
PGM: Applications

Markov Localization

" Monte Carlo Localization for Mobile Robots", Frank Dellaert et. al., ICRA 1999

PGM: Applications

Part of Speech Tagging



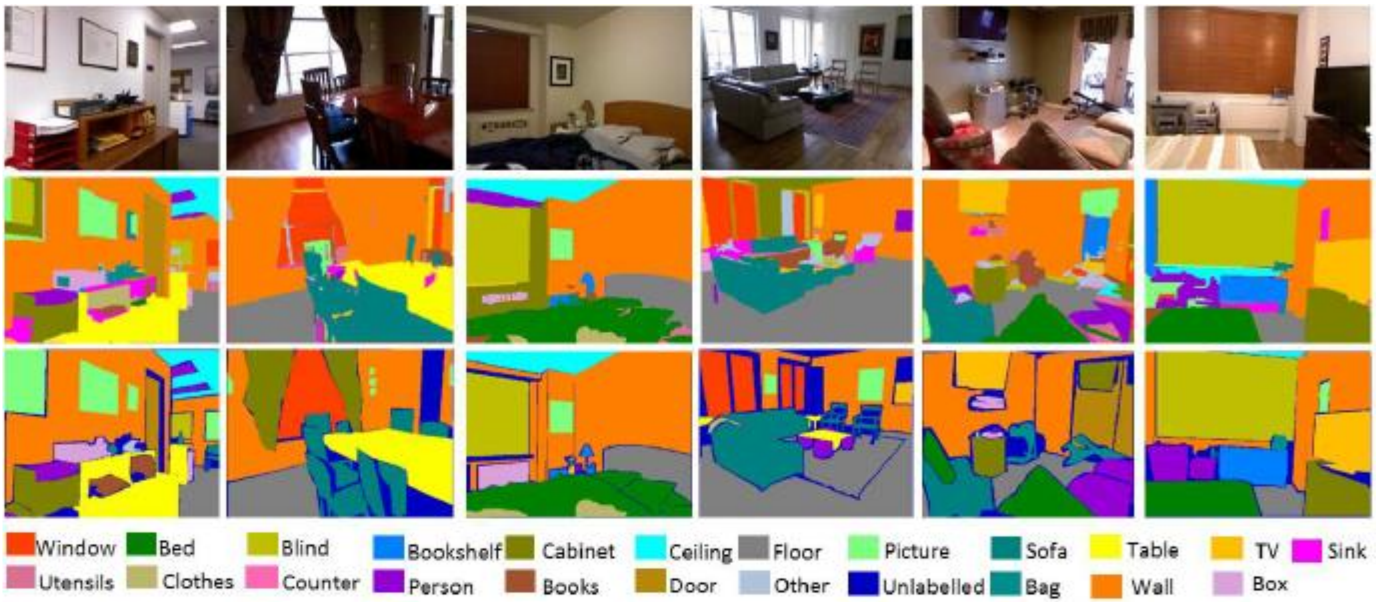
KEY

B	Begin noun phrase	*V*	Verb
I	Within noun phrase	*IN*	Preposition
O	Not a noun phrase	*PRP*	Possesive pronoun
N	Noun	*DT*	Determiner (e.g., a, an, the)
ADJ	Adjective		

D. Koller et. al. 2009

PGM: Applications

Scene Understanding



"Geometry Driven Semantic Labeling of Indoor Scenes", Salman Hameed Khan et. Al. ECCV 2014

NUS National University of Singapore | School of Computing

CS5340
Uncertainty Modelling in AI

Lecture 1:
Introduction to probabilities and probability distributions

Asst. Prof. Lee Gim Hee

AY 2018/19

Semester 1

Course Schedule

Week	Date	Topic	Remarks
1	15 Aug	Introduction to probabilities and probability distributions	
2	22 Aug	Fitting probability models	Hari Raya Haji*
3	29 Aug	Bayesian networks (Directed graphical models)	
4	05 Sep	Markov random Fields (Undirected graphical models)	
5	12 Sep	I will be traveling	No Lecture
6	19 Sep	Variable elimination and belief propagation	
-	26 Sep	Recess week	No lecture
7	03 Oct	Factor graph and the junction tree algorithm	
8	10 Oct	Parameter learning with complete data	
9	17 Oct	Mixture models and the EM algorithm	
10	24 Oct	Hidden Markov Models (HMM)	
11	31 Oct	Monte Carlo inference (Sampling)	
12	07 Nov	Variational inference	
13	14 Nov	Graph-cut and alpha expansion	

* **Make-up lecture**: 25 Aug (Sat), 9.30am-12.30pm, LT 15

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.

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, categoricial, univariate and multivariate normal distributions.

4. Explain the use of conjugate distributions.

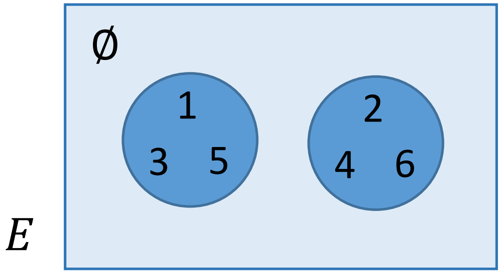
Outcome and Event Spaces

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

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

- Event space $E \subseteq 2^{\Omega}$ is the 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 basis properties:

1. It contains the empty event $\emptyset$, and the trivial event Ω.

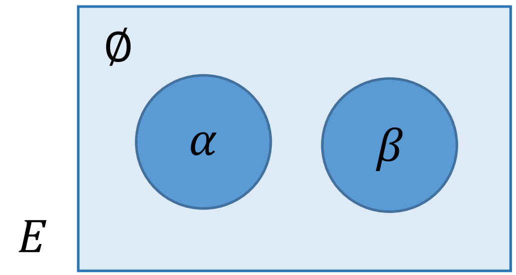
2. It is closed under union, i.e. if $\alpha, \beta \in E$, then so is $\alpha \cup \beta$.

3. It is closed under complementation, i.e. if $\alpha \in E$, then so is $\Omega - \alpha$.

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 $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$.



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 $X\colon \Omega \rightarrow E$ that maps a set of possible outcomes Ω to a event space E.

- The probability that X takes on a value in a measurable set $S \subseteq E$ is written as:

$$P(X \in S) = P(\{\omega \in \Omega \mid X(\omega) \in S\})$$

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 used to refer to a generic value of a random variable X, a.k.a. realization of the random variable.

  Example: We write $P(X = x) \geq 0$ for all $x \in Val(X)$.

- $P(x)$ is often used as a shorthand notation for $P(X = x)$.

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

Random Variables

- 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.

