CS5340
Uncertainty Modelling in AI

Lecture 1:
Introduction to Probabilistic Reasonings

Asst. Prof. Lee Gim Hee

AY 2020/21

Semester 1

Course Schedule

Week	Date	Topic	Remarks
1	12 Aug	Introduction to probabilistic reasoning	**1830hrs: MS Teams (Live Introduction)**
2	19 Aug	Bayesian networks (Directed graphical models)	
3	26 Aug	Markov random Fields (Undirected graphical models)	**1830hrs: MS Teams discussions**
4	02 Sep	Variable elimination and belief propagation	**Assignment 1:** Belief propagation and maximal probability (15%)
5	09 Sep	Factor graph and the junction tree algorithm	
6	16 Sep	Parameter learning with complete data	**Assignment 1:** Due **Assignment 2:** Junction tree and parameter learning (15%) **1830hrs: MS Teams discussions**
-	23 Sep	Recess week	**No lecture**
7	30 Sep	Mixture models and the EM algorithm	**Assignment 2:** Due **Online quiz 1 (20%)**
8	07 Oct	Hidden Markov Models (HMM)	**Assignment 3:** Hidden Markov model (15%)
9	14 Oct	Monte Carlo inference (Sampling)	**1830hrs: MS Teams discussions**
10	21 Oct	Variational inference	**Assignment 3:** Due **Assignment 4:** MCMC Sampling (15%)
11	28 Oct	Variational Auto-Encoder and Mixture Density Networks	
12	04 Nov	Graph-cut and alpha expansion	**Assignment 4:** Due **1830hrs: MS Teams discussions**
-	11 Nov	--	**Online quiz 2 (20%)**

NUS School of Computing | National University of Singapore

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.

Probability Space

- A probability space (Ω, E, P) models a process consisting of outcomes that occur randomly.

- Consists of three parts:

    1. Outcome or sample space Ω
    2. Event space E
    3. Probability distribution $P: E \rightarrow \mathbb{R}$

Outcome/Sample and Event Spaces

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

- 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.

Outcome and Event Spaces

- Event space must satisfy three basis properties:

    1. It **must contain** the empty event ϕ, and the trivial event Ω.

    2. It is closed under countable unions, i.e. if $\alpha_i \in E \ \forall \ i = 1, 2, \dots$, then so is $\bigcup_{i=1}^{\infty} \alpha_i$.

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

Outcome/Sample and Event Spaces

Example 1:

Let's consider a 6-faced dice. The outcome/sample space is given by $\Omega = \{1, 2, 3, 4, 5, 6\}$.

A possible event space is $E = \{\{1,3,5\}, \{2,4,6\}, \emptyset, \{1,2,3,4,5,6\}\}$, i.e. event of a throw is even or odd.

Check:

1. E contains the empty $\emptyset$ and trivial $\{1,2,3,4,5,6\}$ sets.

2. Let $\alpha_1 = \{1,3,5\}$ and $\alpha_2 = \{2,4,6\}$, i.e., $\alpha_1, \alpha_2 \in E$, then $\alpha_1 \cup \alpha_2 = \{1,2,3,4,5,6\} \in E$ because E is closed under countable unions.

3. Let $\alpha = \{2,4,6\} \in E$, then $\{1,2,3,4,5,6\} - \{2,4,6\} = \{1,3,5\} = \{\Omega - \alpha\} \in E$ because E is closed under complement.

**Remark:** Check for yourself that 2 and 3 are always true $\forall\, \alpha \in E$!

Outcome/Sample and Event Spaces

Example 2:

Let's consider measuring the lifetime of a lightbulb. The outcome/sample space is given by $\Omega = [0, \infty)$.

A possible event space is $E = \{[0, 90), [90, \infty), \emptyset, [0, \infty)\}$, i.e. event of lightbulb lifespan ≥ 90.

Check:

1. E contains the empty $\emptyset$ and trivial $[0, \infty)$ sets.

2. Let $\alpha_1 = \emptyset$ and $\alpha_2 = [0, \infty)$, i.e., $\alpha_1, \alpha_2 \in E$, then $\alpha_1 \cup \alpha_2 = [0, \infty) \in E$ because E is closed under countable unions.

3. Let $\alpha = [0, 90) \in E$, then $[0, \infty) - [0, 90) = [90, \infty) = \{\Omega - \alpha\} \in E$ because E is closed under complement.

Remark: Check for yourself that 2 and 3 are always true $\forall \, \alpha \in E$!

Outcome/Sample and Event Spaces

Example 3:

Picking 2 marbles, one at a time, from a bag that contains many blue and red marbles. Find the sample space?

Tree Diagram

```
            B ──── BB
        B
            R ──── BR
            B ──── RB
        R
            R ──── RR
```
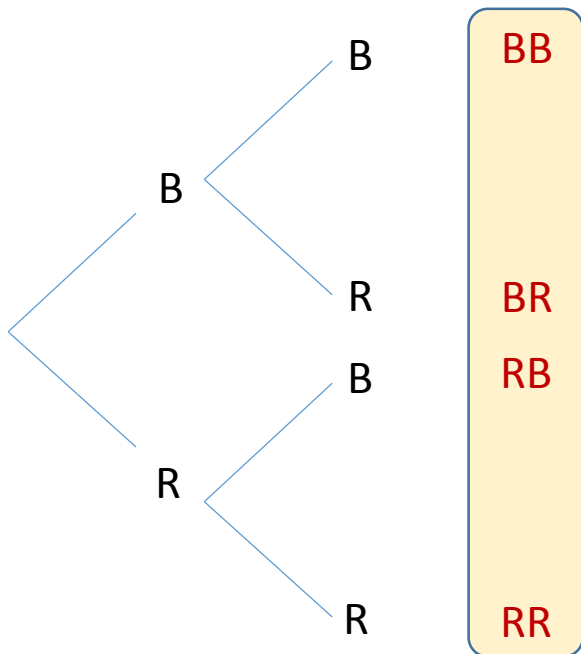
List

{BB, BR, RB, RR}

Table

	B	R
B	BB	BR
R	RB	RR

Probability Distributions

- A probability distribution P over (Ω, E) is a mapping from events in E to real values ($P: E \rightarrow \mathbb{R}$) 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: given a probability space (Ω, E, P), a random variable is a function $X \colon \Omega \to S$ that maps a set of possible outcomes Ω to a measurable space S.

- Typically, S is the set of real numbers, i.e. $S \in \mathbb{R}$.

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 $P(X = x)$ 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. takes values that are real numbers.

Examples:

Random variables with discrete values

- Rolling a six-faced die: $Val(X) = \{1, 2, \ldots, 6\}$
- Weather conditions: $Val(X) = \{"rain", "cloud", "snow", "sun", "wind"\}$
- 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, $P(x)$
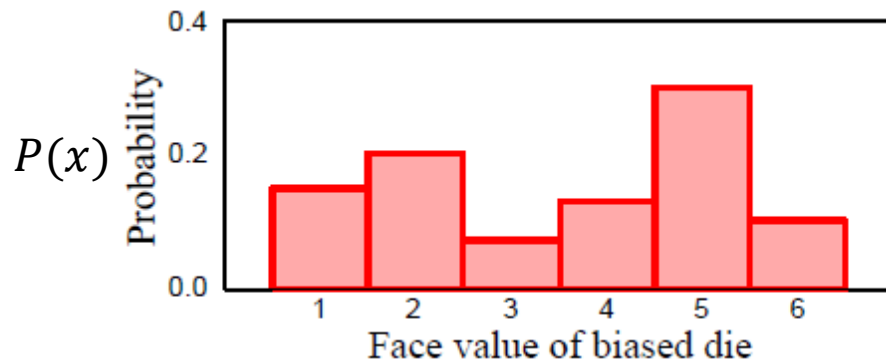
$Val(X) = \{1,2,3,4,5,6\}$

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

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

$$\forall \, i = 1, \dots K,$$

where $K = |Val(X)|$

Probability Distributions: Discrete Vs Continuous

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

$Val(X) = [0,2]$
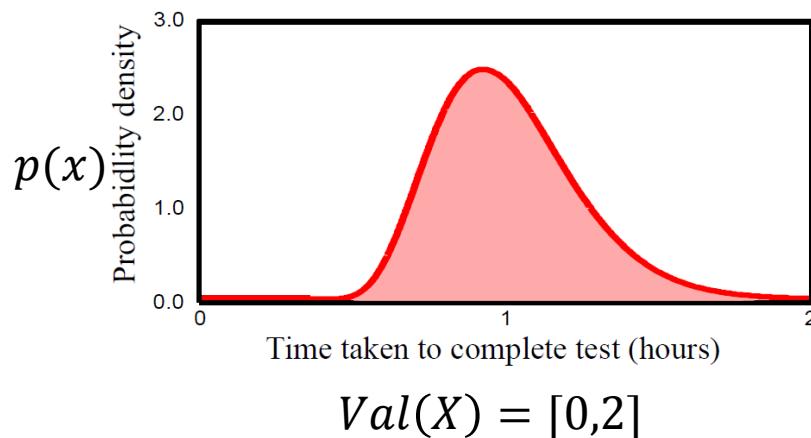
$$\int_{Val(X)} p(x)dx = 1;$$

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

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) $p(x): \mathbb{R} \to \mathbb{R}_{\geq 0}$.

