Introduction Probabilistic Learning & Models

Generative and Deep Learning (GDL)

Davide Bacciu (davide.bacciu@unipi.it)

Università di Pisa

Lecture Outline

Introduction to the module

Probabilistic models and their graphical formalism

Probability refresher

Inference in probabilistic learning models

Introduction

Probabilistic learning models

- ML models that represent knowledge inferred from data under the form of probabilities
  - Probabilities can be sampled: new data can be generated
  - Supervised, unsupervised, weakly supervised learning tasks
  - Incorporate prior knowledge on data and tasks
  - Interpretable knowledge (how data is generated)
- Most of the modern task comprises large numbers of variables
  - Modeling the joint distribution of all variables can become impractical
  - Exponential size of the parameter space
  - Computationally impractical to train and predict

Graphical Models Framework

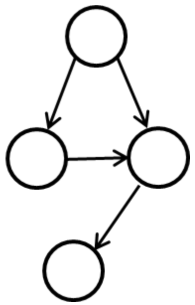
- Representation
  - Graphical models are a compact way to represent exponentially large probability distributions
  - Encode conditional independence assumptions
  - Different classes of graph structures imply different assumptions/capabilities
- Inference
  - How to query (predict with) a graphical model?
  - Probability of unknown X given observations $\boldsymbol{d}$, $P(X|\boldsymbol{d})$
  - Most likely hypothesis
- Learning
  - Find the right model parameter
  - An inference problem after all

Graphical Model Representation

A graph whose **nodes** (vertices) are **random variables** whose **edges** (links) represent **probabilistic relationships** between the variables

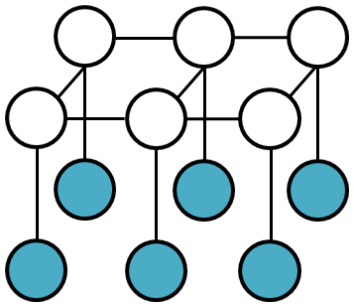
Different classes of graphs

Directed Models



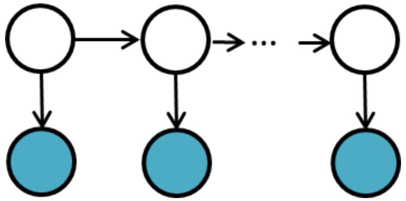
Directed edges express causal relationships

Undirected Models



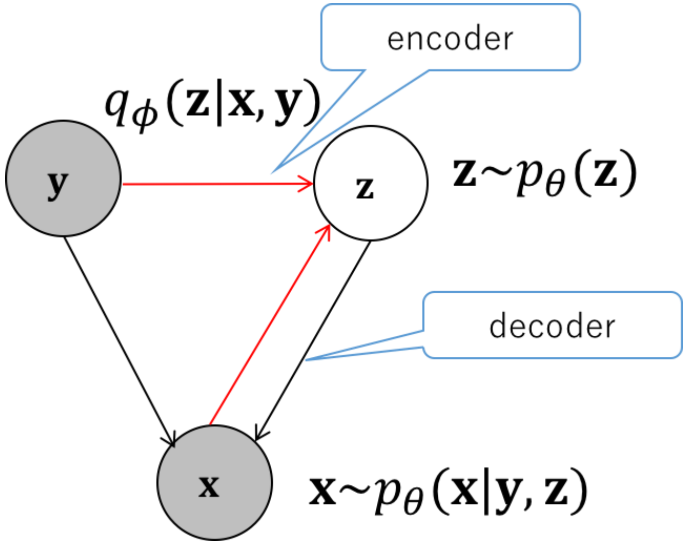
Undirected edges express soft constraints

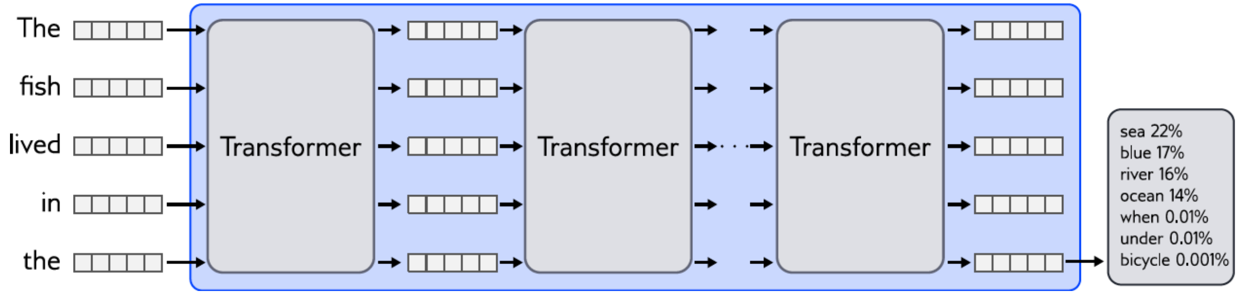
Dynamic Models



Structure changes to reflect dynamic processes

Probabilistic Models in Deep Learning





Probabilistic (generative) learning necessary to understand Generative Deep Learning

Module I - Fundamentals of probabilistic models and causality

Lesson 1	Introduction: Probabilistic Learning & Models
Lesson 2	Graphical models: representation
Lesson 3	Graphical models: Markov properties
Lesson 4	Graphical causal models
Lesson 5	Structure learning

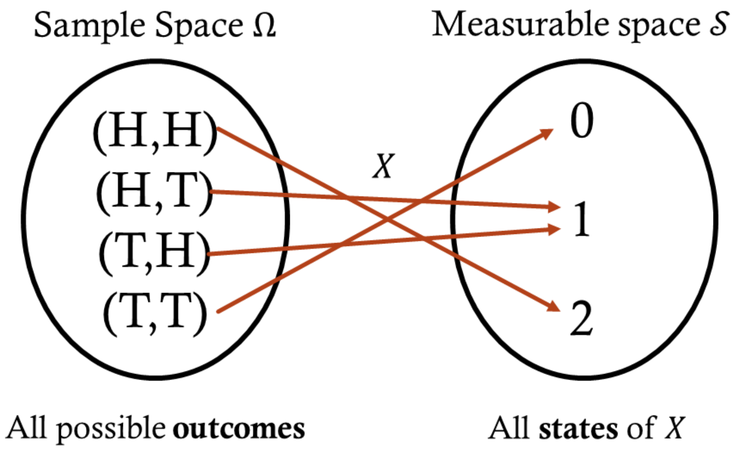
By Riccardo Massidda (Lessons 2–5)

Module content is fully covered by David Barber's book

Probability Refresher

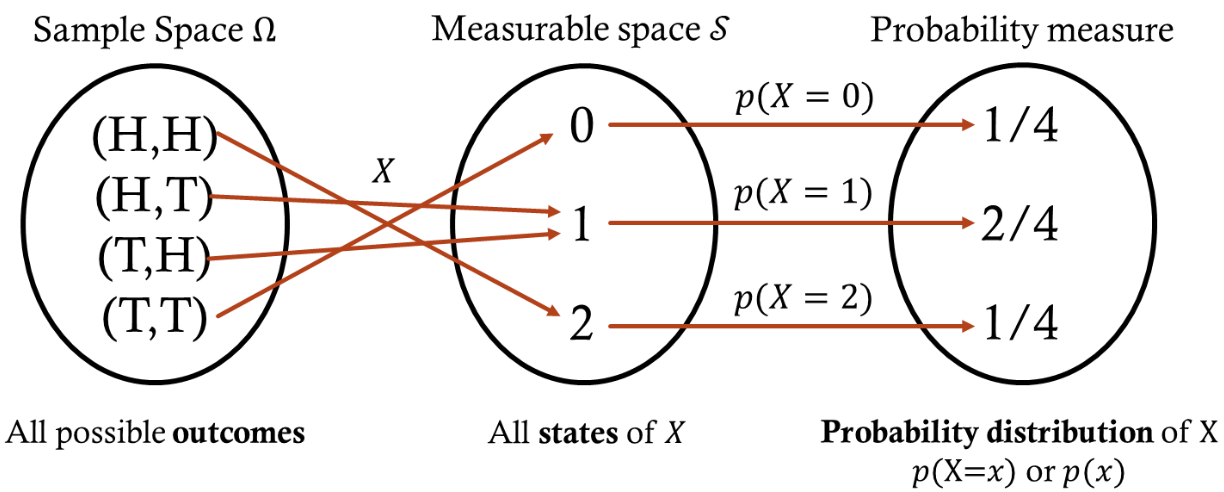
Random Variables

- A Random Variable (RV) is a function describing the outcome of a random process by assigning unique values to all possible outcomes of the experiment
- A RV models an attribute of our data (e.g. age, speech sample,...)
- Use uppercase to denote a RV, e.g. X, and lowercase to denote a value (observation), e.g. x
- A discrete (categorical) RV is defined on a finite or countable list of values
- A continuous RV can take infinitely many values