Examples:

Random variables with discrete values

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

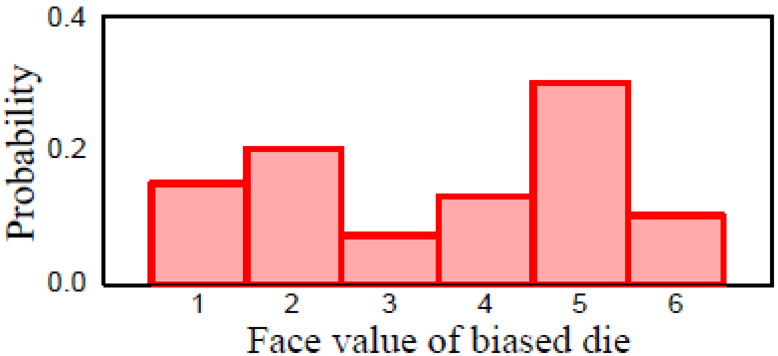
Continuous random variables

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

Probability Distributions: Discrete Vs Continuous

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

$P(x)$



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

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

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

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

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

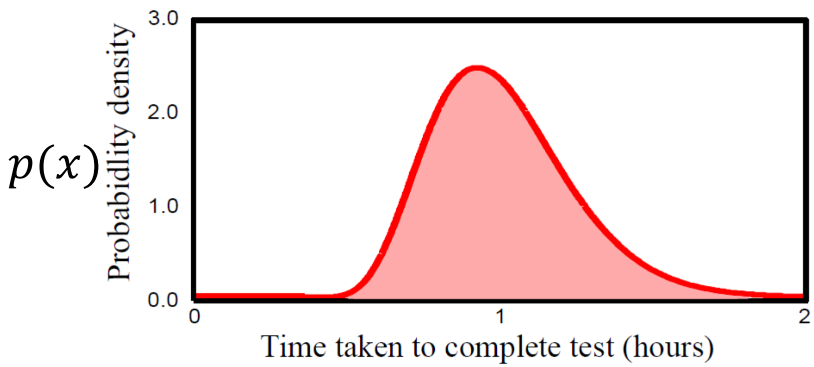
Probability Distributions: Discrete Vs Continuous

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

$$\int_{Val(X)} p(x)dx = 1$$

$$p(X = x^i) \geq 0, \quad \forall \; x^i \in Val(X)$$

$$p(\Omega) \neq 1$$



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

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

$$P(X = x^i) = 0, \quad \forall \; x^i \in Val(X)$$

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

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

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

Probability Distributions: Discrete Vs Continuous

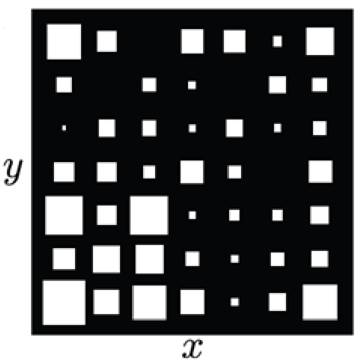
In this course, we abuse the notation by denoting both the probability mass function and probability density function as the lower case $p(x)$!

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

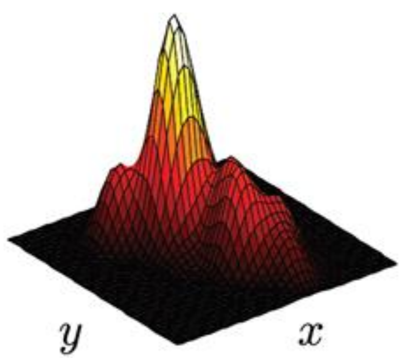
Probability: Joint Probability

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

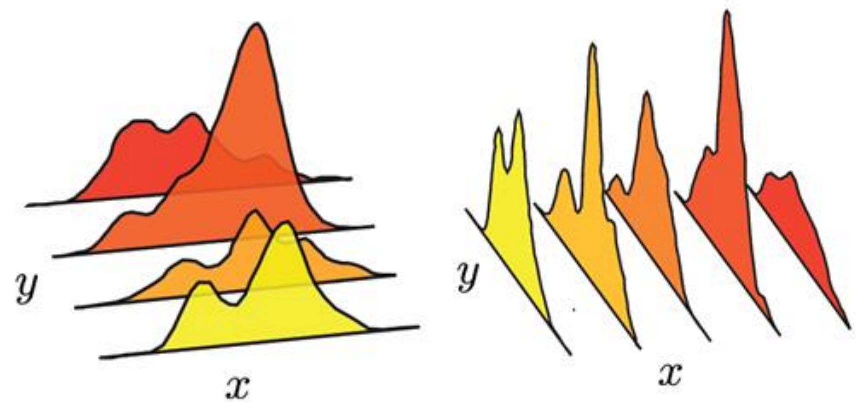
Discrete



Continuous



Discrete-Continuous



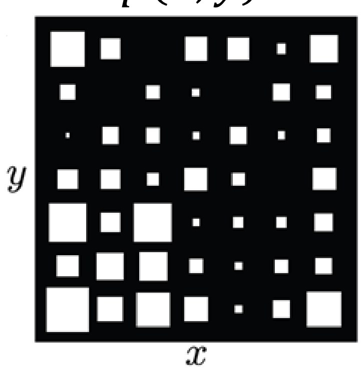
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

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. $p(a, b, c, \dots)$.

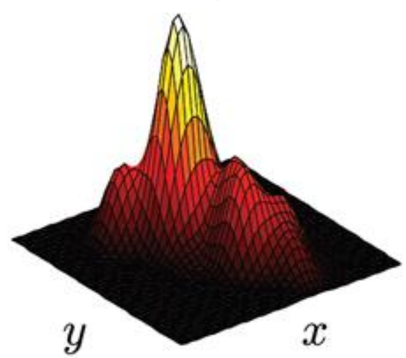
Discrete

$p(x, y)$



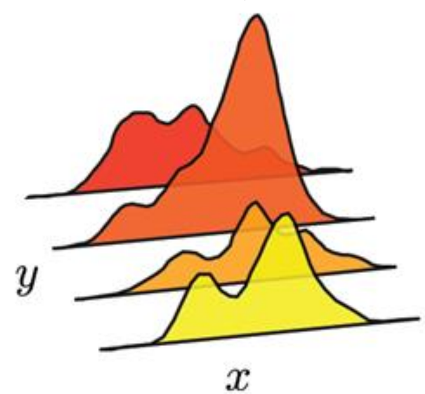
Continuous

$p(x, y)$

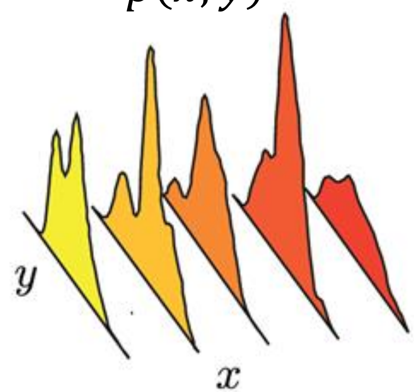


Discrete-Continuous

$p(x, y)$



$p(x, y)$



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

Probability: Marginalization

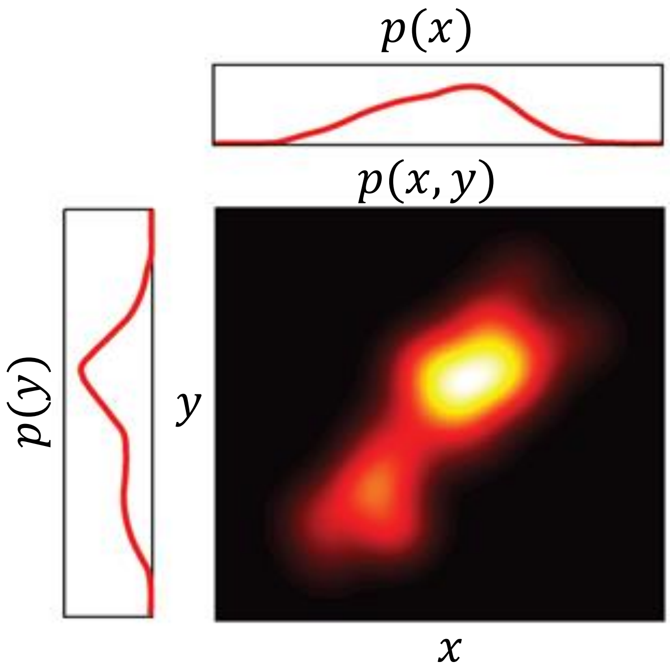
- Recover probability distribution of any variable in a joint distribution by integrating (or summing) over all other variables.

