Preface

Use the template *preface.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your preface in the Springer layout.

A preface is a book's preliminary statement, usually written by the *author or editor* of a work, which states its origin, scope, purpose, plan, and intended audience, and which sometimes includes afterthoughts and acknowledgments of assistance.

When written by a person other than the author, it is called a foreword. The preface or foreword is distinct from the introduction, which deals with the subject of the work.

Customarily *acknowledgments* are included as last part of the preface.

Place(s), *Firstname Surname*
month year *Firstname Surname*

Contents

List of Contributors

Firstname Surname
ABC Institute, 123 Prime Street, Daisy Town, NA 01234, USA, e-mail: smith@smith.edu

Firstname Surname
XYZ Institute, Technical University, Albert-Schweitzer-Str. 34, 1000 Berlin, Germany, e-mail: meier@tu.edu

Acronyms

Use the template *acronym.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your list(s) of abbreviations or symbols in the Springer layout.

Lists of abbreviations, symbols and the like are easily formatted with the help of the Springer-enhanced `description` environment.

ABC	Spelled-out abbreviation and definition
BABI	Spelled-out abbreviation and definition
CABR	Spelled-out abbreviation and definition

Notation

Introduction

It is very difficult to come up with a single, consistent notation to cover the wide variety of data, models and algorithms that we discuss. Furthermore, conventions differ between machine learning and statistics, and between different books and papers. Nevertheless, we have tried to be as consistent as possible. Below we summarize most of the notation used in this book, although individual sections may introduce new notation. Note also that the same symbol may have different meanings depending on the context, although we try to avoid this where possible.

General math notation

Symbol	Meaning
$\lfloor x \rfloor$	Floor of x, i.e., round down to nearest integer
$\lceil x \rceil$	Ceiling of x, i.e., round down to nearest integer
$\boldsymbol{x} \otimes \boldsymbol{y}$	Convolution of $\boldsymbol{x}$ and $\boldsymbol{y}$
$\boldsymbol{x} \odot \boldsymbol{y}$	Hadamard (elementwise) product of $\boldsymbol{x}$ and $\boldsymbol{y}$
$a \wedge b$	logical AND
$a \vee b$	logical OR
$\neg a$	logical NOT
$\mathbb{I}(x)$	Indicator function, $\mathbb{I}(x) = 1$ if x is true, else $\mathbb{I}(x) = 0$
∞	Infinity
$\rightarrow$	Tends towards, e.g., $n \rightarrow \infty$
$\propto$	Proportional to, so $y = ax$ can be written as $y \propto x$
$\lvert x \rvert$	Absolute value
$\lvert \mathcal{S} \rvert$	Size (cardinality) of a set
$n!$	Factorial function
∇	Vector of first derivatives
∇^2	Hessian matrix of second derivatives
$\triangleq$	Defined as
$O(\cdot)$	Big-O: roughly means order of magnitude
$\mathbb{R}$	The real numbers
$1:n$	Range (Matlab convention): $1:n = 1, 2, ..., n$
$\approx$	Approximately equal to
$\arg\max_x f(x)$	Argmax: the value x that maximizes f
$B(a,b)$	Beta function, $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
$B(\boldsymbol{\alpha})$	Multivariate beta function, $\dfrac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$
$\binom{n}{k}$	n choose k , equal to $n!/(k!(nk)!)$
$\delta(x)$	Dirac delta function,$\delta(x) = \infty$ if $x = 0$, else $\delta(x) = 0$
$\exp(x)$	Exponential function e^x
$\Gamma(x)$	Gamma function, $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u$
$\Psi(x)$	Digamma function,$Psi(x) = \dfrac{d}{dx} \log \Gamma(x)$

$\mathcal{X}$	A set from which values are drawn (e.g.,$\mathcal{X} = \mathbb{R}^D$)

Linear algebra notation

We use boldface lower-case to denote vectors, such as $\boldsymbol{x}$, and boldface upper-case to denote matrices, such as $\boldsymbol{X}$. We denote entries in a matrix by non-bold upper case letters, such as X_{ij}.

Vectors are assumed to be column vectors, unless noted otherwise. We use $(x_1, \cdots, x_D)$ to denote a column vector created by stacking D scalars. If we write $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, where the left hand side is a matrix, we mean to stack the $\boldsymbol{x}_i$ along the columns, creating a matrix.