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

$Val(X) = [0,2]$
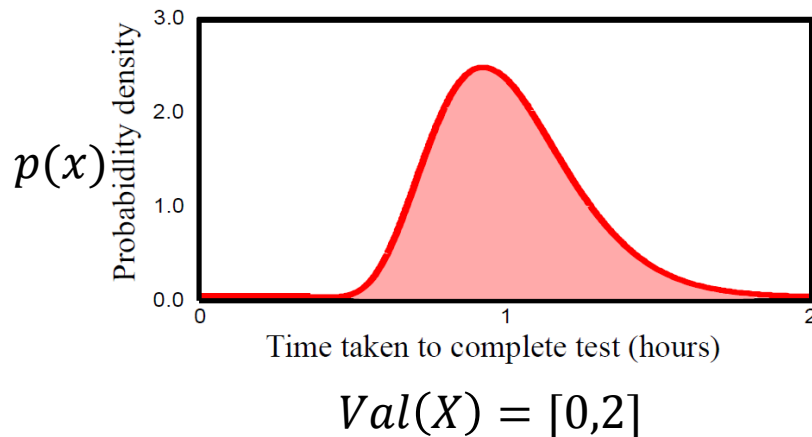
$$P(X \leq a) = \int_{-\infty}^{a} p(x)\, dx$$

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

$$P(X = x^i) = \int_{x^i}^{x^i} p(x)\, dx = 0,$$

$$\forall\ x^i \in Val(X)$$

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.

Probabilistic Reasoning

Probabilistic Modeling:

- The central paradigm of probabilistic reasoning is to:

  1. Identify all relevant variables $X_1, \ldots X_N$ of the environment, and

  2. make a probabilistic model $p(X_1, \ldots X_N)$ of their interactions.

Probabilistic Reasoning

Probabilistic Inference:

- Reasoning (inference) is then performed by:

  1. Introducing evidence that sets variables in known state, and

  2. subsequently computing probabilities of interest, conditioned on this evidence.

Probabilistic Reasoning

- To this end, we require the definitions of joint probability, marginalization, conditional probability, Bayes' rule, and independence.

- In this lecture, we look at the use of these definitions for probabilistic modeling and inference of a small number of variables.

- In the subsequent lectures, we will look at the use of these definitions with **graphical models** for large number of variables.

Probability: Joint Probability

- Consider all combination of events of two random variables X and Y.

- Some combinations of outcomes are more likely than others.
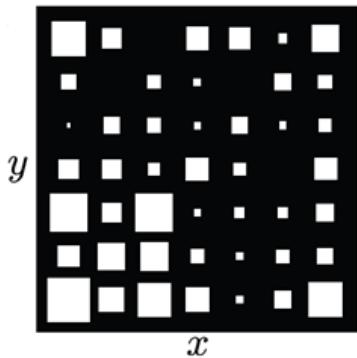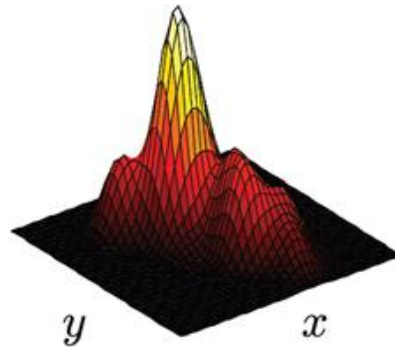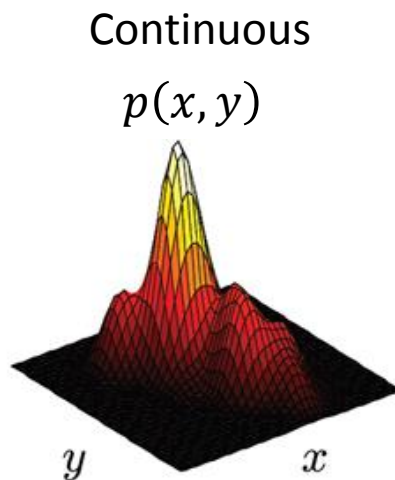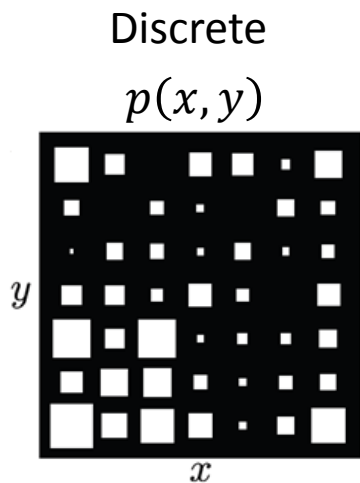
Discrete	Continuous	Discrete-Continuous
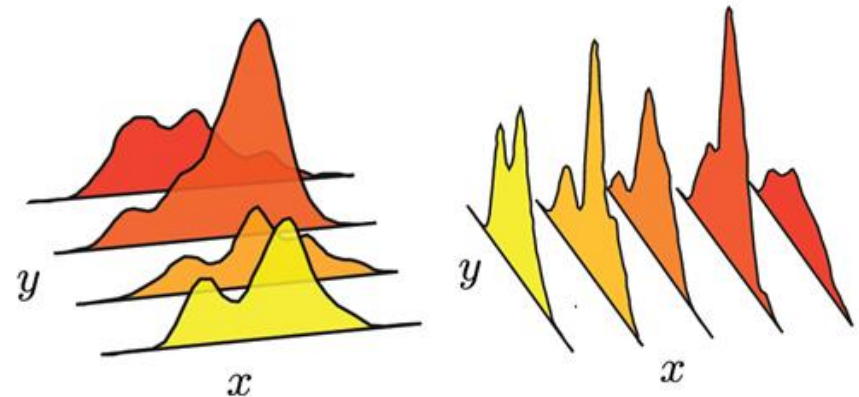
Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

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. $p(a, b, c, \dots)$.
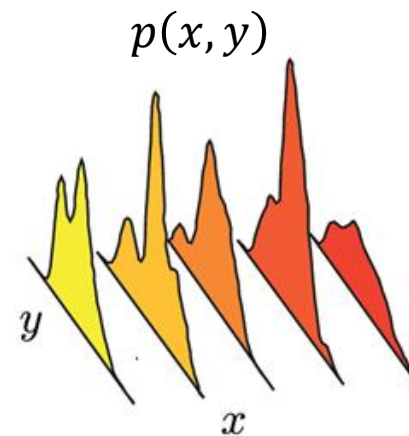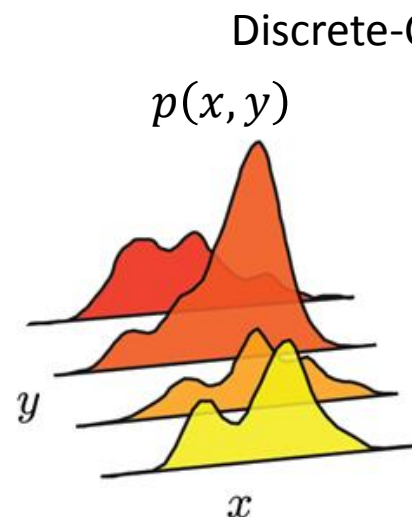
Discrete
$p(x, y)$

Continuous
$p(x, y)$

Discrete-Continuous
$p(x, y)$ $p(x, y)$

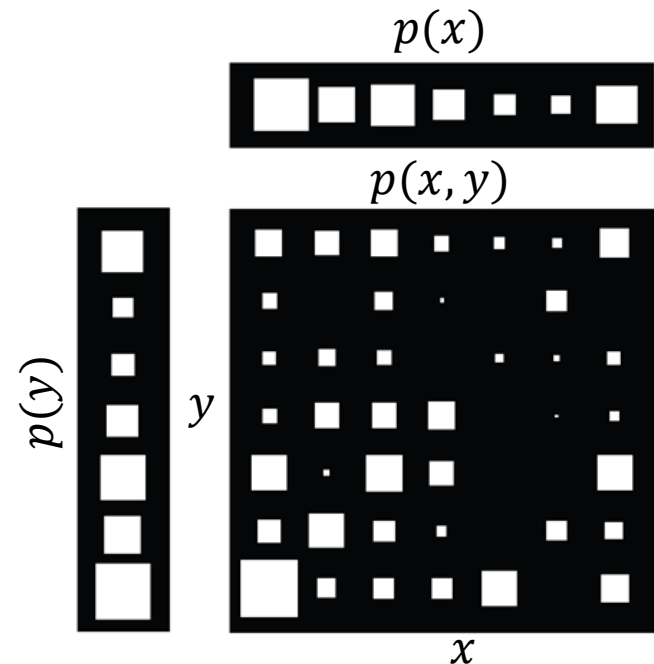Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