- 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

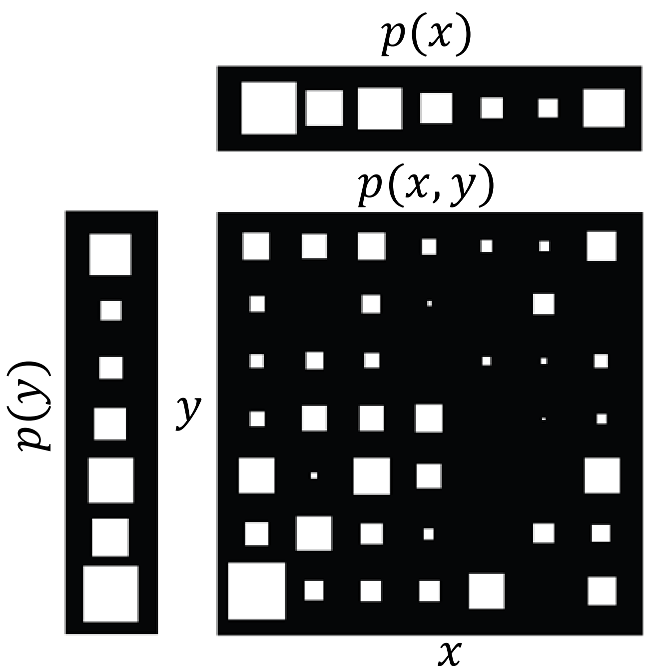
- Recover probability distribution of any variable in a joint distribution by integrating (or summing) over all other variables.

- 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

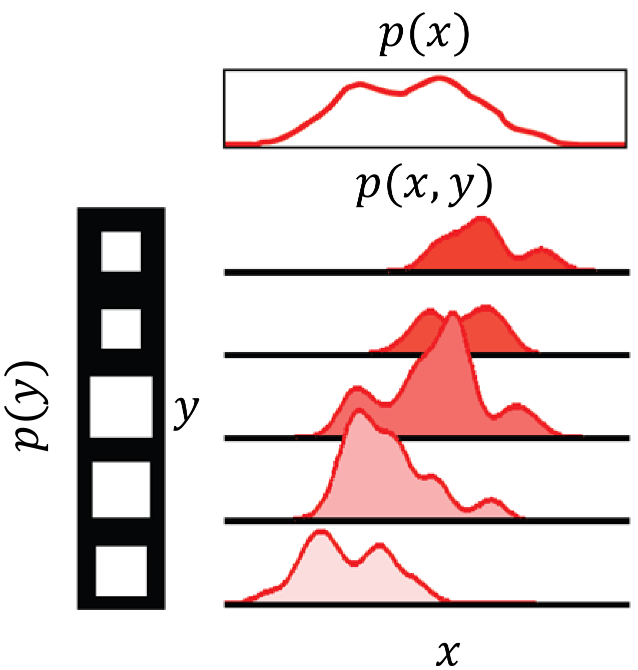
- Recover probability distribution of any variable in a joint distribution by integrating (or summing) over all other variables.

- 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

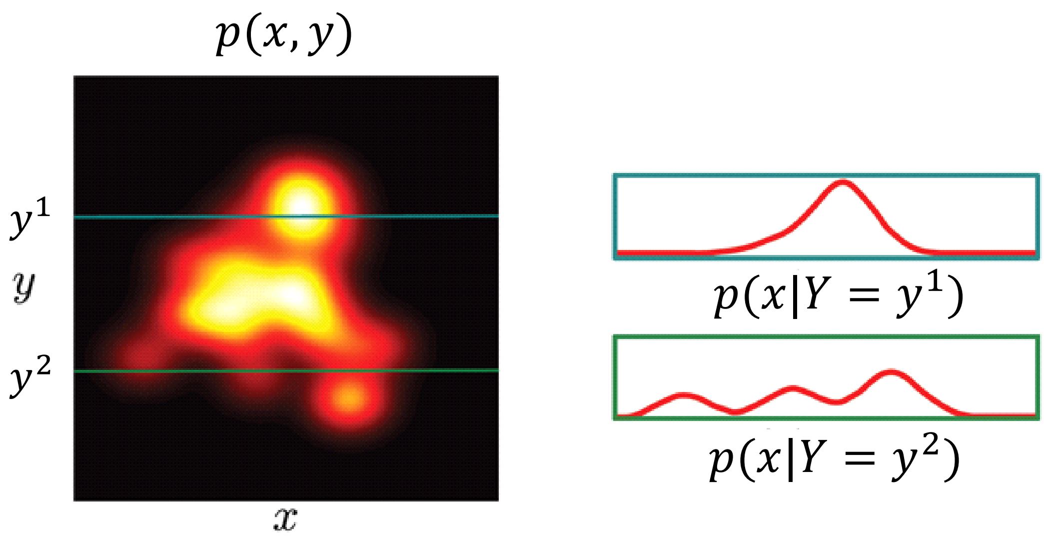
- Works in higher dimensions too!

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 $= y^*$".
- Also known as "chain rule" or "product rule" of probability.
- Relative propensity of the random variable X to take different outcomes given that the random variable Y is fixed to value y^*.

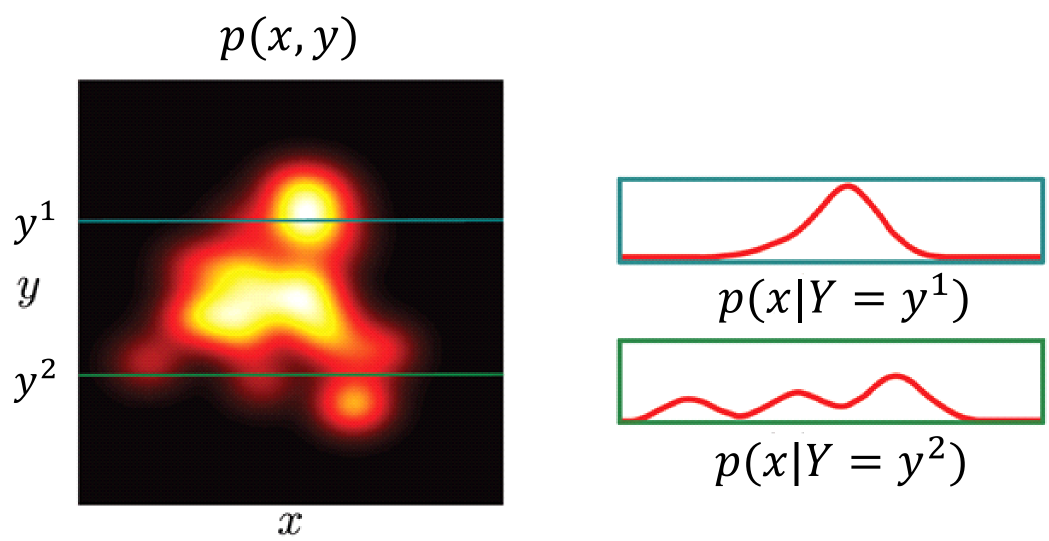


Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Conditional Probability

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

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



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:

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

Hence, the name "product rule"!

Probability: Conditional Probability

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

- Works for higher dimensions too!