Symbol	Meaning
$\boldsymbol{X} \succ 0$	$\boldsymbol{X}$ is a positive definite matrix
$tr(\boldsymbol{X})$	Trace of a matrix
$det(\boldsymbol{X})$	Determinant of matrix $\boldsymbol{X}$
$\|\boldsymbol{X}\|$	Determinant of matrix $\boldsymbol{X}$
$\boldsymbol{X}^{-1}$	Inverse of a matrix
$\boldsymbol{X}^{\dagger}$	Pseudo-inverse of a matrix
$\boldsymbol{X}^T$	Transpose of a matrix
$\boldsymbol{x}^T$	Transpose of a vector
$\text{diag}(x)$	Diagonal matrix made from vector $\boldsymbol{x}$
$\text{diag}(X)$	Diagonal vector extracted from matrix $\boldsymbol{X}$
$\boldsymbol{I}$ or $\boldsymbol{I}_d$	Identity matrix of size $d \times d$ (ones on diagonal, zeros of)
$\mathbf{1}$ or $\mathbf{1}_d$	Vector of ones (of length d)
$\mathbf{0}$ or $\mathbf{0}_d$	Vector of zeros (of length d)
$\|\|\boldsymbol{x}\|\| = \|\|\boldsymbol{x}\|\|_2$	Euclidean or ℓ_2 norm $\sqrt{\sum_{j=1}^{d} x_j^2}$
$\|\|\boldsymbol{x}\|\|_1$	ℓ_1 norm $\sum_{j=1}^{d} \|x_j\|$
$\boldsymbol{X}_{:,j}$	j'th column of matrix
$\boldsymbol{X}_{i,:}$	transpose of i'th row of matrix (a column vector)
$\boldsymbol{X}_{i,j}$	Element (i, j) of matrix $\boldsymbol{X}$
$\boldsymbol{x} \otimes \boldsymbol{y}$	Tensor product of $\boldsymbol{x}$ and $\boldsymbol{y}$

Probability notation

We denote random and fixed scalars by lower case, random and fixed vectors by bold lower case, and random and fixed matrices by bold upper case. Occasionally we use non-bold upper case to denote scalar random variables. Also, we use $p()$ for both discrete and continuous random variables

Symbol	Meaning
X, Y	Random variable
$P()$	Probability of a random event
$F()$	Cumulative distribution function(CDF), also called distribution function
$p(x)$	Probability mass function(PMF)
$f(x)$	probability density function(PDF)
$F(x,y)$	Joint CDF
$p(x,y)$	Joint PMF
$f(x,y)$	Joint PDF

$p(X\|Y)$	Conditional PMF, also called conditional probability
$f_{X\|Y}(x\|y)$	Conditional PDF
$X \perp Y$	X is independent of Y
$X \not\perp Y$	X is not independent of Y
$X \perp Y\|Z$	X is conditionally independent of Y given Z
$X \not\perp Y\|Z$	X is not conditionally independent of Y given Z
$X \sim p$	X is distributed according to distribution p
$\boldsymbol{\alpha}$	Parameters of a Beta or Dirichlet distribution
$\text{cov}[X]$	Covariance of X
$\mathbb{E}[X]$	Expected value of X
$\mathbb{E}_q[X]$	Expected value of X wrt distribution q
$\mathbb{H}(X)$ or $\mathbb{H}(p)$	Entropy of distribution $p(X)$
$\mathbb{I}(X;Y)$	Mutual information between X and Y
$\mathbb{KL}(p\|\|q)$	KL divergence from distribution p to q
$\ell(\boldsymbol{\theta})$	Log-likelihood function
$L(\boldsymbol{\theta}, a)$	Loss function for taking action a when true state of nature is $\boldsymbol{\theta}$
λ	Precision (inverse variance) $\lambda = 1/\sigma^2$
$\boldsymbol{\Lambda}$	Precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$
$\text{mode}[\boldsymbol{X}]$	Most probable value of $\boldsymbol{X}$
μ	Mean of a scalar distribution
$\boldsymbol{\mu}$	Mean of a multivariate distribution
$\boldsymbol{\Phi}$	cdf of standard normal
ϕ	pdf of standard normal
$\boldsymbol{\pi}$	multinomial parameter vector, Stationary distribution of Markov chain
ρ	Correlation coefficient
$\text{sigm}(x)$	Sigmoid (logistic) function, $\frac{1}{1+e^{-x}}$
σ^2	Variance
$\boldsymbol{\Sigma}$	Covariance matrix
$\text{var}[x]$	Variance of x
ν	Degrees of freedom parameter
Z	Normalization constant of a probability distribution

Machine learning/statistics notation

In general, we use upper case letters to denote constants, such as C, K, M, N, T, etc. We use lower case letters as dummy indexes of the appropriate range, such as $c = 1 : C$ to index classes, $i = 1 : M$ to index data cases, $j = 1 : N$ to index input features, $k = 1 : K$ to index states or clusters, $t = 1 : T$ to index time, etc.

We use x to represent an observed data vector. In a supervised problem, we use y or $\boldsymbol{y}$ to represent the desired output label. We use $\boldsymbol{z}$ to represent a hidden variable. Sometimes we also use q to represent a hidden discrete variable.