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:

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

$$p(y) = \int p(x, y) \, dx$$
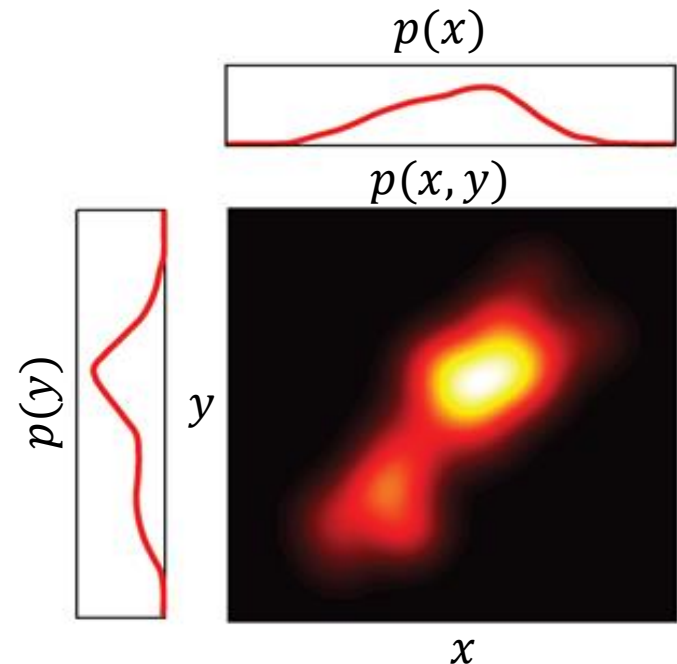
Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince
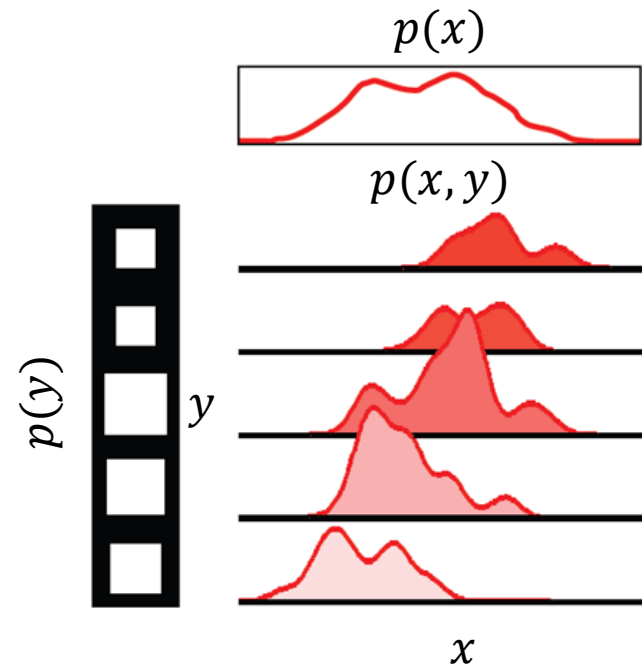
Probability: Marginalization

- 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:

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

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

$p(x)$

$p(x,y)$

$p(y)$

y

x

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
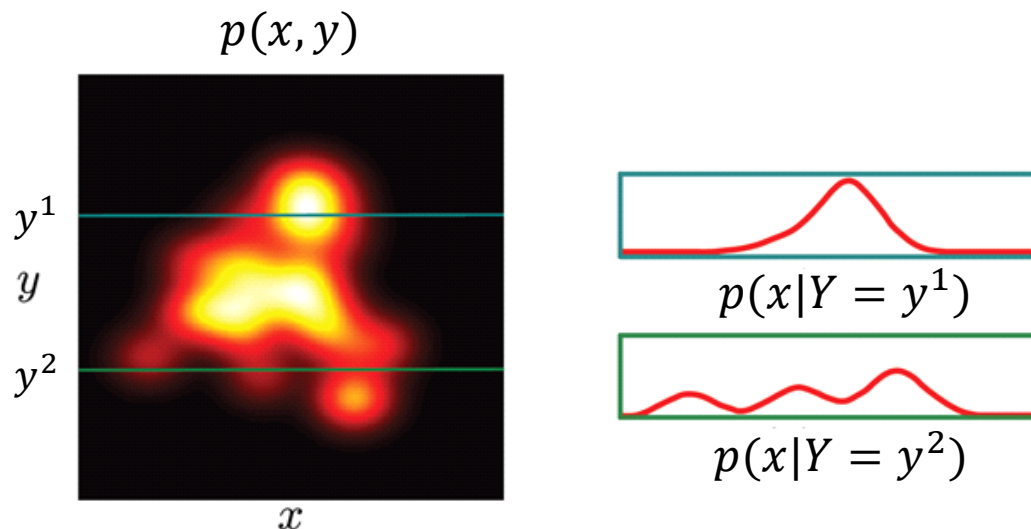
Probability: Marginalization

- 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^*.
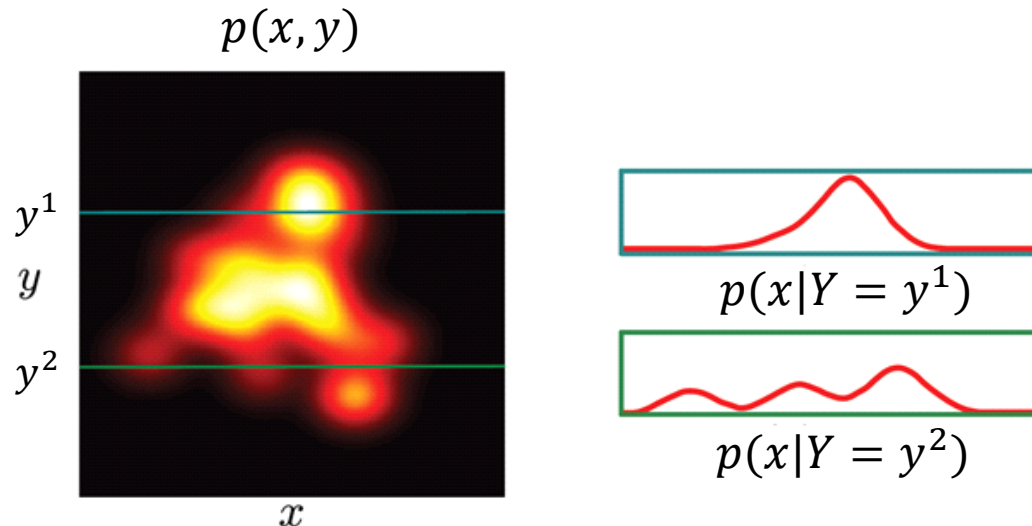

Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Probability: Conditional Probability

- Conditional probability can be extracted from joint probability.

- Extract appropriate slice and normalize (so that the area is 1):

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

$p(x, y)$

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

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

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:

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

Hence, the name "product rule"!

Probability: Conditional Probability

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

- Works for higher dimensions too!