  Example:

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

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

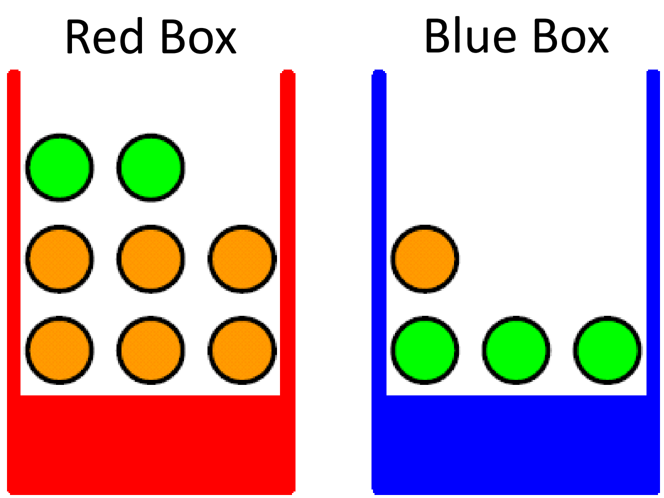
$$p(y|x) = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

Posterior – what we know about Y after observing X

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

Probability: Example

Let random variables B and F represent the box color and type of fruit respectively, where $Val(B) = \{r, b\}$ and $Val(F) = \{a, o\}$.



2 apples,
6 oranges

3 apples,
1 orange

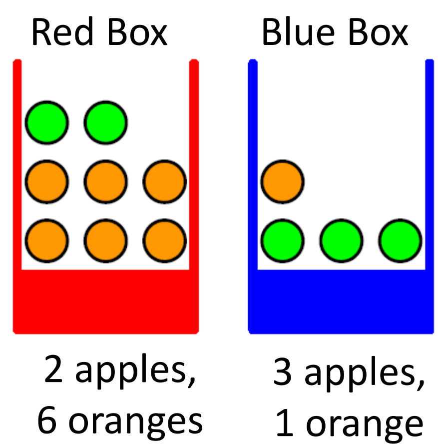
Image source: “Pattern Recognition and Machine Learning”, Christopher Bishop

Probability: Example

Given:

- Probabilities of selecting either the red or the blue boxes,

$$p(B = r) = 0.4$$
$$p(B = b) = 0.6$$



- Conditional probabilities for the type of fruit, given the selected box,

$$p(F = a|B = r) = 0.25$$
$$p(F = o|B = r) = 0.75$$
$$p(F = a|B = b) = 0.75$$
$$p(F = o|B = b) = 0.25$$

Image source: “Pattern Recognition and Machine Learning”, Christopher Bishop

Probability: Example

Find:

a) The overall probability of choosing an apple.

b) Identify the color of the box if we observed that an orange has been selected.

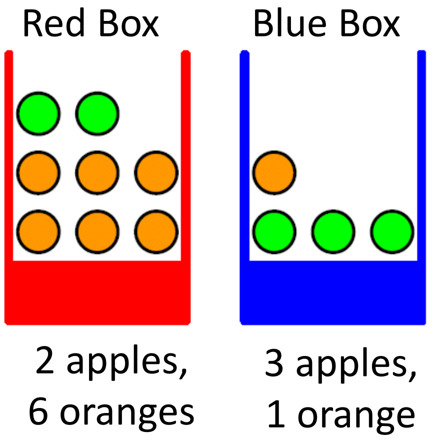


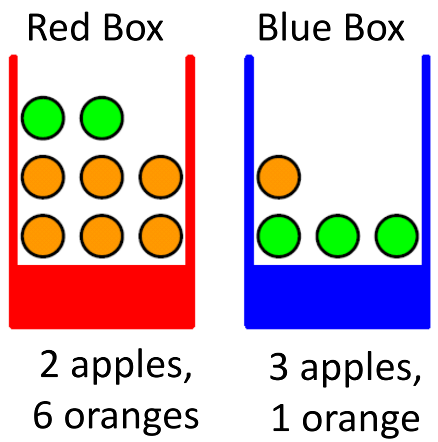
Image source: "Pattern Recognition and Machine Learning", Christopher Bishop

Probability: Example

Solution:

a) The overall probability of choosing an apple.

Using the sum and product rules of probability:



$$p(F = a) = \sum_B p(F = a|B)p(B)$$

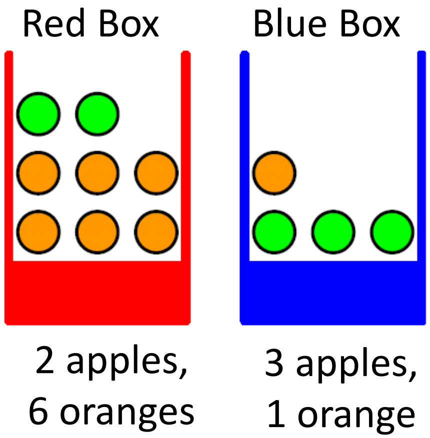
$$= p(F = a|B = r)p(B = r) + p(F = a|B = b)p(B = b)$$

$$= (0.25)(0.4) + (0.75)(0.6) = 0.55$$

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

Probability: Example

Solution:



b) Identify the color of the box if we observed that an orange has been selected.

Using Bayes' theorem:

$$p(B=r|F=o) = \frac{p(F=o|B=r)p(B=r)}{p(F=o)}$$

$$= \frac{p(F=o|B=r)p(B=r)}{1-p(F=a)} = \frac{(0.75)(0.4)}{1-0.55}$$

$$= 0.667$$

$$p(B=b|F=o) = 1 - p(B=r|F=o) = 1 - 0.667 = 0.333$$

The orange is more likely to be selected from the red box!

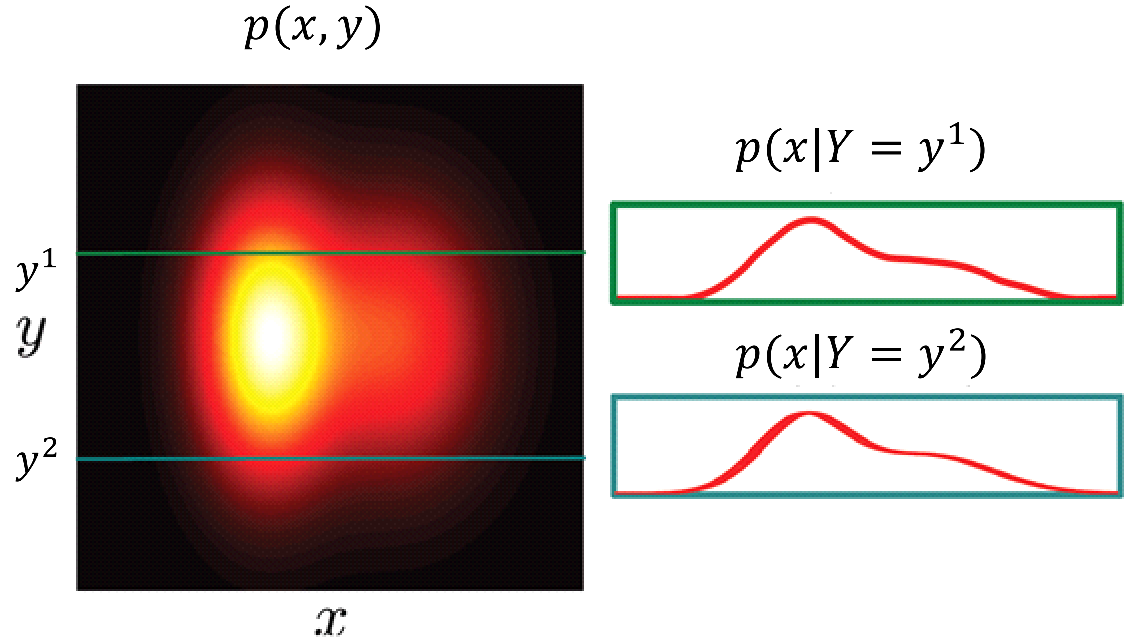
Image source: "Pattern Recognition and Machine Learning", Christopher Bishop

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.