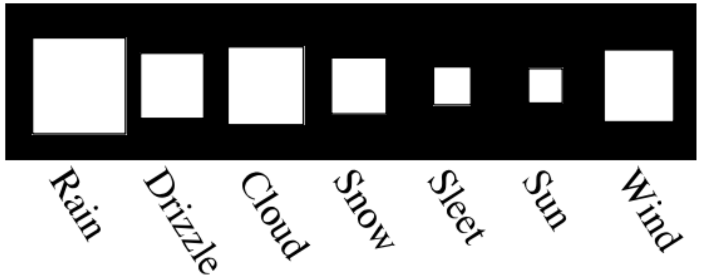
Random Variables

- A Random Variable (RV) is a function describing the outcome of a random process by assigning unique values to all possible outcomes of the experiment
- A RV models an attribute of our data (e.g. age, speech sample,…)
- Use uppercase to denote a RV, e.g. X, and lowercase to denote a value (observation), e.g. x
- A discrete (categorical) RV is defined on a finite or countable list of values
- A continuous RV can take infinitely many values

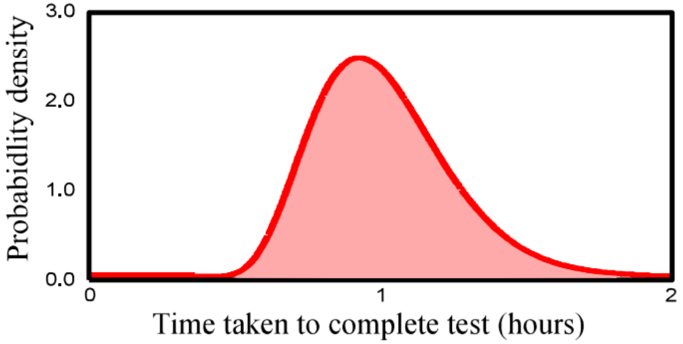


Probability Functions

Hinton diagram of a discrete RV



PDF of a continuous RV



- Discrete Random Variables
  - A probability function $P(X = x) \in [0,1]$ measures the probability of a RV X attaining the value x
  - Subject to sum-rule $\sum_{x \in \Omega} P(X = x) = 1$
- Continuous Random Variables
  - A density function $p(t)$ describes the relative likelihood of a RV to take on a value t
  - Subject to sum-rule $\int_{\Omega}^{t} p(t)dt = 1$
  - Defines a probability distribution, e.g. $P(X \leq x) = \int_{-\infty}^{x} p(t)dt$
- Shorthand $P(x)$ for $P(X = x)$ or $P(X \leq x)$

Joint and Conditional Probabilities

If a discrete random process is described by a set of RVs $X_1, \ldots, X_N$, then the joint probability writes

$$P(X_1 = x_1, \ldots, X_N = x_n) = P(x_1 \wedge \cdots \wedge x_n)$$

The joint conditional probability of $x_1, \ldots, x_n$ given y

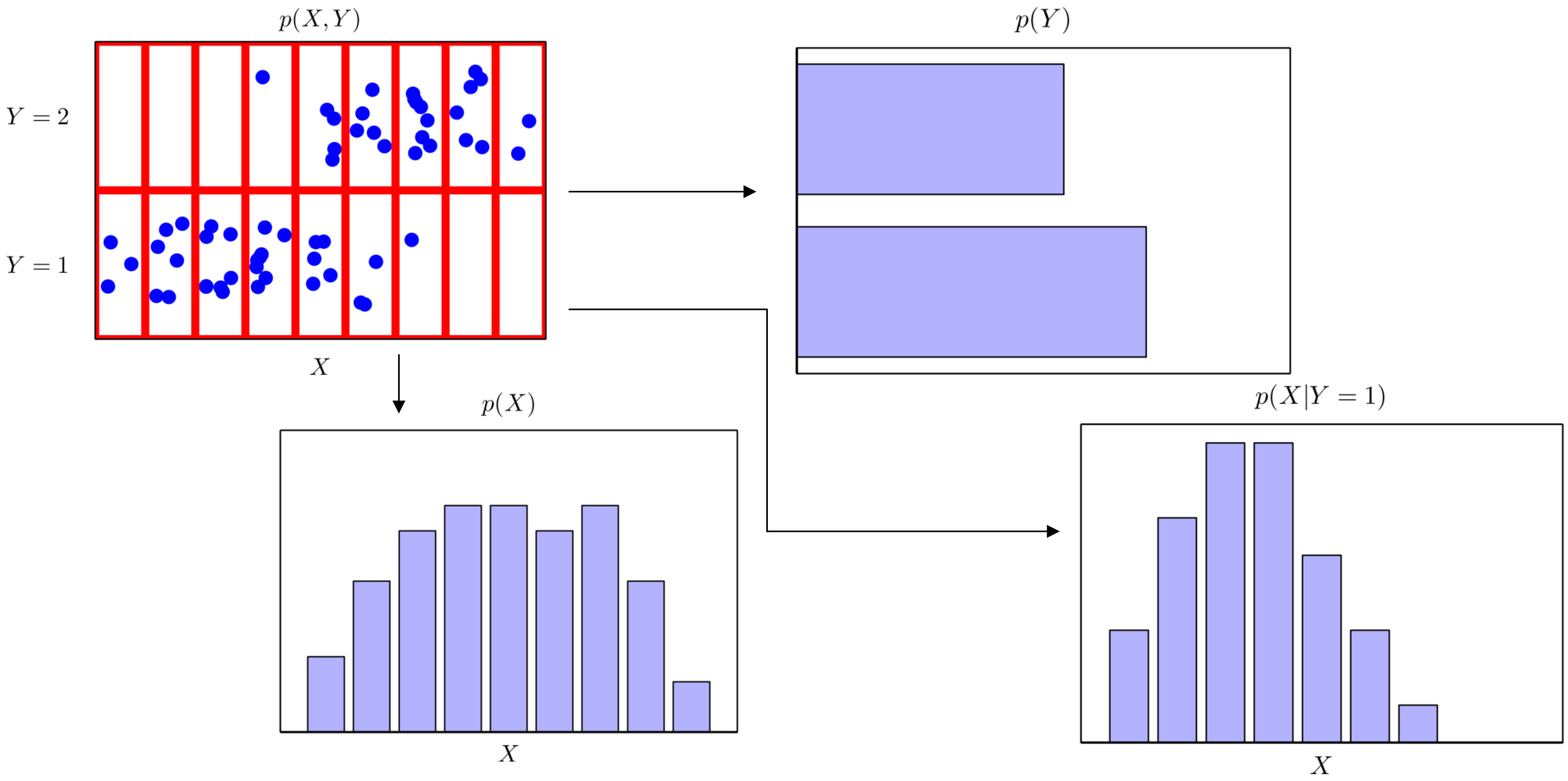
$$P(x_1, \ldots, x_n | y)$$

measures the effect of the realization of an event y on the occurrence of $x_1, \ldots, x_n$

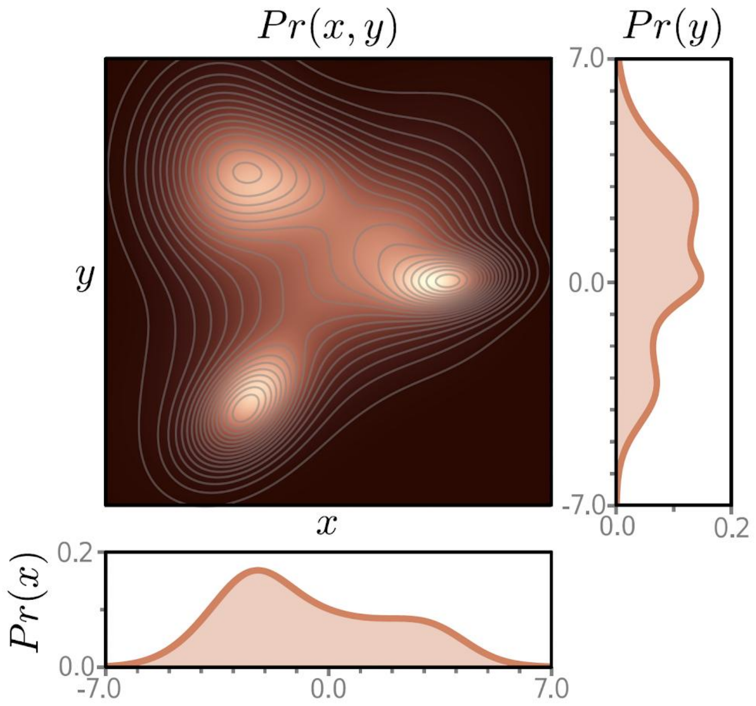
A conditional distribution $P(x|y)$ is actually a family of distributions