Example:

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

Probability: Bayes' Rule

- Recall:

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

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

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

Thomas Bayes
1701–1761

- 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

NUS | School of Computing
National University of Singapore

Probability: Example

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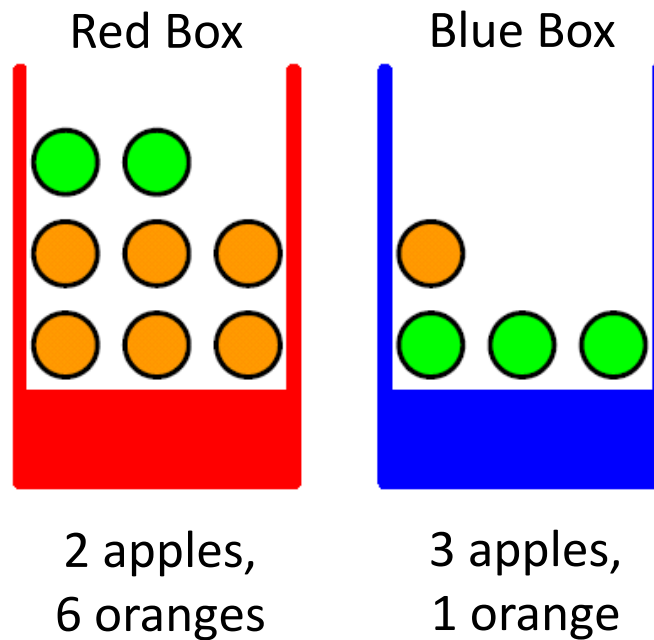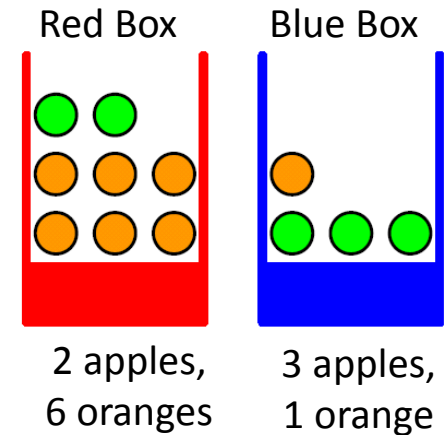
Red Box   Blue Box

2 apples,   3 apples,
6 oranges   1 orange

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

Probability: Example

Given:

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

Red Box   Blue Box

2 apples, 6 oranges   3 apples, 1 orange

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

- 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.

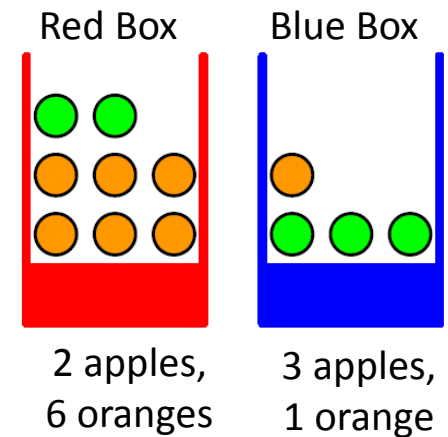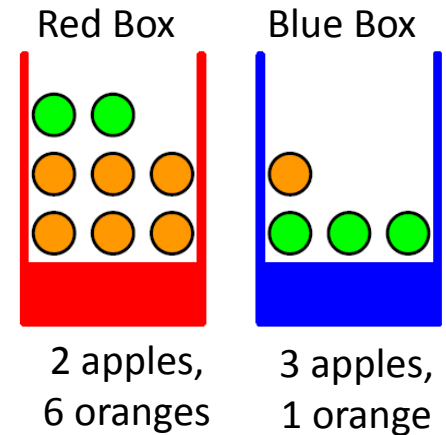b) Identify the color of the box if we observed that an orange has been selected.

Red Box   Blue Box

2 apples,   3 apples,
6 oranges   1 orange

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:

Red Box — 2 apples, 6 oranges

Blue Box — 3 apples, 1 orange

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

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

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

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!

Red Box   Blue Box
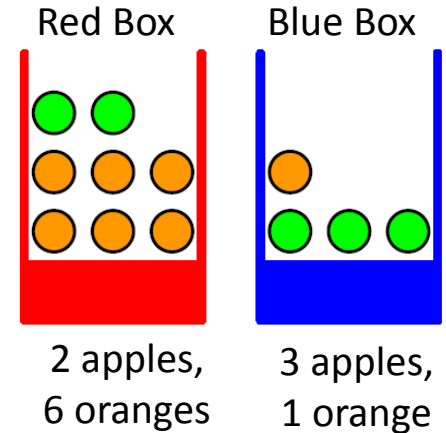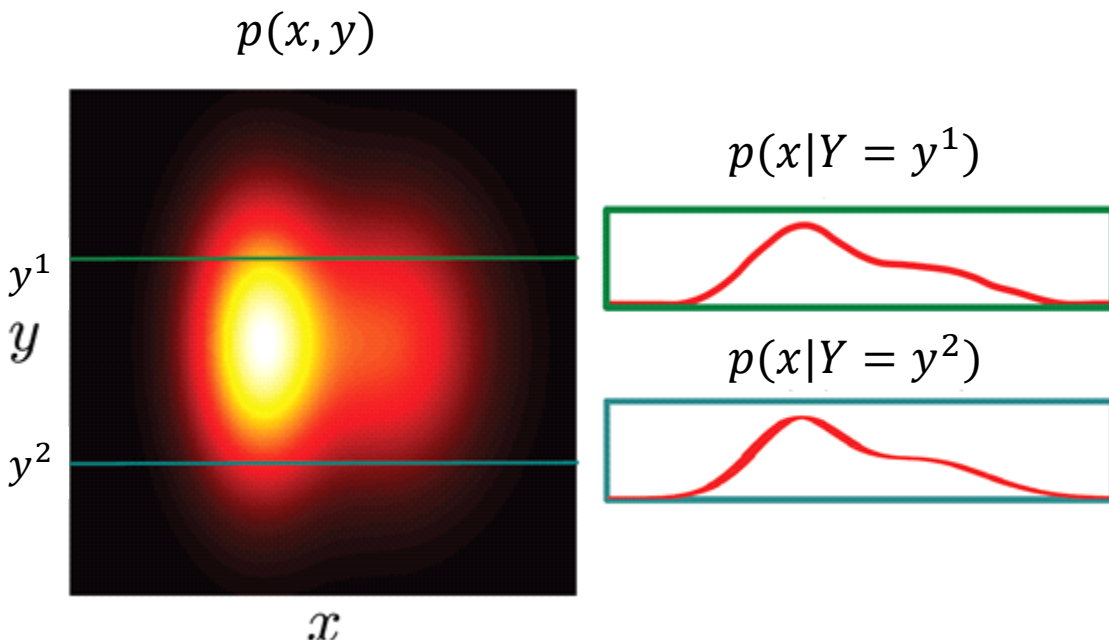
2 apples,   3 apples,
6 oranges   1 orange

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.

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

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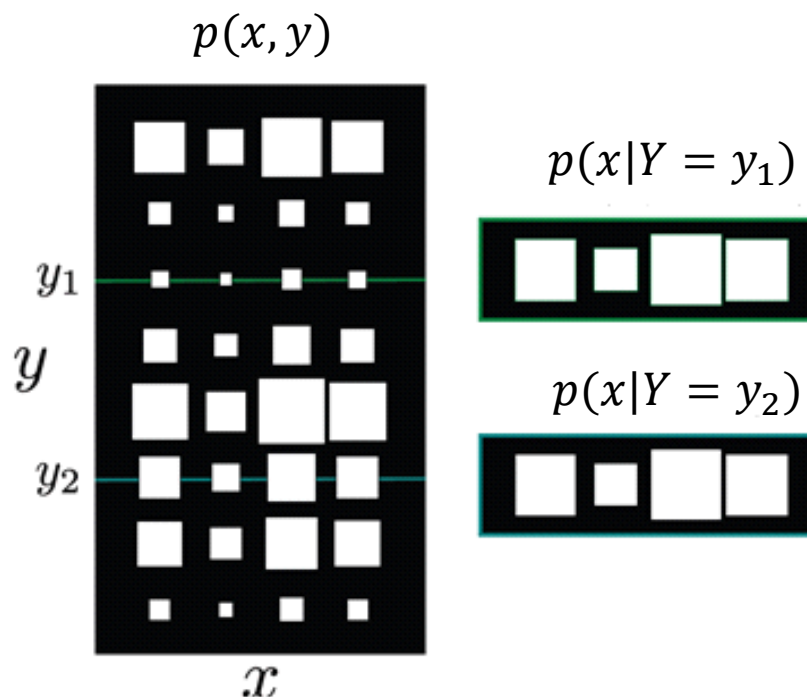
$p(x,y)$

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

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

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.

$p(x,y)$

$p(x|Y = y_1)$

$p(x|Y = y_2)$

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

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

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.