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

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



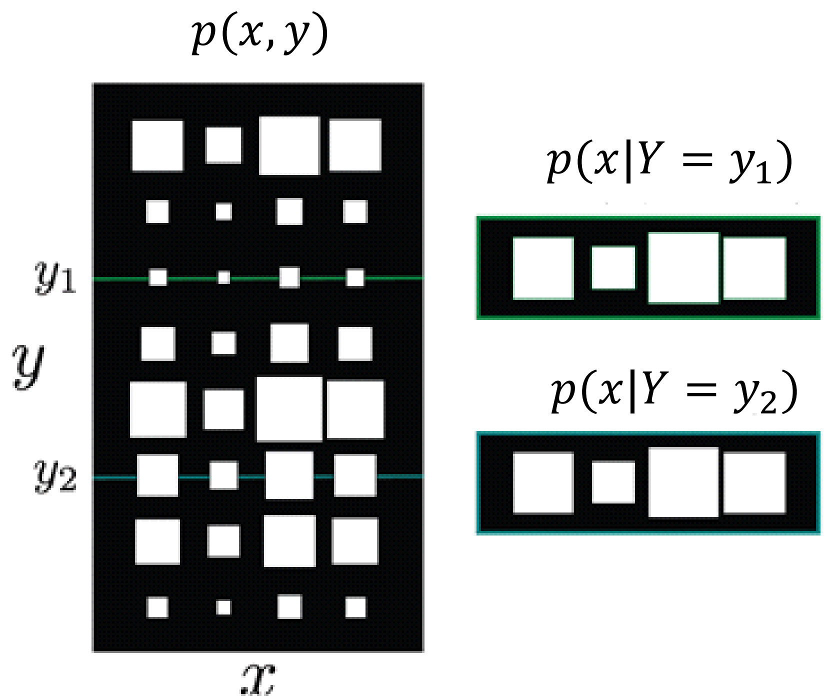
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Probability: Independence

- 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:

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

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.

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

Probability: Rules of Expectation

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

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

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

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

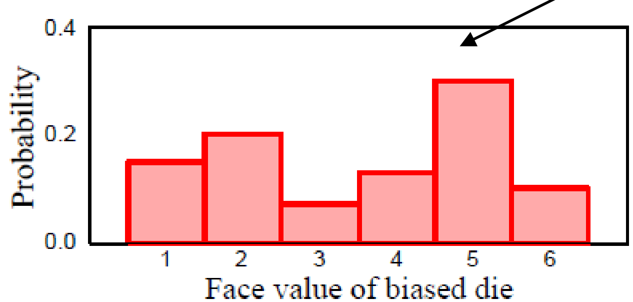
if X and Y are independent

Probability Distributions

- We have seen the definitions of random variables, probability, and rules for manipulating probabilities.
- One question that remains unanswered is: “How do we assign the values of $p(X = x^i)$?”

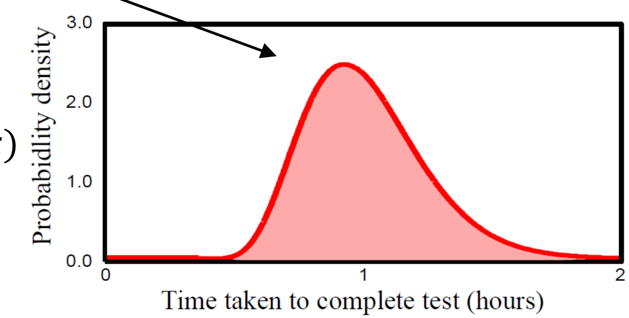
How to get $p(X = x^i)$?

$p(x)$



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

$p(x)$



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

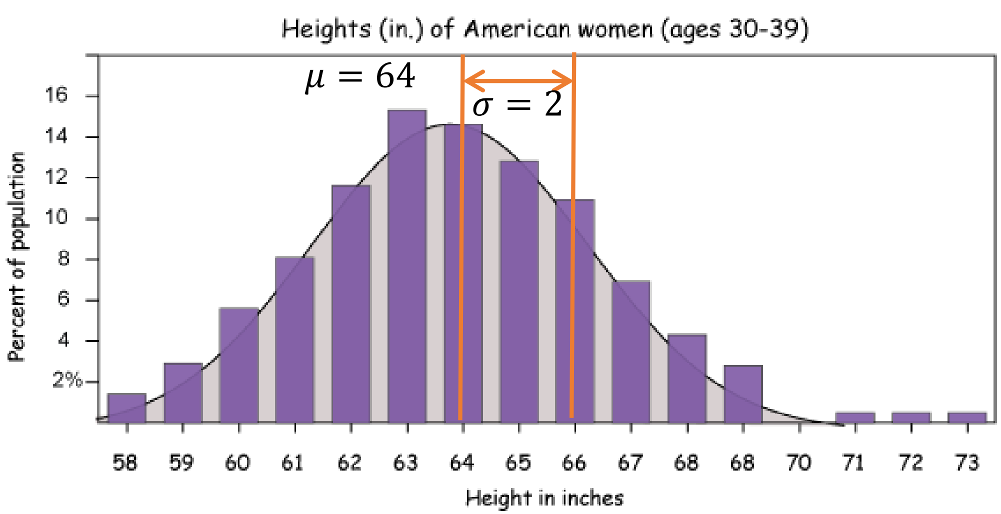
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Probability Distributions

Q: "How do we assign the probability values?"

A: Use **probability distributions** defined over some **parameters** learned from data!

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	$x \in \{1, 2, \ldots, K\}$	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	$\mathbf{x} = [x_1, x_2, \ldots, x_K]^T$ $x_k \in [0, 1], \sum_{k=1}^{K} x_k = 1$	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

Bernoulli Distribution

Jacob Bernoulli
1654–1705

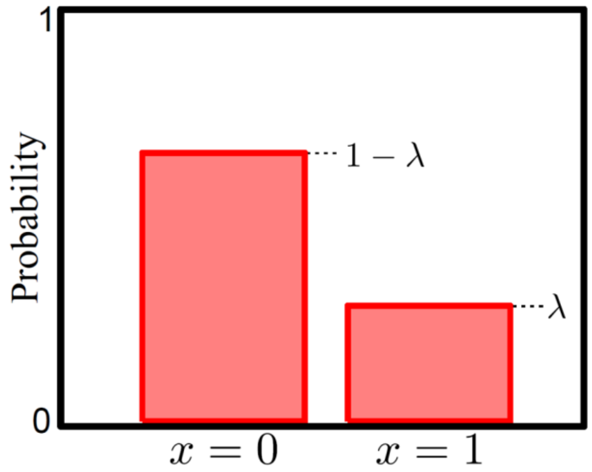
- Single binary random variable X, i.e. $x \in \{0,1\}$
- A single parameter $\lambda \in [0,1]$.