- For each y, there is a distribution $P(x|y)$

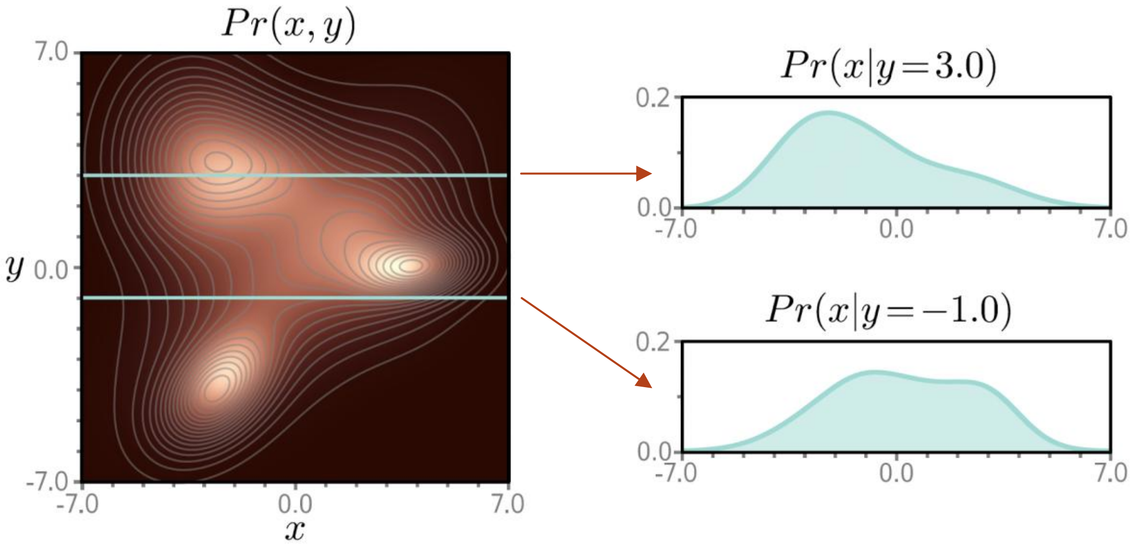
Probabilities Visually



Continuous Distributions



Joint and marginal distributions



Joint and conditional distributions

Chain Rule

Definition (Product Rule a.k.a. Chain Rule)

$$P(x_1, \dots, x_i, \dots, x_n | y) = \prod_{i=1}^{N} P(x_i \,\middle|\, x_1, \dots, x_{i-1}, y)$$

Definition (Marginalization)

Using the sum and product rules together yield to the complete probability

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1 | X_2 = x_2) P(X_2 = x_2)$$

Bayes Rule (a ML interpretation)

Given hypothesis $h_i \in H$ and observations $\boldsymbol{d}$

$$P(h_i|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|h_i)P(h_i)}{P(\boldsymbol{d})} = \frac{P(\boldsymbol{d}|h_i)P(h_i)}{\sum_j P(\boldsymbol{d}|h_j)P(h_j)}$$

- $P(h_i)$ is the prior probability of h_i
- $P(\boldsymbol{d}|h_i)$ is the conditional probability of observing $\boldsymbol{d}$ given that hypothesis h_i is true (likelihood).
- $P(\boldsymbol{d})$ is the marginal probability of $\boldsymbol{d}$
- $P(h_i|\boldsymbol{d})$ is the posterior probability that hypothesis is true given the data and the previous belief about the hypothesis

Independence and Conditional Independence

- Two RV X and Y are independent if knowledge about X does not change the uncertainty about Y and vice versa

$$I(X,Y) \Leftrightarrow P(X,Y) = P(X|Y)P(Y) \\ = P(Y|X)P(X) = P(X)P(Y)$$

- Two RV X and Y are conditionally independent given Z if the realization of X and Y is an independent event of their conditional probability distribution given Z

$$I(X,Y|Z) \Leftrightarrow P(X,Y|Z) = P(X|Y,Z)P(Y|Z) \\ = P(Y|X,Z)P(X|Z) = P(X|Z)P(Y|Z)$$

- Shorthand $X \perp Y$ for $I(X,Y)$ and $X \perp Y|Z$ for $I(X,Y|Z)$

Expectation

Discrete RV X with n possible realizations:

$$\mathbb{E}_{x \sim p(X)}[f(x)] = \sum_{i=0}^{n} p(x_i) f(x_i)$$

The average value of a function f after considering the probability of seeing different values of x

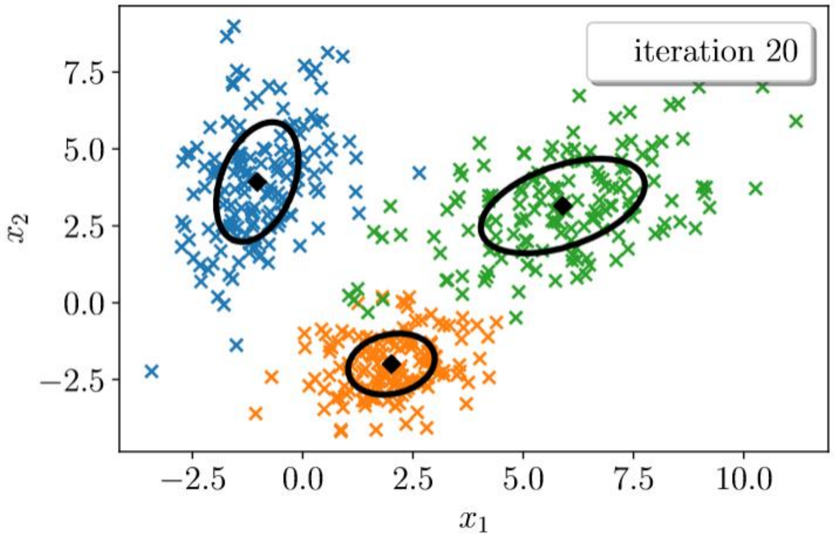
❖ If n is finite, the expectation can be computed in **closed form** ☺

Continuous RV X:

$$\mathbb{E}_{x \sim p(X)}[f(x)] = \int p(x) f(x) dx$$

❖ If an analytical solution does not exist, **we need approximations** ☹

The resulting value depends on the codomain of f



More on Expectation

- Easily generalizes to multivariate cases

$$\mathbb{E}_{x,y\sim p(X,Y)}[f(x,y)] = \int\int p(x,y)f(x)dxdy$$

- Expectation is a linear operator

$$\mathbb{E}_x[k] = k$$

$$\mathbb{E}_x[k \cdot f(x)] = k\mathbb{E}_x[f(x)]$$

$$\mathbb{E}_x[f(x) + g(x)] = \mathbb{E}_x[f(x)] + \mathbb{E}_x[g(x)]$$

$$\mathbb{E}_{x,y}[f(x) \cdot g(y)] = \mathbb{E}_x[f(x)] \cdot \mathbb{E}_y[g(y)] \text{ (if } x, y \text{ independent)}$$

...