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

Probability: Rules of Expectation

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

$$E\big[f[x] + g[x]\big] = E\big[f[x]\big] + E\big[g[x]\big]$$
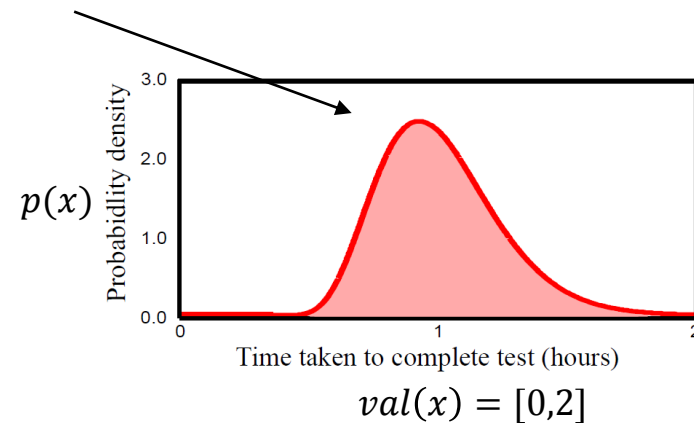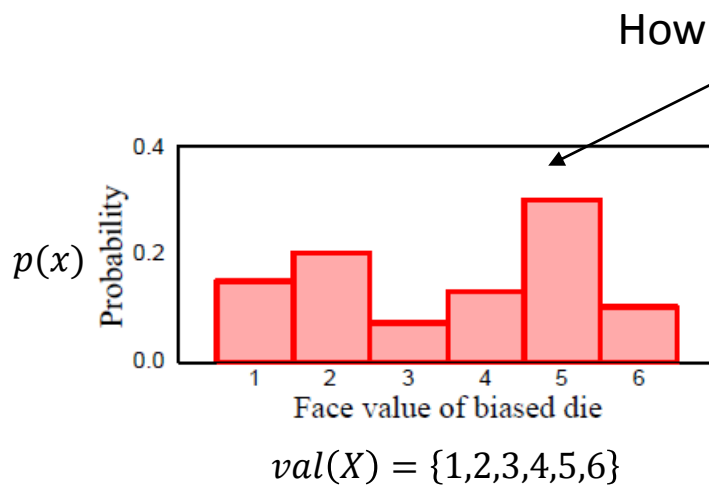
- **Rule 4**: Expectation of product of functions in variables X and Y is product of expectations of functions if X and Y are independent.

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

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 $p(x)$?"

How to get $p(x)$?

$p(x)$

Probability

0.4

0.2

0.0

1 2 3 4 5 6

Face value of biased die

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

$p(x)$

Probability density

3.0

2.0

1.0

0.0

0   1   2

Time taken to complete test (hours)

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

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

NUS National University of Singapore | School of Computing

Probability Distributions

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

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

Example:

Heights (in.) of American women (ages 30-39)

$\mu = 64$

$\sigma = 2$
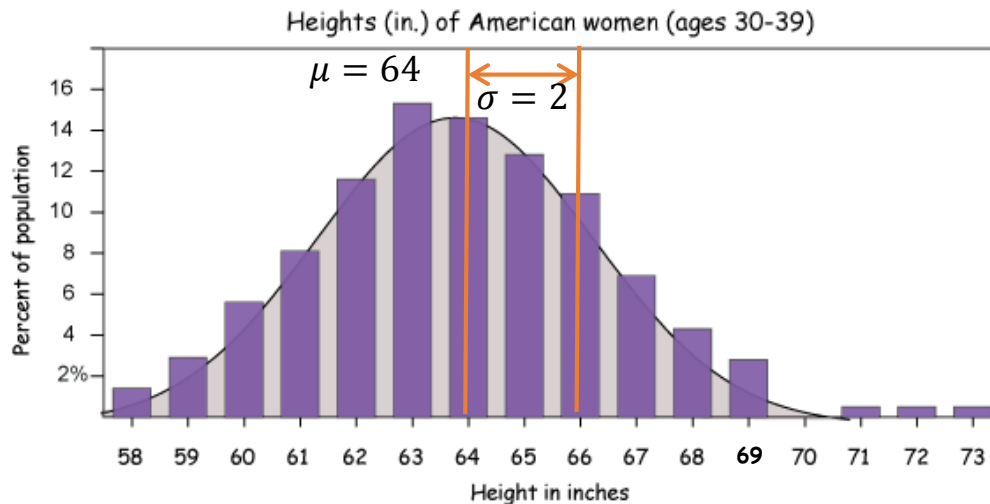
Fitting a Normal distribution to the heights of a population:

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

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

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

$$p(X = 0 \mid \lambda) \quad = \quad 1 - \lambda$$
$$p(X = 1 \mid \lambda) \quad = \quad \lambda$$

Jacob Bernoulli
1654–1705

Or

$$p(x) = \lambda^x (1 - \lambda)^{1-x} \, ,$$
$$p(x) = \text{Bern}_x[\lambda]$$

Example:

X is the outcome of flipping a coin, $X = 1$ represents 'heads', and $X = 0$ represents 'tails'.
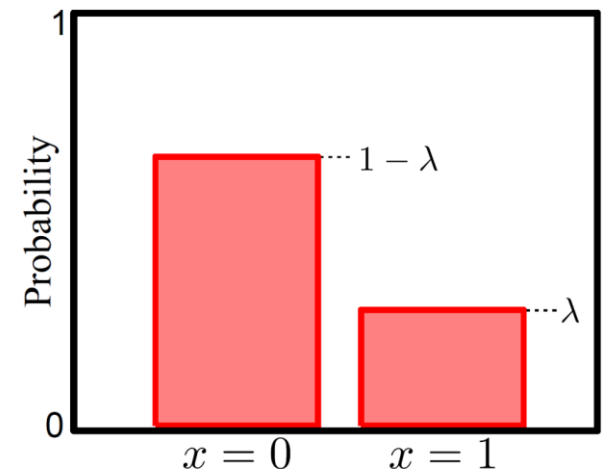
Images source: "Pattern Recognition and Machine Learning", Christopher Bishop
"Computer Vision: Models, Learning, and Inference", Simon Prince
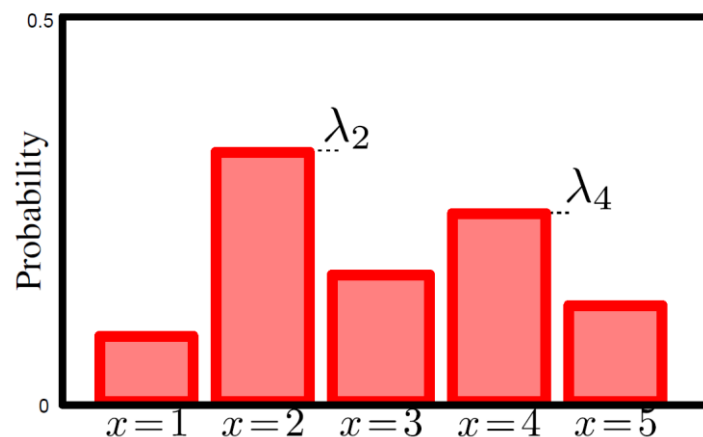
Categorical Distribution

- Discrete variables X that take on 1-of-K possible mutually exclusive states, e.g. a K-faced die.

- 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 $x = \mathbf{e}_3 = [0,0,1,0,0]^T$.

- K parameters $\lambda = [\lambda_1, \dots, \lambda_K]^T$, where $\lambda \geq 0$, $\sum_k \lambda_k = 1$.

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

Or

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

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

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

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

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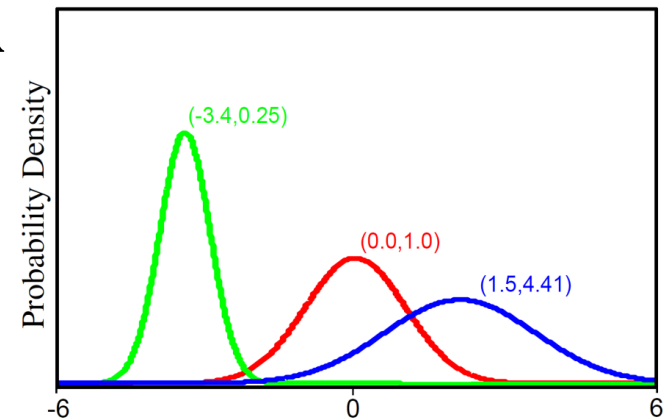
Probability Density

(-3.4,0.25)

(0.0,1.0)

(1.5,4.41)