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

Or

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

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



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

Categorical Distribution

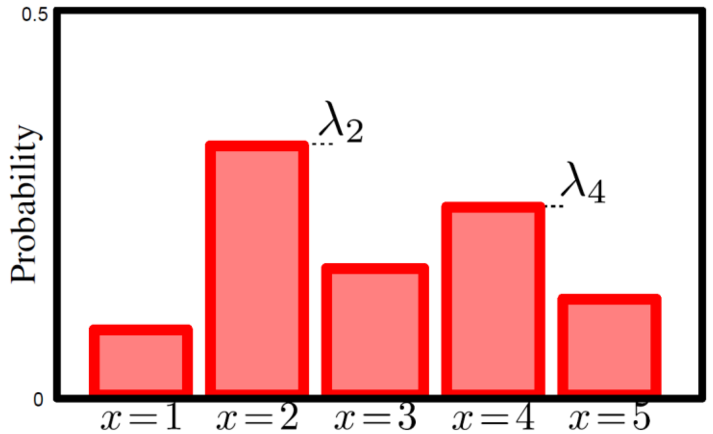
- Discrete variables $\boldsymbol{X}$ that take on 1-of-K possible mutually exclusive states, e.g. a K-faced die.
- $\boldsymbol{x}$ is represented by a K-dimensional vector $\mathbf{e}_k$ in which one of the elements $x_k = 1$, and $\sum_{k=1}^{K} x_k = 1$.
- e.g. $K = 5$, and $\boldsymbol{x} = \mathbf{e}_3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

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

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

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

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

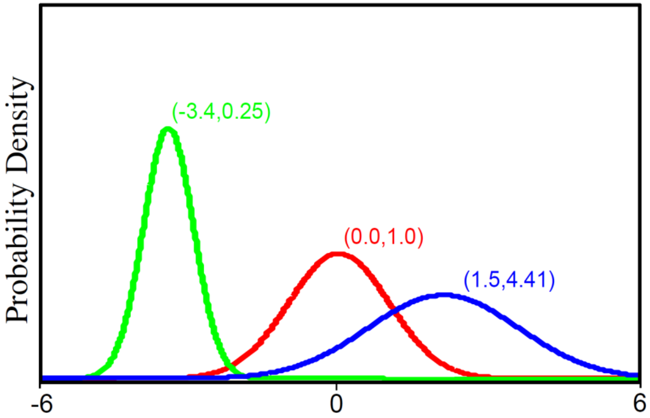


Image sources: "Pattern Recognition and Machine Learning", Christopher Bishop
"Computer Vision: Models, Learning, and Inference", Simon Prince

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

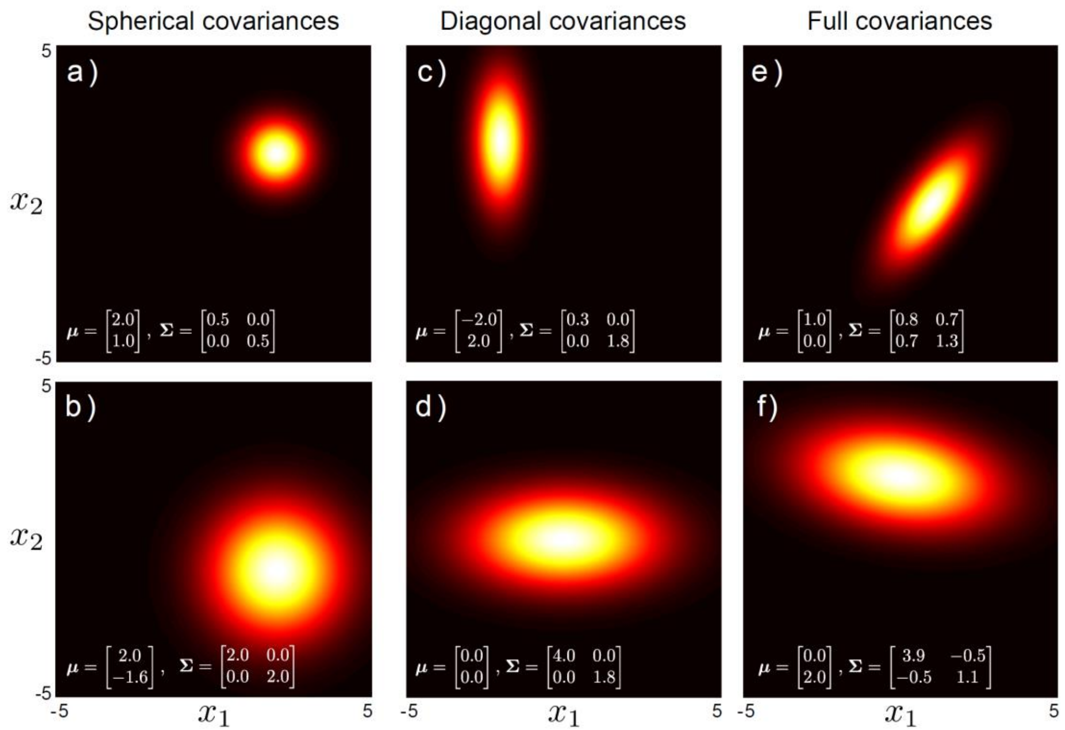
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{ -0.5(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \}$$

$$p(\boldsymbol{x}) = \mathrm{Norm}_{\boldsymbol{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$$

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

Conjugate Distributions

- Conjugate distributions model the parameters of the 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.

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

1. Learning the parameters θ of a probability distribution:

Recall the Bayes' Rule:

1. Choose prior that is conjugate to likelihood

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

2. Implies that posterior must have same form as conjugate prior distribution, i.e. closed-form.

3. Posterior must be a distribution which implies that evidence must equal constant κ from conjugate relation.

Importance of Conjugate Distributions

2. Marginalizing over parameters:

$$p(x^*|\boldsymbol{x}) = \int p(x^*|\theta\,)p(\theta|\boldsymbol{x})d\theta$$

2. Integral becomes easy --the product becomes a constant times a distribution.

1. Chosen as conjugate to other term.

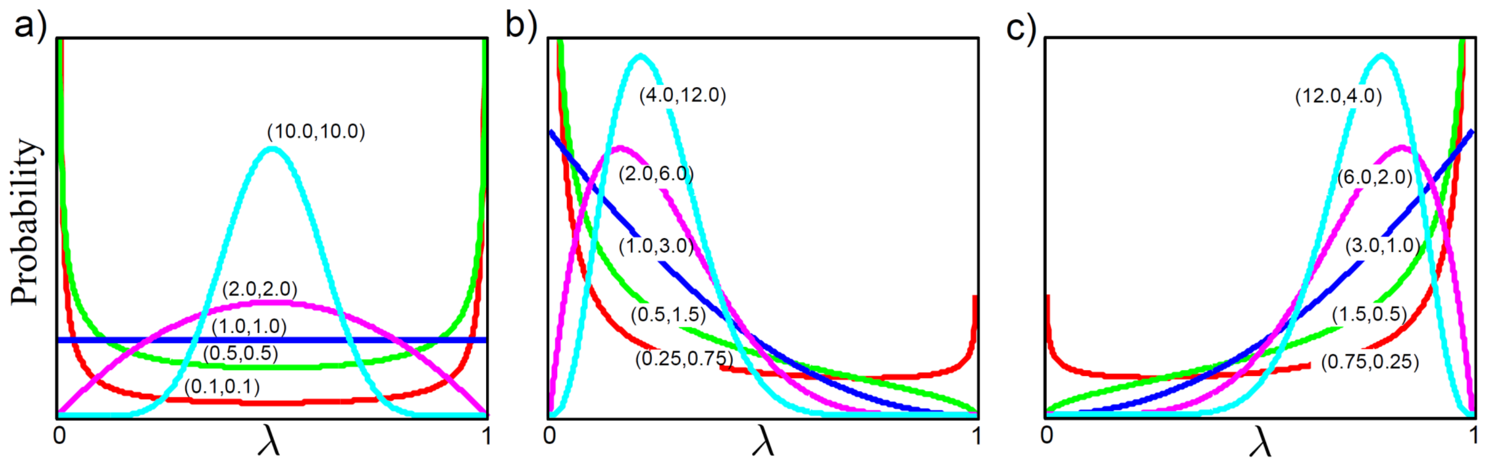
Integral of constant times probability distribution
= constant times integral of <u>probability distribution</u>
= constant x <u>1</u> = constant