Common Probability Distributions

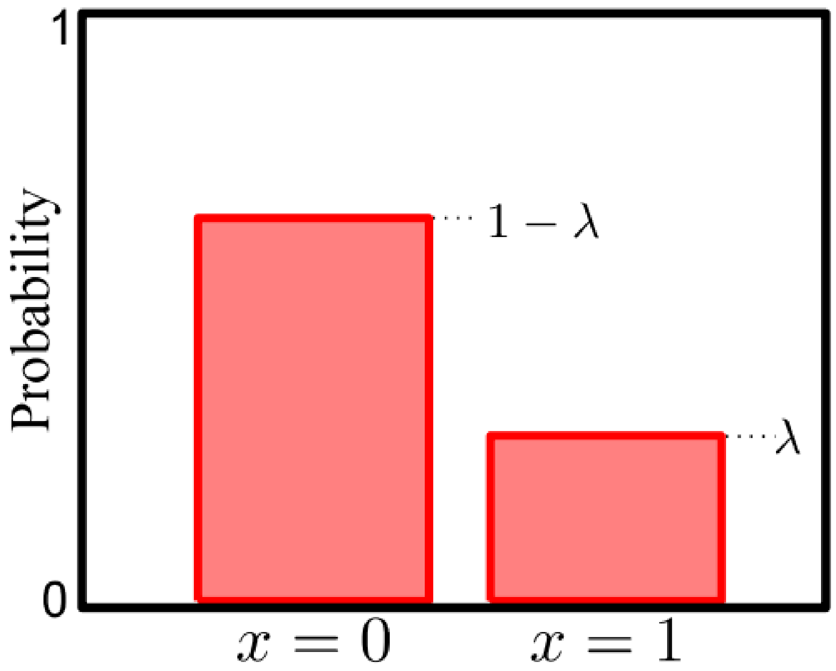
Many distributions, depending on data domain

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

Discrete distribution with two possible outcomes $x \in \{0,1\}$, governed by parameter λ, i.e. the probability of success $P(x = 1) = \lambda$

Has a matching Binomial distribution for measuring number of success in N samples/trials

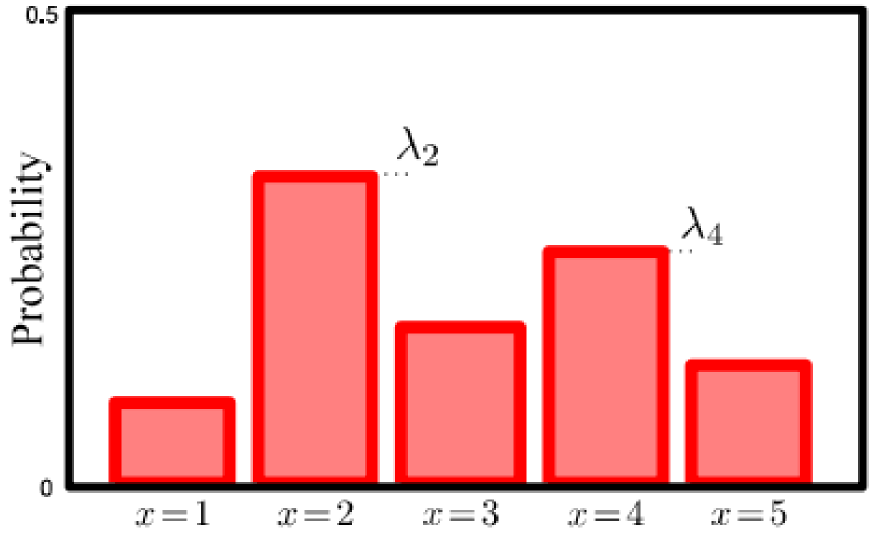


Categorical

Discrete distribution determining the probability of observing K possible outcomes $x \in \{1, .., K\}$

Governed by parameters $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_K]$, where $\mathrm{P(x = k)} = \lambda_k$

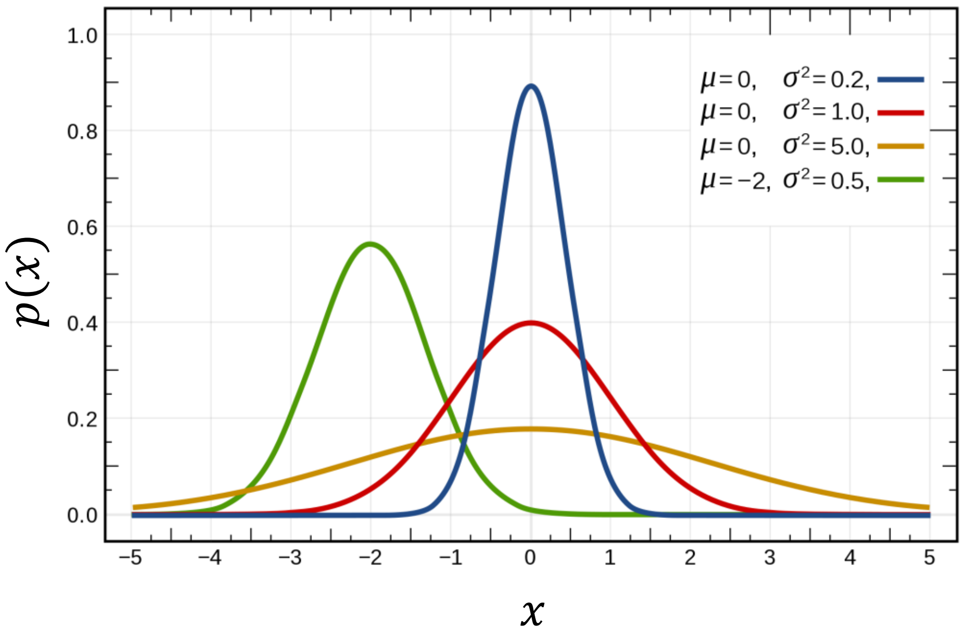
Has a matching Multinomial distribution for the counts of categories in N samples/trials



Univariate Gaussian (Normal)

Probability that an event has a continuous value x written as $\mathcal{N}(x \mid \mu, \sigma^2)$

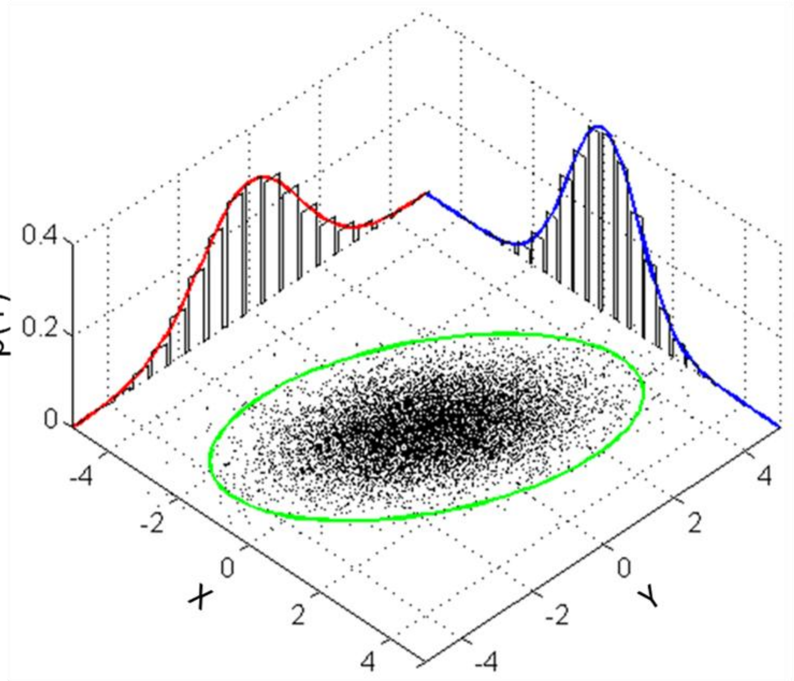
$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \left(\frac{(x-\mu)^2}{2\sigma^2} \right)$$