-6   0   6

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

NUS School of Computing
National University of Singapore

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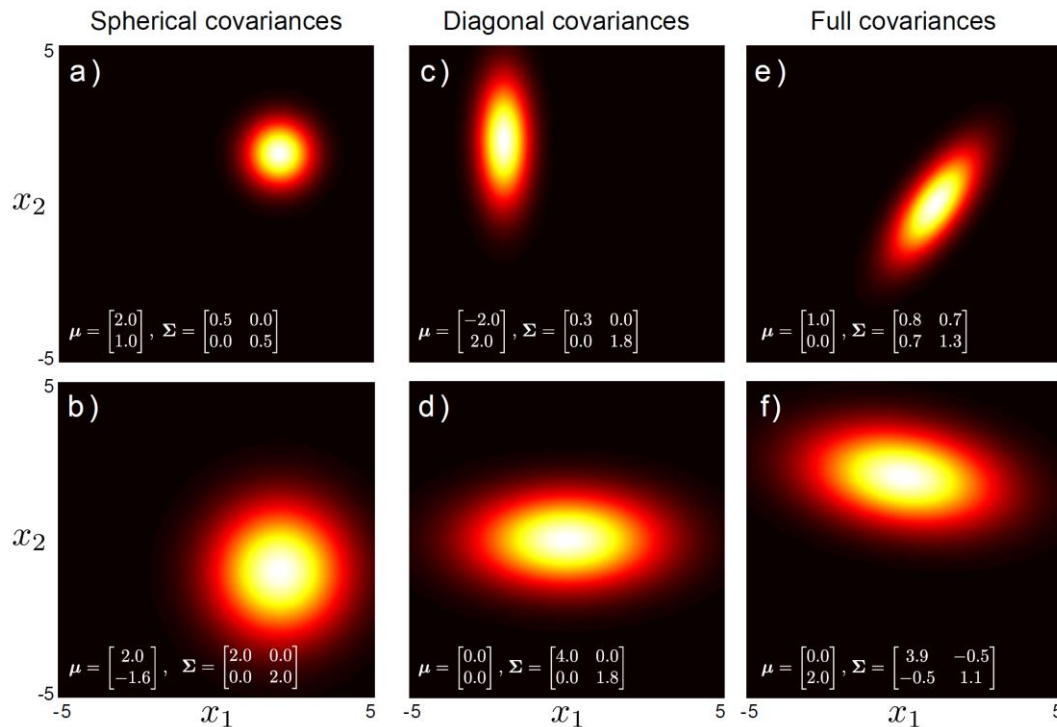
$$p(\boldsymbol{X} = \boldsymbol{a} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-0.5(\boldsymbol{a} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu})\}, \quad \boldsymbol{a} \in \mathbb{R}^D$$

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.

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

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 its 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:

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

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

i. 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

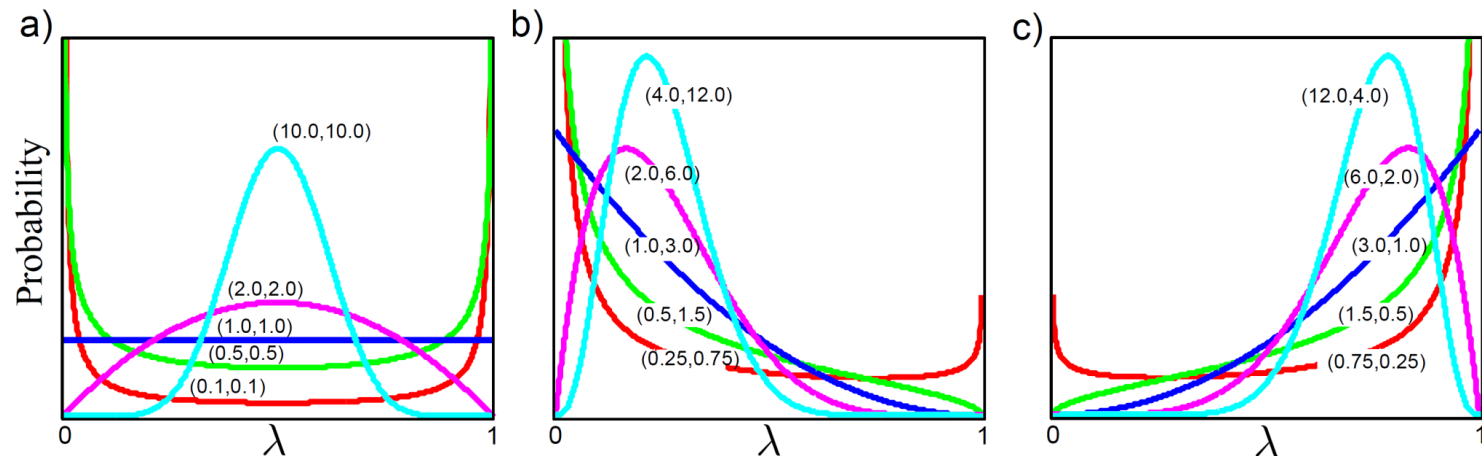# Importance of Conjugate Distributions

Proof:

$$p(x^*|x) = \frac{p(x^*, x)}{p(x)} \qquad \text{(Conditional probability)}$$

$$= \frac{\int p(x^*, x, \theta) d\theta}{p(x)} \qquad \text{(Marginal probability)}$$

$$= \frac{\int p(x^*, \theta|x) p(x) d\theta}{p(x)} \qquad \text{(Conditional probability)}$$

$$= \int p(x^*|x, \theta) p(\theta|x) d\theta \qquad \text{(Conditional probability)}$$

$$= \int p(x^*|\theta) p(\theta|x) d\theta \qquad \text{(Conditional Independence)}$$

Conjugate Distribution: Beta Distribution

- Conjugate distribution of Bernoulli distribution.

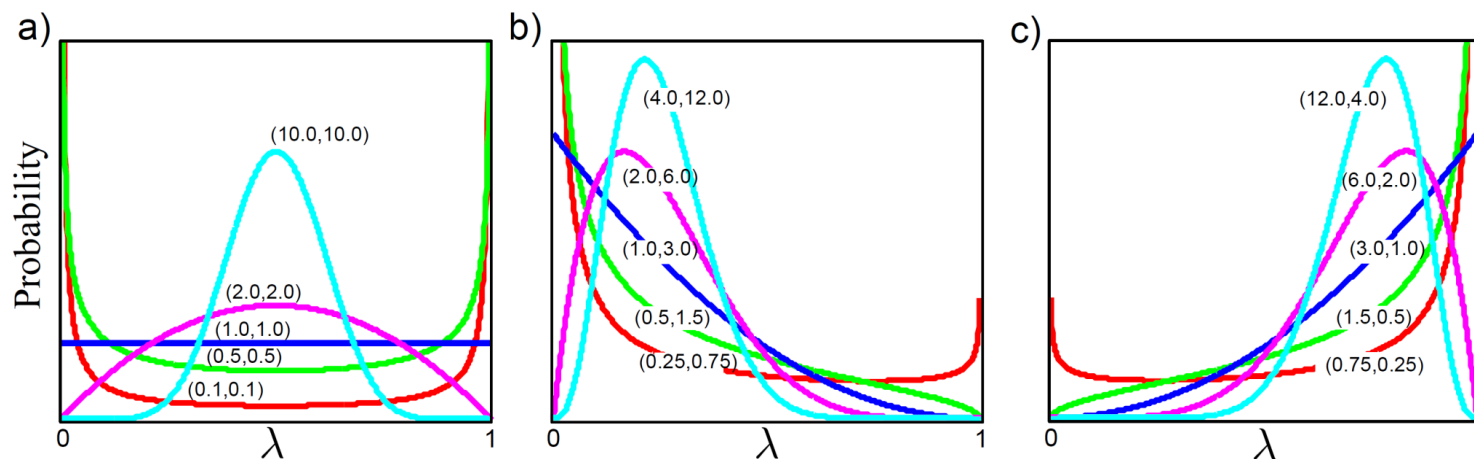- Defined over parameter of the Bernoulli distribution $\lambda \in [0,1]$.

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

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

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

Conjugate Distribution: Beta Distribution

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

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

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 $\alpha, \beta > 0$.

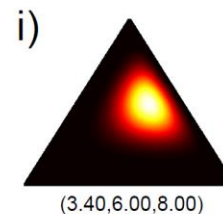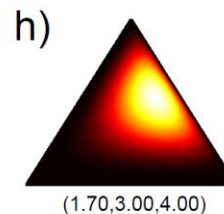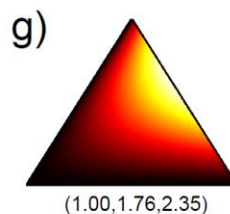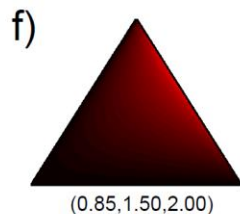Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Conjugate Distribution: Dirichlet Distribution

- Conjugate distribution of categorical distribution.