Conjugate Distribution: Beta Distribution

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

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

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



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

Conjugate Distribution: Beta Distribution

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

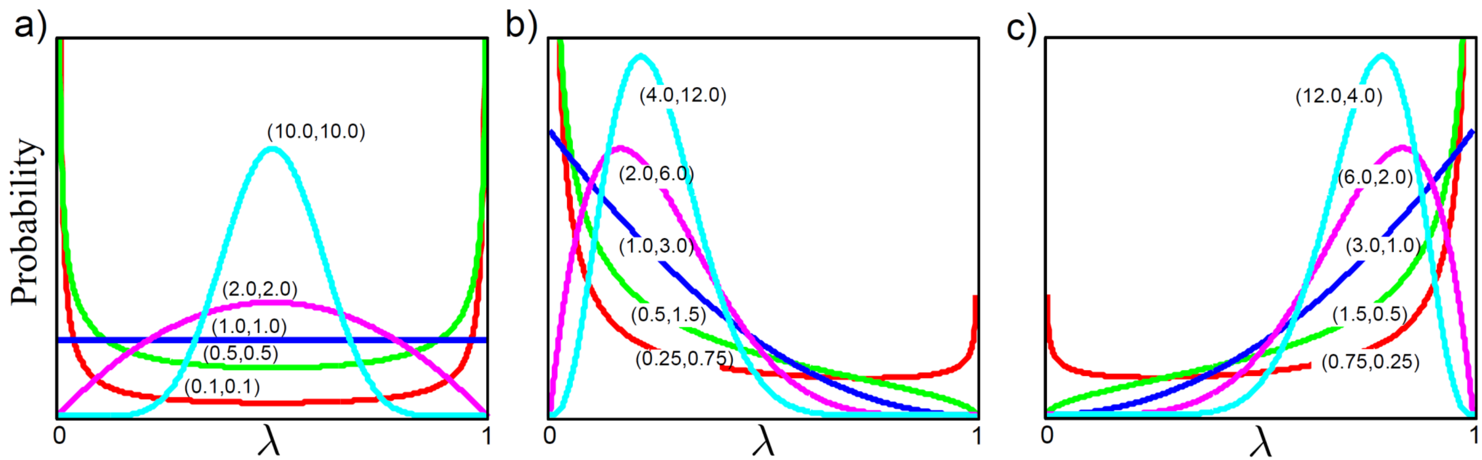
$$p(\lambda) = \mathrm{Beta}_{\lambda}[\alpha,\beta]$$

Gamma Function:

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

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

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



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

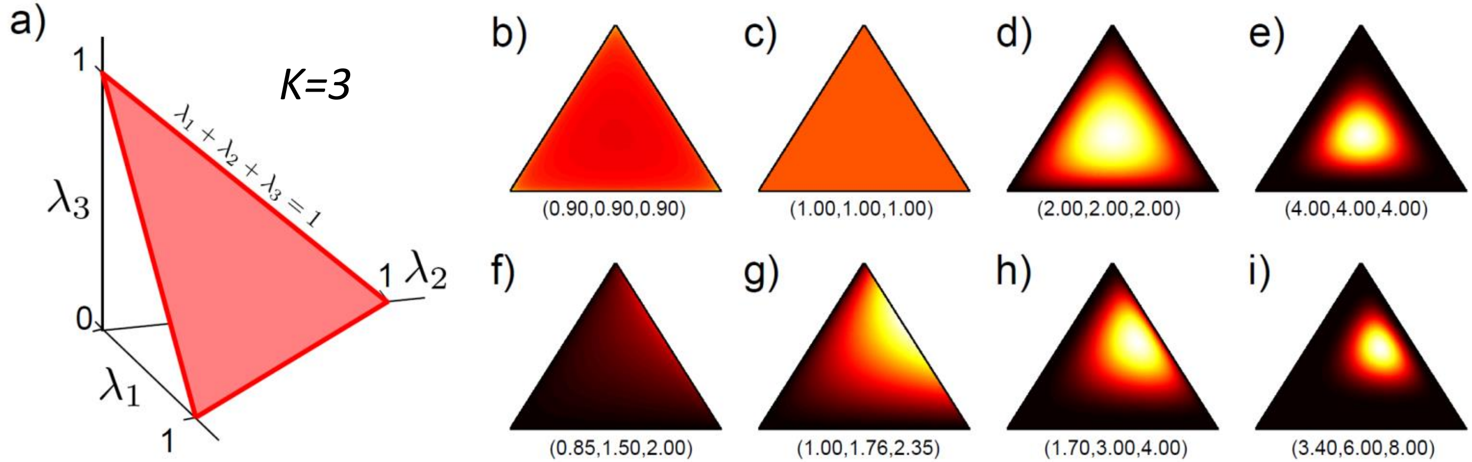
Conjugate Distribution: Dirichlet Distribution

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

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

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

Peter Gustav Lejeune Dirichlet (1805-1859)



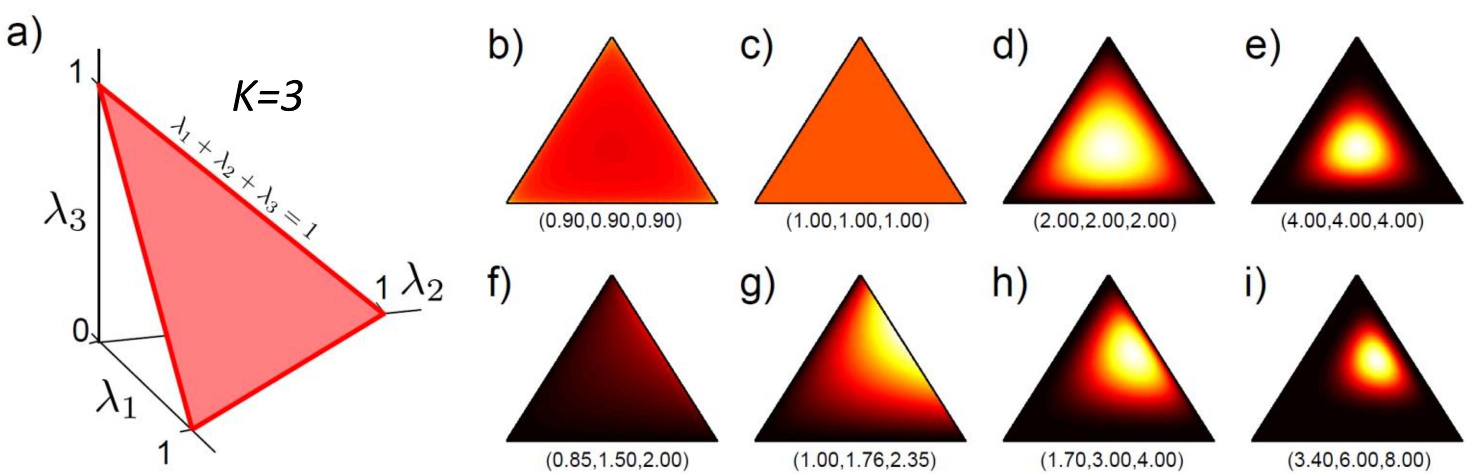
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince
http://www.amt.edu.au/biogdirichlet.html

Conjugate Distribution: 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 $\alpha_k > 0$.