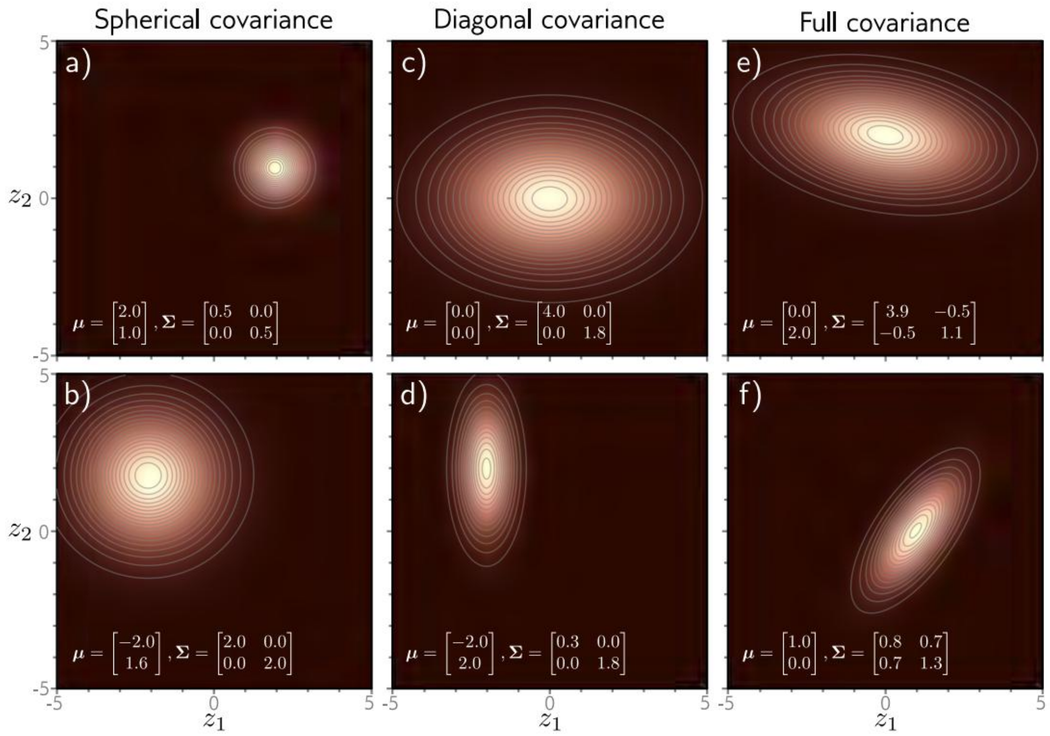


Multivariate Gaussian

Probability that an event has a continuous vector $\boldsymbol{x}$ written as $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

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



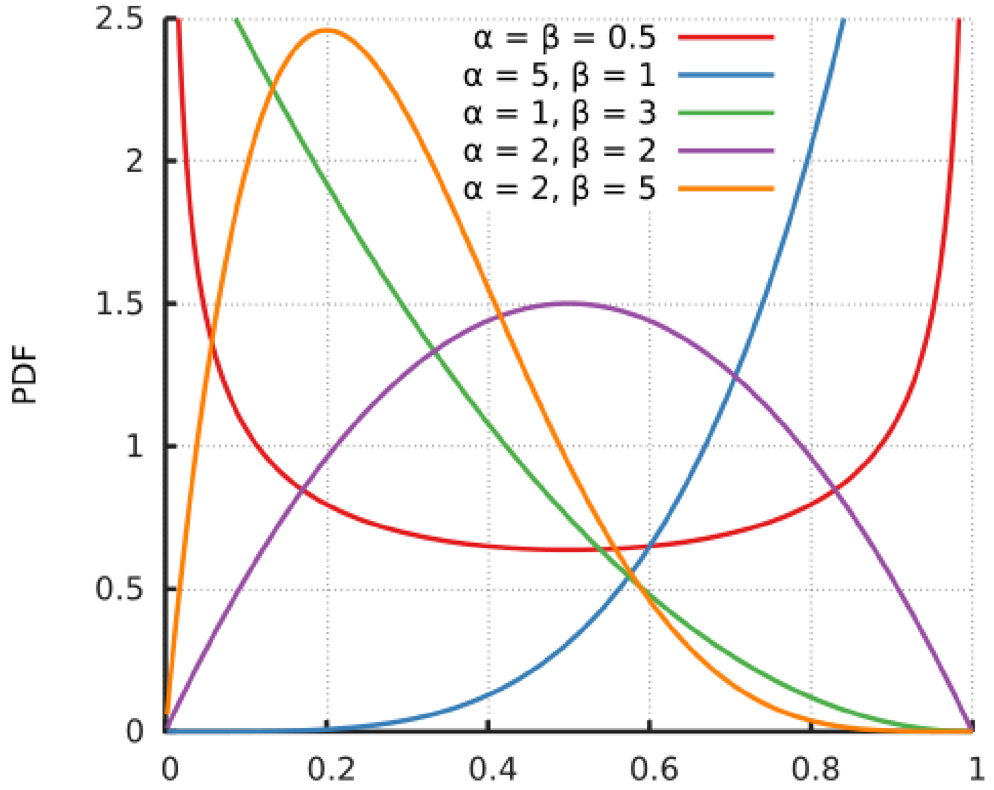


Beta

Continuous univariate distribution on $x \in [0,1]$ governed by two parameters $\alpha, \beta \in [0, \infty]$

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

Normalizing factor depending on the $\Gamma[\cdot]$ function (closely related to factorials)



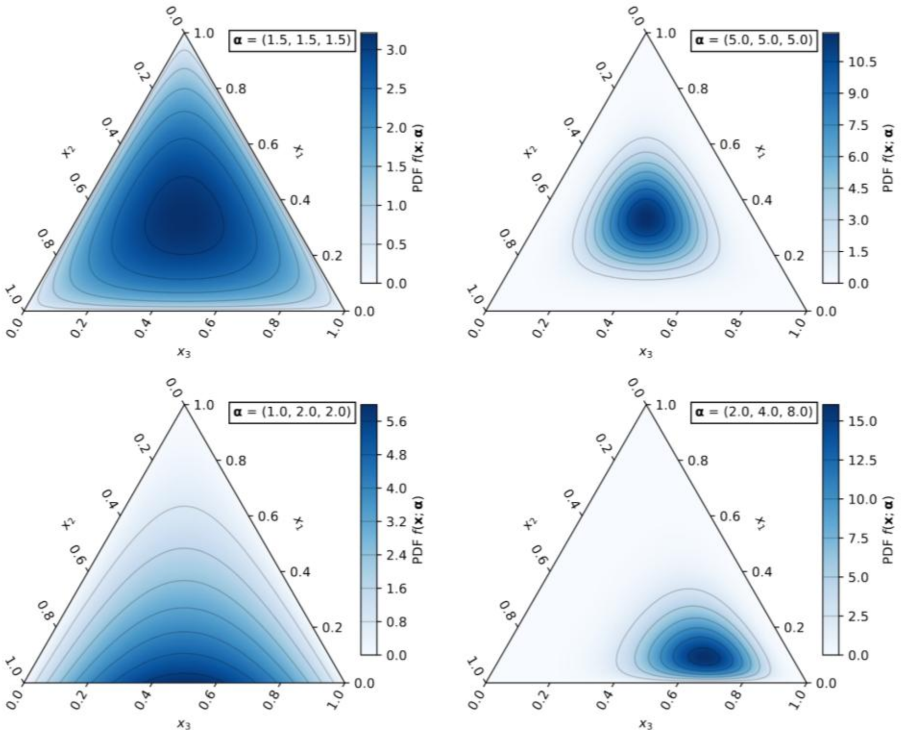
Dirichlet

Continuous multivariate distribution on $\mathbf{x} = [x_1, \dots, x_K]$ where $x_k \in [0,1]$ and $\sum_{k=1}^{K} x_k = 1$

Governed by k parameters $\alpha_1, \dots, \alpha_k$

$$p(\boldsymbol{x}) = \frac{\prod_{k=1}^{k} x_k^{\alpha_k - 1}}{B(\alpha_1, \dots, \alpha_k)}$$

Again a normalizing factor depending on the $\Gamma[\cdot]$ function

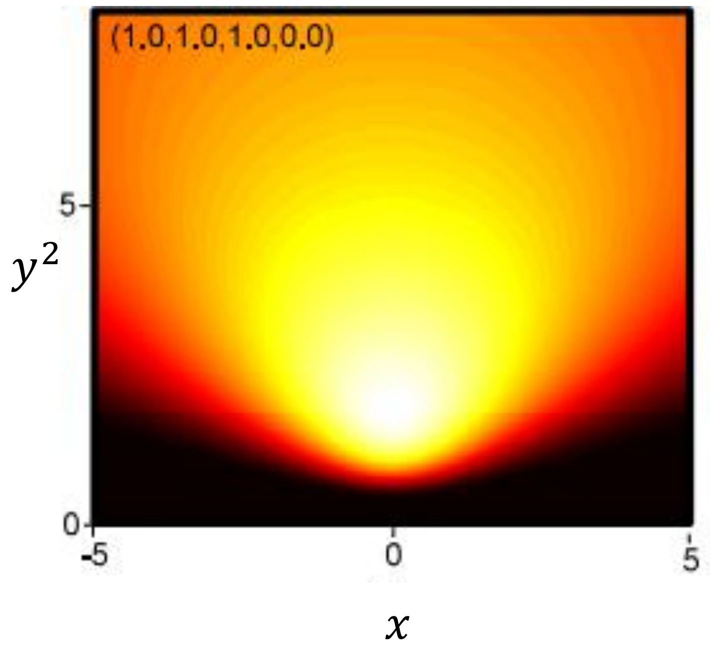


Normal-Scaled Inverse Gamma

Continuous bivariate distribution on x, y with $x \in \mathbb{R}$ and $x \in \mathbb{R}^+$, governed by four parameters $(\alpha, \beta, \gamma, \mu_0)$

$$P(x, y) = \frac{\sqrt{\gamma}}{y\sqrt{2\pi}\Gamma[\alpha]} \left(\frac{1}{y^2}\right)^{\alpha+1} \exp - \left(\frac{2\beta + \gamma(\mu_0 - x)^2}{2y^2}\right)$$