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

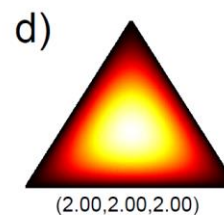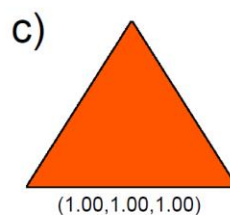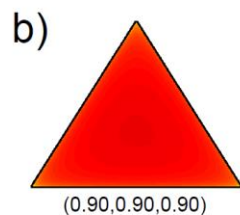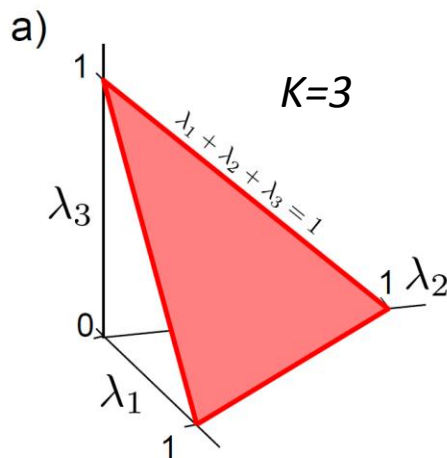$$p(\lambda_1, \ldots, \lambda_K) = \frac{\Gamma[\sum_{k=1}^K \alpha_k]}{\prod_{k=1}^K \Gamma[\alpha_k]} \prod_{k=1}^K \lambda_k^{\alpha_k - 1},$$

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

Peter Gustav Lejeune Dirichlet (1805-1859)
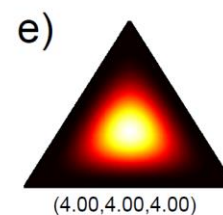
a) $K=3$, $\lambda_1 + \lambda_2 + \lambda_3 = 1$

b) (0.90,0.90,0.90)
c) (1.00,1.00,1.00)
d) (2.00,2.00,2.00)
e) (4.00,4.00,4.00)
f) (0.85,1.50,2.00)
g) (1.00,1.76,2.35)
h) (1.70,3.00,4.00)
i) (3.40,6.00,8.00)

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

NUS National University of Singapore | School of Computing

Conjugate Distribution: Dirichlet Distribution

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

$$p(\lambda_1, \ldots, \lambda_K) = \text{Dir}_{\lambda_{1 \ldots K}}[\alpha_1, \ldots \alpha_K]$$
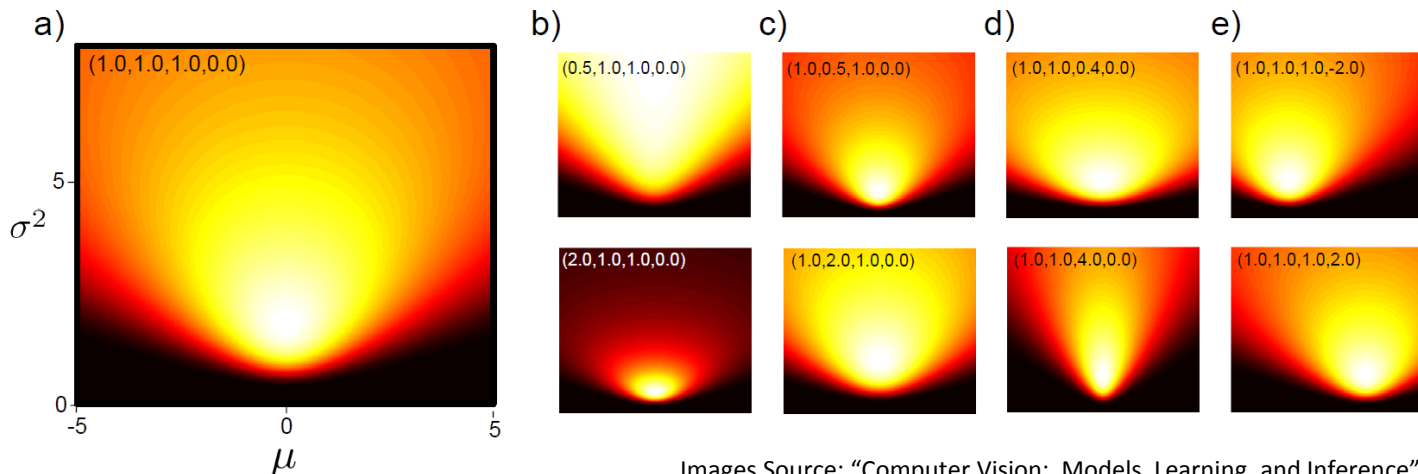
- *K* hyperparameters $\alpha_k > 0.$
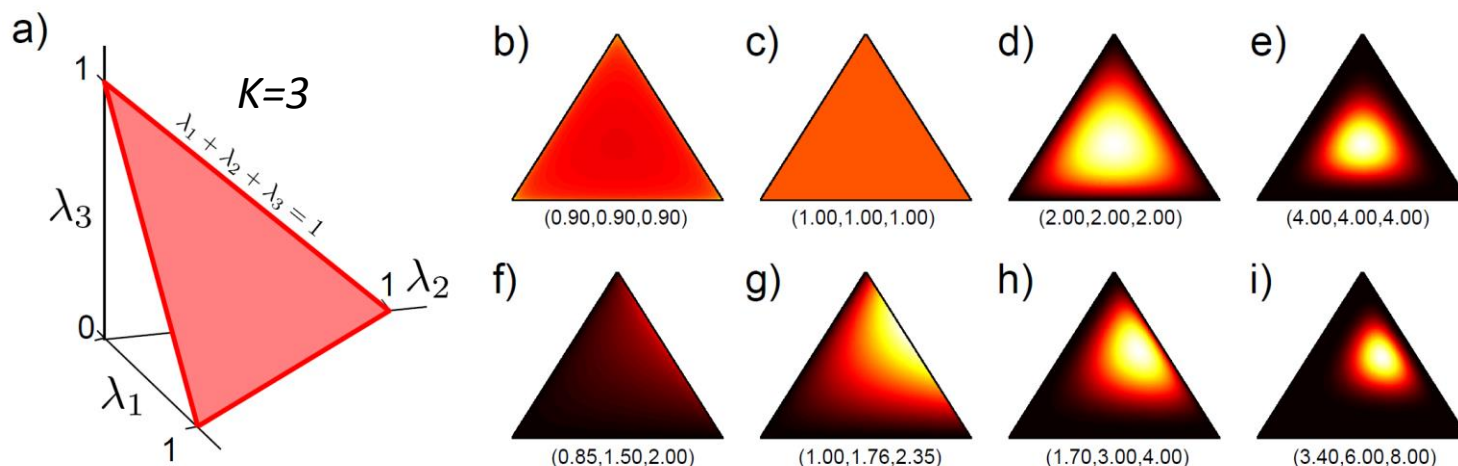
Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Conjugate Distribution:
Normal Inverse Gamma Distribution

- Conjugate distribution of univariate normal distribution.

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

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

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

a) (1.0,1.0,1.0,0.0)
b) (0.5,1.0,1.0,0.0)
c) (1.0,0.5,1.0,0.0)
d) (1.0,1.0,0.4,0.0)
e) (1.0,1.0,1.0,-2.0)
(2.0,1.0,1.0,0.0)
(1.0,2.0,1.0,0.0)
(1.0,1.0,4.0,0.0)
(1.0,1.0,1.0,2.0)

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

NUS National University of Singapore | School of Computing

Conjugate Distribution:
Normal Inverse Gamma Distribution

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

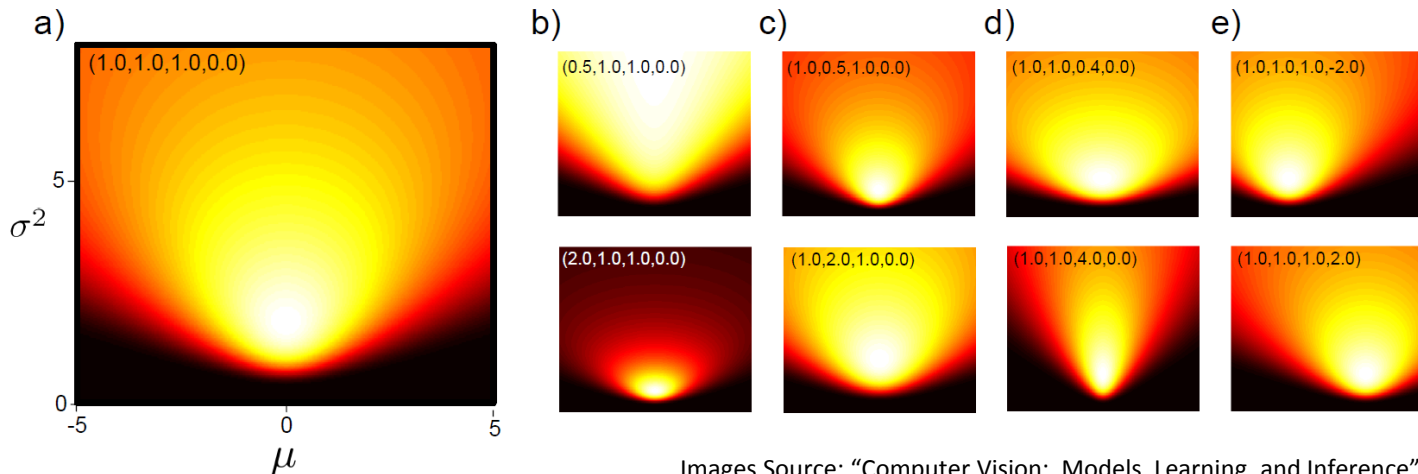$$p(\mu, \sigma^2) = \text{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

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

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

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.