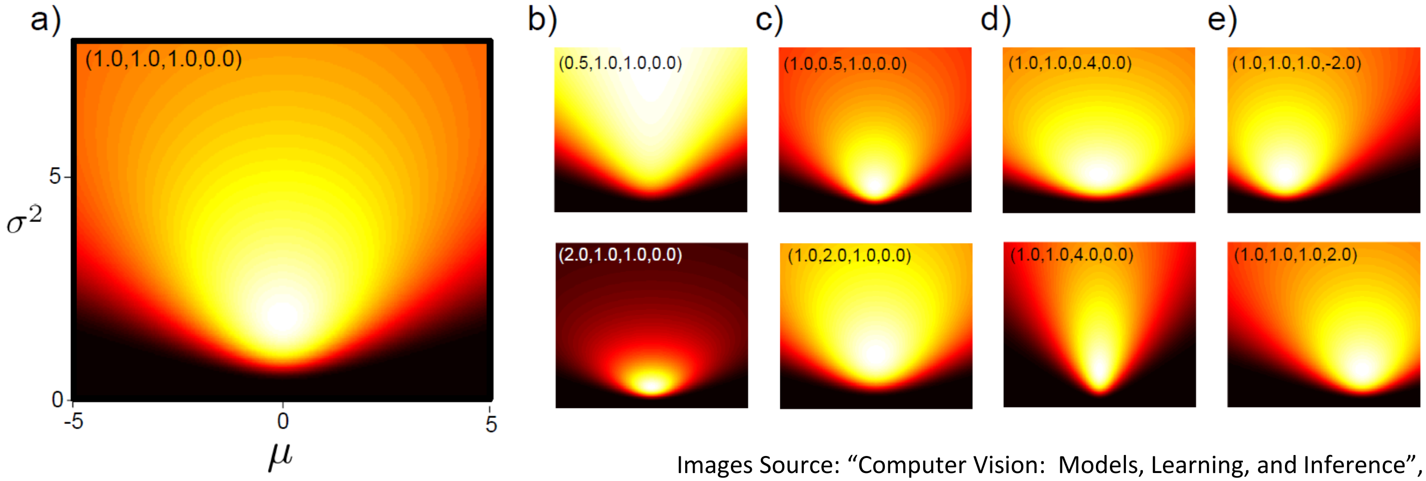
Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Conjugate Distribution: Normal Inverse Gamma Distribution

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

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

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



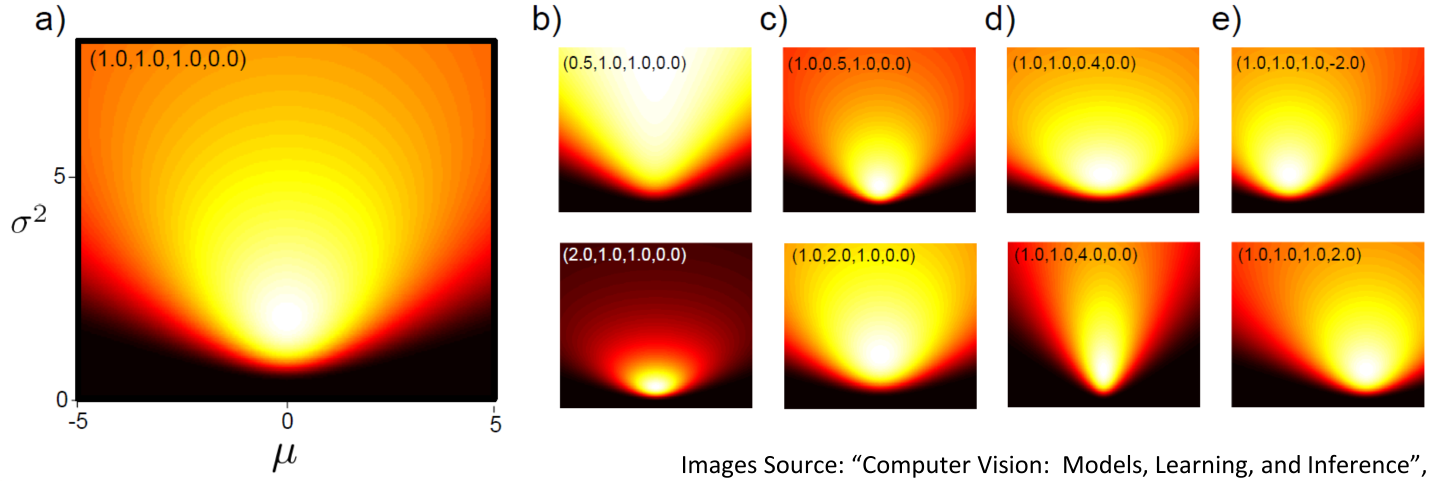
Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Conjugate Distribution: Normal Inverse Gamma Distribution

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

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

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



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

Conjugate Distribution: Normal Inverse Wishart

John Wishart (1898-1956)

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

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

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

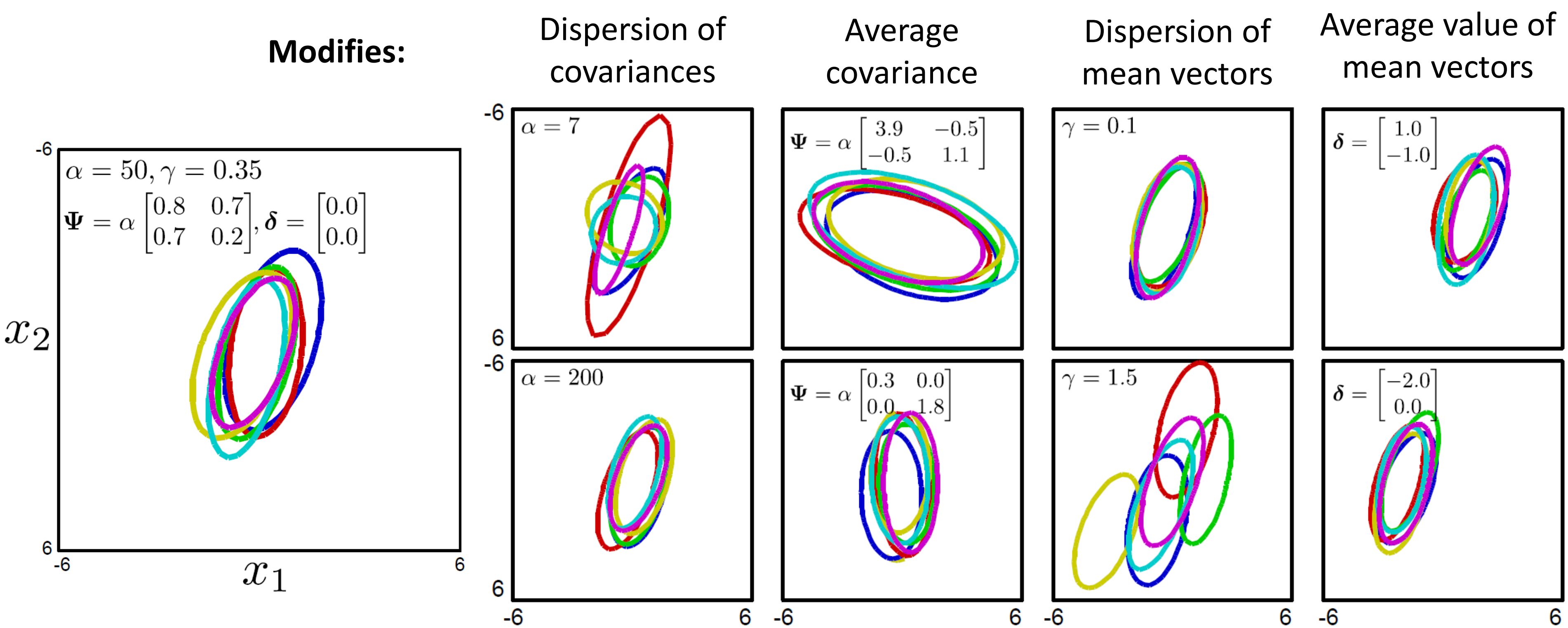
- Four hyperparameters: a positive scalar α, a positive definite matrix $\boldsymbol{\Psi} \in \mathbb{R}_+^{D\times D}$, a positive scalar γ, and a vector $\boldsymbol{\delta} \in \mathbb{R}^D$.

Multivariate gamma function:

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

Conjugate Distribution: Normal Inverse Wishart

- Samples from Normal Inverse Wishart:



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