Generalizes to $\boldsymbol{x}$ being a vector and $\boldsymbol{y}$ a positive definite matrix

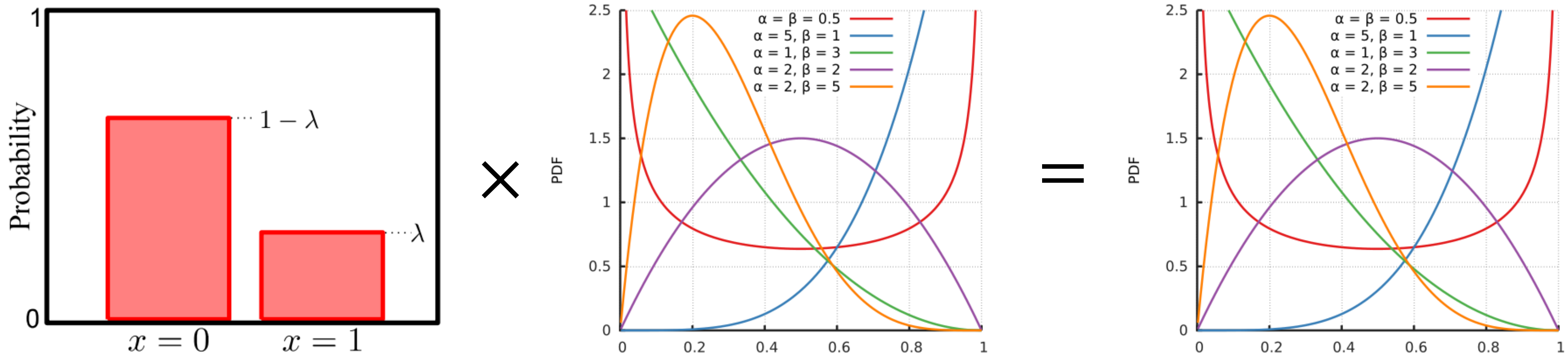


Wait!

Don’t you feel there is something oddly specific about the choice of these distributions?

Conjugate Priors

- When the posterior distribution is of the same shape of the prior, the prior is called a **conjugate distribution** for the likelihood
- Consider a Bernoulli data point with likelihood $p(x\,|\lambda)$, we can place a Beta-distributed prior on its parameter λ, i.e. $p(\lambda\,|\alpha, \beta)$
- The posterior $p(\lambda|x) \propto p(x\,|\lambda)p(\lambda|\alpha,\beta) = Beta(\lambda\,|\,\alpha',\beta')$ → same distribution of the prior!



Leveraging Conjugacy

Gives us guidelines to choose likelihood distributions on data and prior distributions on parameters, so that the posterior can be computed neatly in closed form

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

Inference in Probabilistic Learning Models

Wrapping Up….

- We know how to represent the world and the observations
  - Random Variables $\Longrightarrow X_1, \ldots, X_N$
  - Joint Probability Distribution $\Longrightarrow P(X_1 = x_1, \ldots, X_N = x_n)$
- We have rules for manipulating the probabilistic knowledge
  - Sum-Product
  - Marginalization
  - Bayes
  - Conditional Independence
- In this context, learning is about discovering the values for

$$P(x_1, \ldots, x_n)$$

Inference and Learning in Probabilistic Models

Inference - How can one determine the distribution of the values of one/several RV, given the observed values of others?

$$P(graduate|exam_1, \ldots, exam_n)$$

Machine Learning view - Given a set of observations (data) $\boldsymbol{d}$ and a set of hypotheses $\{h_i\}_i^K = 1$, how can I use them to predict the distribution of a RV X?

Learning - A very specific inference problem!

- Given a set of observations $\boldsymbol{d}$ and a probabilistic model of a given structure, how do I find the parameters θ of its distribution P_θ ?
- Amounts to determining the best hypothesis h_θ regulated by a (set of) parameters θ

3 Approaches to Inference

Bayesian Consider all hypotheses weighted by their probabilities

$$P(X|\boldsymbol{d}) = \sum_i P(X|h_i)P(h_i\,|\boldsymbol{d})$$

MAP Infer X from $P(X|h_{MAP})$ where h_{MAP} is the Maximum a-Posteriori hypothesis given $\boldsymbol{d}$

$$h_{MAP} = \arg\max_{h \in H} P(h|\boldsymbol{d}) = \arg\max_{h \in H} P(\boldsymbol{d}|h)P(h)$$

ML Assuming uniform priors $P(h_i) = P(h_j)$, yields the Maximum Likelihood (ML) estimate $P(X|h_{ML})$

$$h_{ML} = \arg\max_{h \in H} P(\boldsymbol{d}|h)$$

Let’s go to the cinema!

- How do I choose the next movie (prediction)?
- I might ask my friends for their favorite choice given their personal taste (hypothesis)
- Select the movie
  - Bayesian advice? Make a voting from all the friends’ suggestions weighted by their attendance to cinema and taste judgement
  - MAP advice? From the friend who goes often to the cinema and whose taste I trust
  - ML advice? From the friend who goes more often to the cinema

The Candy Box Problem

- A candy manufacturer produces 5 types of candy boxes (hypothesis) that are indistinguishable in the darkness of the cinema

  h_1 100% cherry flavor

  h_2 75% cherry and 25% lime flavor

  h_3 50% cherry and 50% lime flavor

  h_4 25% cherry and 75% lime flavor

  h_5 100% lime flavor

- Given a sequence of candies $\boldsymbol{d} = d_1, \ldots, d_N$ extracted and reinserted in a box (observations), what is the most likely flavor for the next candy (prediction)?

Candy Box Problem: Hypothesis Posterior

- First, we need to compute the posterior for each hypothesis (Bayes)

$$P(h_i|\boldsymbol{d}) = \alpha P(\boldsymbol{d}|h_i)P(h_i)$$

- The manufacturer is kind enough to provide us with the production shares (prior) for the 5 boxes

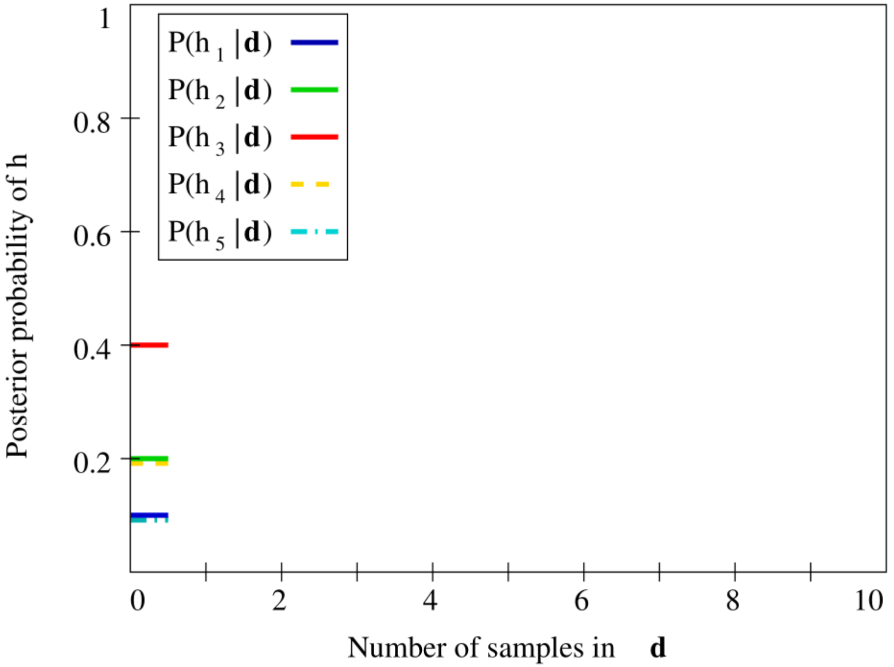
$$P(h_1), P(h_2), P(h_3), P(h_4), P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$$