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

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{NorIWis}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}[\alpha, \boldsymbol{\Psi}, \gamma, \boldsymbol{\delta}]$$
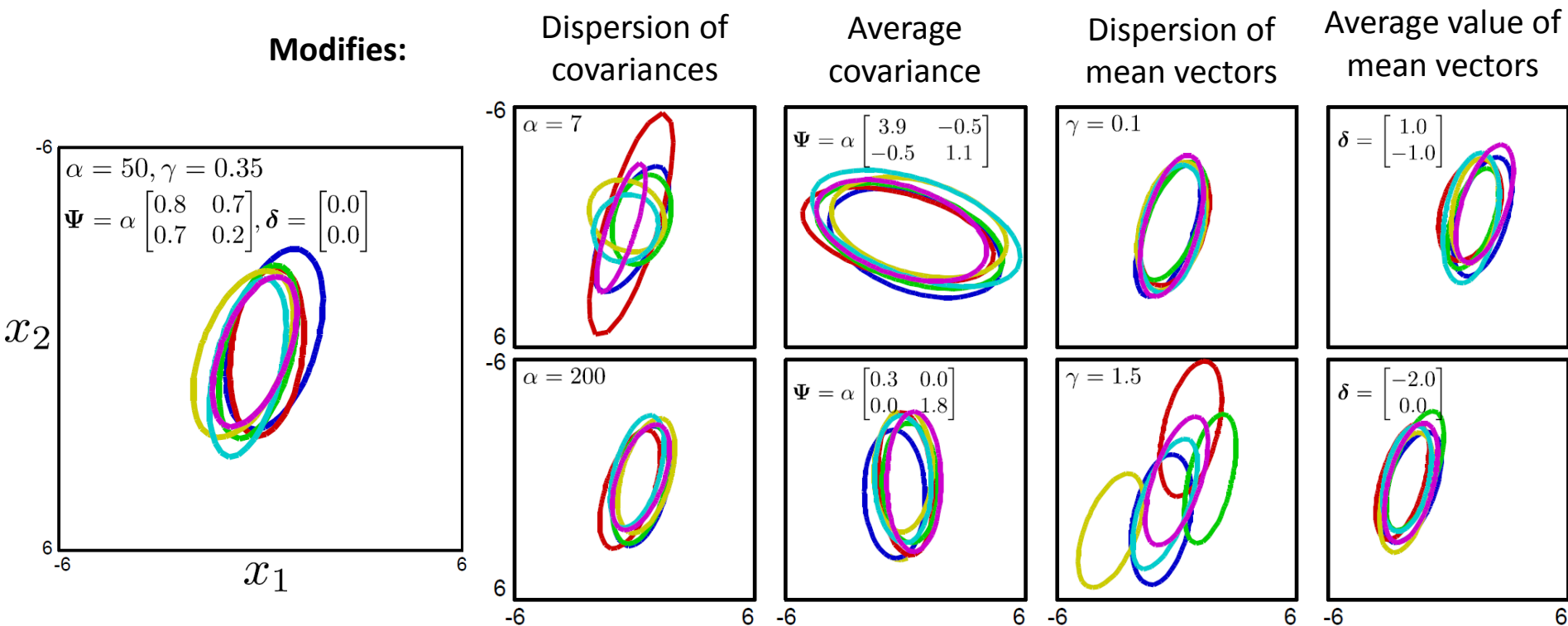
- **Four hyperparameters**: a positive scalar α, a positive definite matrix $\boldsymbol{\Psi} \in \mathbb{R}_+^{D\times D}$, a positive scalar γ, and a vector $\boldsymbol{\delta} \in \mathbb{R}^D$.

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

Example 1: Conjugate Distribution

Find: The posterior distribution of the parameter (μ, σ) from a univariate Gaussian distribution.

Solution: Using Bayes' rule:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta) p(\theta)}{p(x)} = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta) p(\theta)}{\int \prod_{i=1}^{N} p(x[i] \mid \theta) p(\theta) \, d\theta}$$

where:

$$\prod_{i=1}^{N} p(x[i] \mid \theta) p(\theta) = \prod_{i=1}^{N} \text{Norm}_{x[i]}[\mu, \sigma^2] \, \text{NormInvGam}_{\mu, \sigma^2}[\alpha, \beta, \gamma, \delta]$$

Example 1: Conjugate Distribution

We have

$$\prod_{i=1}^{N} p(x[i]|\theta)p(\theta) = \prod_{i=1}^{N} \text{Norm}_{x[i]}[\mu, \sigma^2]\, \text{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

Rearranging:

$$\prod_{i=1}^{N} p(x[i]|\theta)p(\theta) = \underbrace{\kappa[\alpha, \beta, \gamma, \delta, x]}_{\text{Constant}}\text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}]$$

Constant

where

$$\tilde{\alpha} = \alpha + \frac{N}{2},\qquad\qquad \tilde{\delta} = \frac{(\gamma\delta + \sum_i x[i])}{\gamma + N},$$

$$\tilde{\gamma} = \gamma + N,\qquad\qquad \tilde{\beta} = \frac{\sum_i x[i]^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\gamma\delta + \sum_i x[i])^2}{2(\gamma + N)}.$$

Example 1: Conjugate Distribution

Putting into the Bayes' rule, we get:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i]|\theta)p(\theta)}{p(x)} = \frac{\prod_{i=1}^{N} p(x[i]|\theta)p(\theta)}{\int \prod_{i=1}^{N} p(x[i]|\theta)p(\theta)\, d\theta}$$

$$p(\theta|x) = \frac{\cancel{\kappa[\alpha,\beta,\gamma,\delta,x]}\text{NormInvGam}_{\mu,\sigma^2}\big[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}\big]}{\underbrace{\cancel{\kappa[\alpha,\beta,\gamma,\delta,x]}\int\int \text{NormInvGam}_{\mu,\sigma^2}\big[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}\big]d\mu d\sigma^2}_{=\,1}}$$

$$\boxed{p(\theta|x) = \text{NormInvGam}_{\mu,\sigma^2}\big[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}\big]}$$

Same form as conjugate prior,
i.e., Normalized Inverse Gamma!

Example 2: Conjugate Distribution

Find: $p(x^*|\boldsymbol{x})$ for a univariate Gaussian.

Solution:

$$p(x^*|x) = \int \int p(x^*|\mu, \sigma^2) p(\mu, \sigma^2|x) d\mu d\sigma^2$$

$$= \int \int \text{Norm}_{x^*}[\mu, \sigma^2] \text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}] d\mu d\sigma^2$$

$$= \kappa[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}, x^*] \underbrace{\int \int \text{NormInvGam}_{\mu,\sigma^2}[\breve{\alpha}, \breve{\beta}, \breve{\gamma}, \breve{\delta}] d\mu d\sigma^2}_{= 1}$$

$$= \kappa[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}, x^*]$$

This is a constant (function of $\{x^*, x[1], \dots, x[N]\}$)!

Example 2: Conjugate Distribution

This is a constant (function of $\{x^*, x[1], \dots, x[N]\}$)!

$$p(x^*|x) = \kappa\left[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}, x^*\right] = \frac{1}{\sqrt{2\pi}} \frac{\tilde{\beta}\tilde{\alpha}\sqrt{\tilde{\gamma}}}{\breve{\beta}\breve{\alpha}\sqrt{\breve{\gamma}}} \frac{\Gamma[\breve{\alpha}]}{\Gamma[\tilde{\alpha}]}$$

where

$$\breve{\alpha} = \tilde{\alpha} + 1/2, \qquad \breve{\gamma} = \tilde{\gamma} + 1$$

$$\breve{\beta} = \frac{x^{*2}}{2} + \tilde{\beta} + \frac{\tilde{\gamma}\tilde{\delta}^2}{2} - \frac{(\tilde{\gamma}\tilde{\delta} + x^*)^2}{2(\tilde{\gamma} + 1)}.$$

Summary

You have learned how 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.