- Data likelihood can be computed under the assumption that observations are independently and identically distributed (i.i.d.)

$$P(\boldsymbol{d}|h_i) = \prod_{j=1}^{N} P(d_j|h_i)$$

Candy Box Problem: Hypothesis Posterior Computation

Suppose that the bag is a h_5 and consider a sequence of 10 observed lime candies



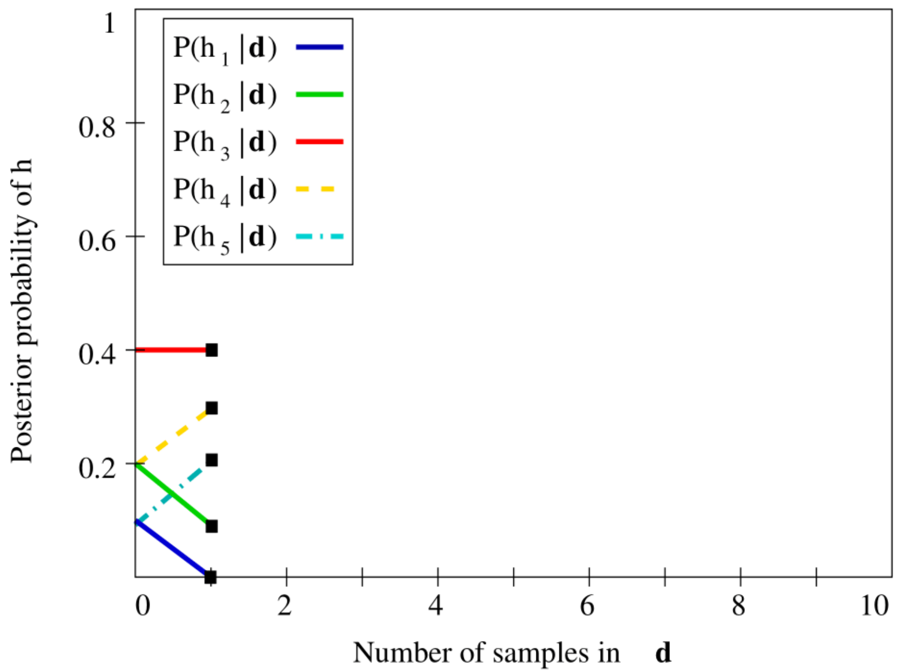
Hyp	d_0
h_1	0.1
h_2	0.2
h_3	0.4
h_4	0.2
h_5	0.1

$$P(h_i|\boldsymbol{d}) = P(h_i)$$

Posteriors start at the value of the prior

Candy Box Problem: Hypothesis Posterior Computation

Suppose that the bag is a h_5 and consider a sequence of 10 observed lime candies



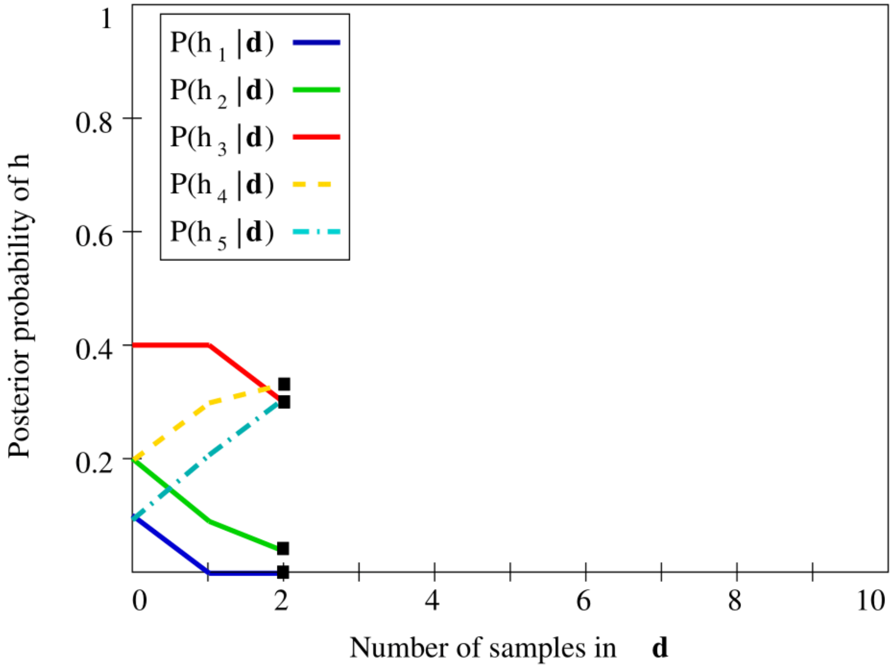
Hyp	d_0	d_1
h_1	0.1	0
h_2	0.2	0.1
h_3	0.4	0.4
h_4	0.2	0.3
h_5	0.1	0.2

$$P(h_i|\boldsymbol{d}) = \alpha P(h_i)P(d_1 = l|h_i)$$

Most likely MAP hypothesis is re-evaluated as more data comes in

Candy Box Problem: Hypothesis Posterior Computation

Suppose that the bag is a h_5 and consider a sequence of 10 observed lime candies



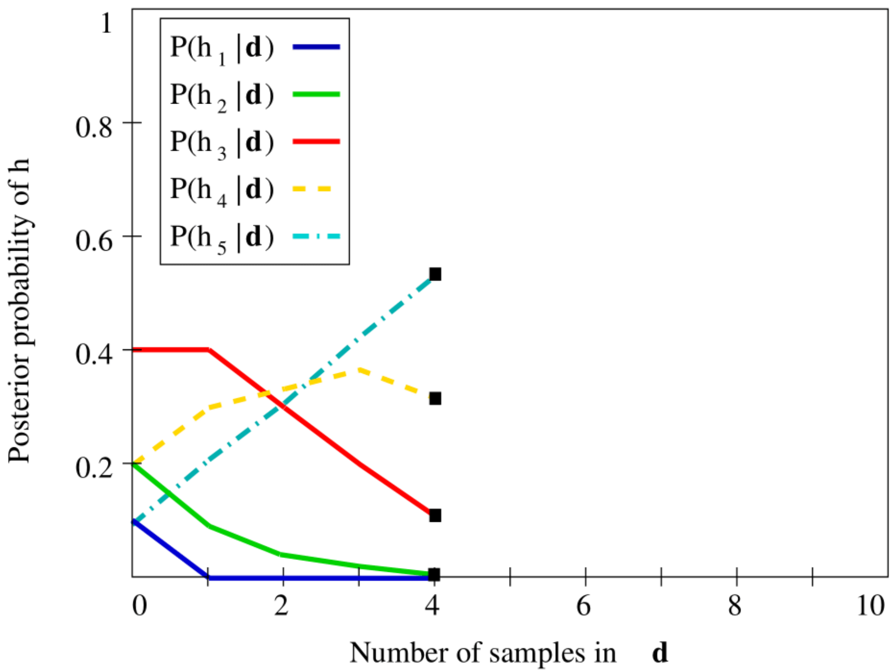
Hyp	d_0	d_1	d_2
h_1	0.1	0	0
h_2	0.2	0.1	0.03
h_3	0.4	0.4	0.30
h_4	0.2	0.3	0.35
h_5	0.1	0.2	0.31

$$P(h_i|\boldsymbol{d}) = \alpha P(h_i)P(d_1 = l|h_i) \times P(d_2 = l|h_i)$$

Most likely MAP hypothesis is re-evaluated as more data comes in

Candy Box Problem: Hypothesis Posterior Computation

Suppose that the bag is a h_5 and consider a sequence of 10 observed lime candies



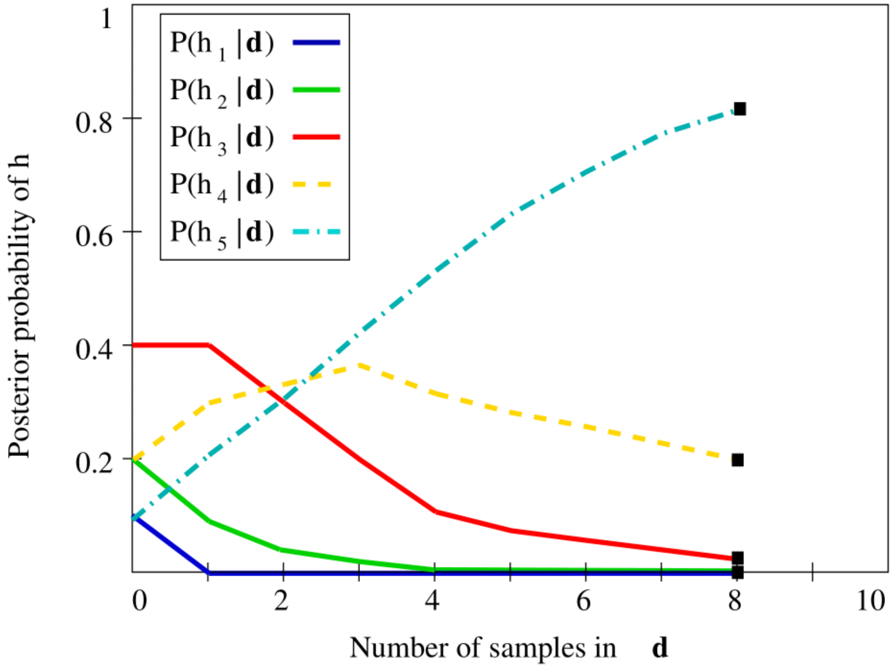
Hyp	d_0	d_1	d_2
h_1	0.1	0	0
h_2	0.2	0.1	0.03
h_3	0.4	0.4	0.30
h_4	0.2	0.3	0.35
h_5	0.1	0.2	0.31

$$P(h_i|\boldsymbol{d}) = \alpha P(h_i) P(d = l|h_i)^N$$

Most likely MAP hypothesis is re-evaluated as more data comes in

Candy Box Problem: Hypothesis Posterior Computation

Suppose that the bag is a h_5 and consider a sequence of 10 observed lime candies

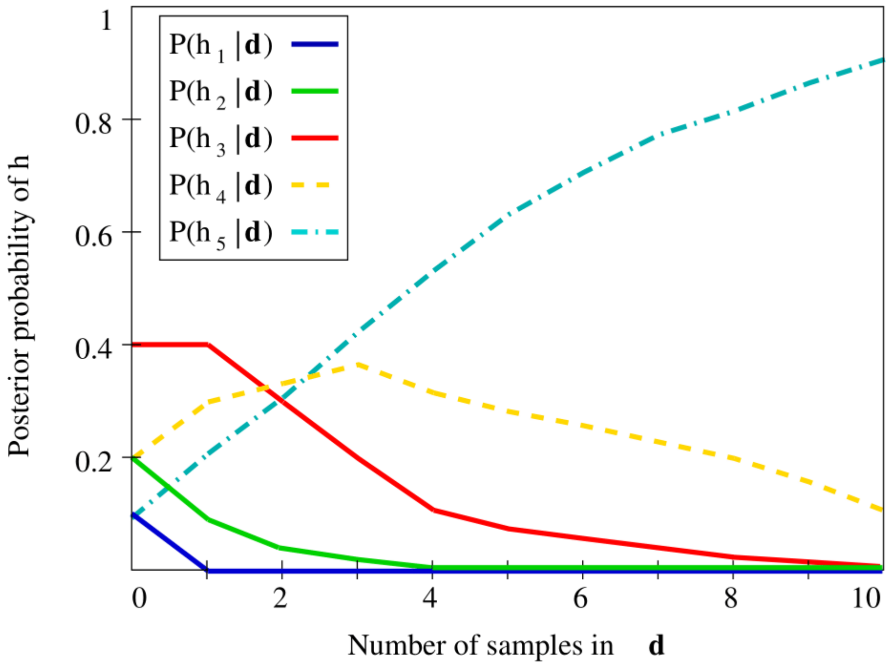


Hyp	d_0	d_1	d_2
h_1	0.1	0	0
h_2	0.2	0.1	0.03
h_3	0.4	0.4	0.30
h_4	0.2	0.3	0.35
h_5	0.1	0.2	0.31

Posteriors of false hypothesis eventually vanish

Candy Box Problem: Hypothesis Posterior Computation

Suppose that the bag is a h_5 and consider a sequence of 10 observed lime candies

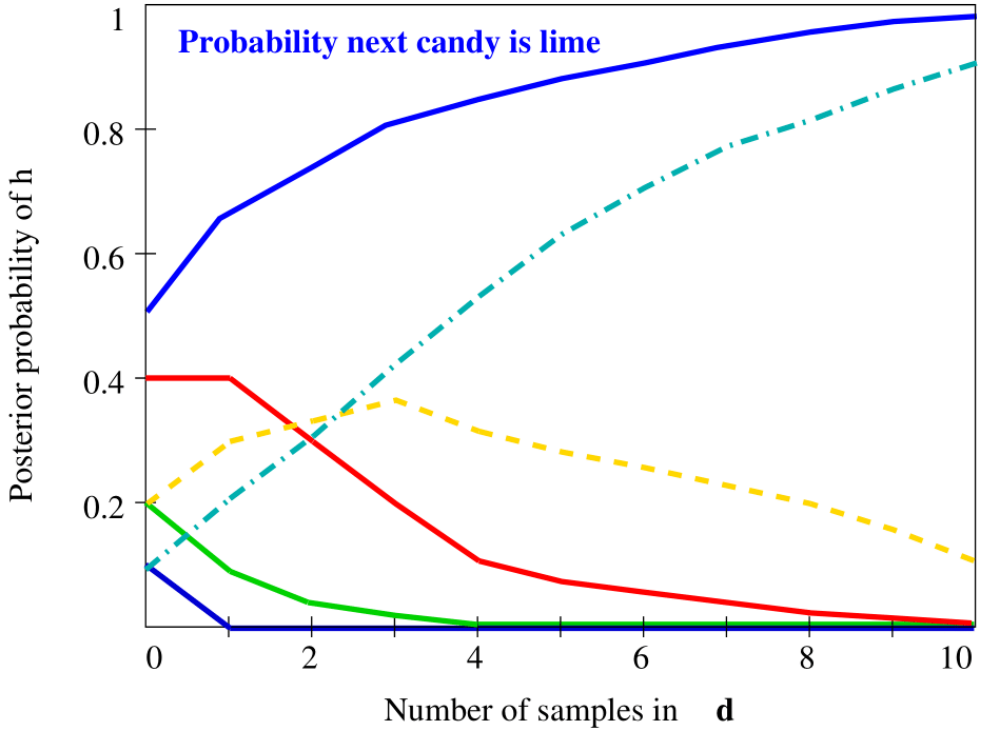


Hyp	d_0	d_1	d_2
h_1	0.1	0	0
h_2	0.2	0.1	0.03
h_3	0.4	0.4	0.30
h_4	0.2	0.3	0.35
h_5	0.1	0.2	0.31

Posteriors of false hypothesis eventually vanish

Candy Box Problem: Comparing Predictions

Bayesian learning seeks $P(d_{11} = l | d_1 = l, \ldots, d_{10} = l) = \sum_{i=1}^{5} P(d_{11} = l | h_i) P(h_i \,| \boldsymbol{d})$



Considerations About Bayesian Inference

- The Bayesian approach is optimal but poses computational and analytical tractability issues

$$P(X|\boldsymbol{d}) = \int_H P(X|h)P(h|\boldsymbol{d})dh$$

- ML and MAP are point estimates of the Bayesian since they infer based only on one most likely hypothesis

Conclusion

Take Home Messages

- Generative models as a gateway for next-gen deep learning
- Everything is an inference problem, including learning
- Graphical models represent probabilistic relationships between RV and conditional probabilities in compact way

Next 2 Lectures (24-25/03/2025)

Conditional independence: representation and learning

- Bayesian Networks
- Markov properties in Bayesian Networks
- Conditional independence as a graph-theoretic concept
- Conditional independence in undirected models
- Learning conditional independence relationships from data

Followed by lectures on causality and inferring conditional independence (and more) from data

Lectures by Riccardo Massidda