Symbol	**Meaning**
C	Number of classes
D	Dimensionality of data vector (number of features)
N	Number of data cases
N_c	Number of examples of class c,$N_c = \sum_{i=1}^{N} \mathbb{I}(y_i = c)$
R	Number of outputs (response variables)
$\mathcal{D}$	Training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \| i = 1 : N\}$
$\mathcal{D}_{test}$	Test data
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space

K	Number of states or dimensions of a variable (often latent)
$k(x,y)$	Kernel function
$\boldsymbol{K}$	Kernel matrix
$\mathcal{H}$	Hypothesis space
L	Loss function
$J(\boldsymbol{\theta})$	Cost function
$f(\boldsymbol{x})$	Decision function
$P(y\|\boldsymbol{x})$	TODO
λ	Strength of ℓ_2 or ℓ_1 *regularizer*
$\phi(x)$	Basis function expansion of feature vector $\boldsymbol{x}$
$\boldsymbol{\Phi}$	Basis function expansion of design matrix $\boldsymbol{X}$
$q()$	Approximate or proposal distribution
$Q(\boldsymbol{\theta},\boldsymbol{\theta}_{old})$	Auxiliary function in EM
T	Length of a sequence
$T(\mathcal{D})$	Test statistic for data
$\boldsymbol{T}$	Transition matrix of Markov chain
$\boldsymbol{\theta}$	Parameter vector
$\boldsymbol{\theta}^{(s)}$	s'th sample of parameter vector
$\hat{\boldsymbol{\theta}}$	Estimate (usually MLE or MAP) of $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}_{MLE}$	Maximum likelihood estimate of $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}_{MAP}$	MAP estimate of $\boldsymbol{\theta}$
$\bar{\boldsymbol{\theta}}$	Estimate (usually posterior mean) of $\boldsymbol{\theta}$
$\boldsymbol{w}$	Vector of regression weights (called $\boldsymbol{\beta}$ in statistics)
b	intercept (called $\boldsymbol{\varepsilon}$ in statistics)
$\boldsymbol{W}$	Matrix of regression weights
x_{ij}	Component (i.e., feature) j of data case i ,for $i = 1:N, j = 1:D$
$\boldsymbol{x}_i$	Training case, $i = 1:N$
$\boldsymbol{X}$	Design matrix of size $N \times D$
$\bar{\boldsymbol{x}}$	Empirical mean $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$
$\tilde{\boldsymbol{x}}$	Future test case
$\boldsymbol{x}_*$	Feature test case
$\boldsymbol{y}$	Vector of all training labels $\boldsymbol{y} = (y_1,...,y_N)$
z_{ij}	Latent component j for case i

Chapter 1
Introduction

1.1 Types of machine learning

$$\begin{cases} \text{Supervised Learning} \begin{cases} \text{Classification} \\ \text{Regression} \end{cases} \\ \text{Unsupervised Learning} \begin{cases} \text{Discovering clusters} \\ \text{Discovering latent factors} \\ \text{Discovering graph structure} \\ \text{Matrix completion} \end{cases} \\ \text{Reinforcement Learning} \end{cases}$$

In the **predictive** or **supervised learning** approach,the goal is to learn a **mapping** from **inputs x** to **outputs** y,given a labeled set of input-output paris $D = \{(\mathbf{x_i}, y_i)\}_{i=1}N$.Here D is called the **training set**,and N is the number of training examples. In the simplest setting,each training input $\mathbf{x_i}$ is a D-dimensional vector of numbers,representing,say,the height and weight of a person,which are called **features,attributes,or covariates**.

1.2 Supervised learning

Similarly the form of the output or **response variable** can in principle be anything,but most methods assume that y_i is **categorical or nominal** variable from some finite set,$y_i \in \{1,...,C\}$.When y_i is categorical,the problem is known as **classification or pattern recognition**,and when real-valued,known as **regression**.

1.2.1 Classification

Here the goal is to learn a mapping from inputs x to outputs y,where $y \in \{1,...,C\}$,with C being the number of classes.If $C = 2$,this is called **binary classification**;if $C > 2$,this is called **multiclass classification**.if the class labels are not mutually exclusive,we call it **multi-label classification**,but this is best viewed as predicting multiple related binary class labels(a so-called **multiple output model**). One way to formalize the problem is as **function approximation**:assume $y = f(\mathbf{x})$ for some unknown function f,and the goal of learning is to estimate the function f given a labeled training set,and then to make predictions(estimate) using $\hat{y} = \hat{f}(\mathbf{x})$.Our main goal is to make predictions on novel inputs,meaning ones that we have not seen before(**generalization**.

1.2.1.1 Probabilistic predictions

Given a probabilistic output,we can always compute out "best guess" as to the "true label" using

$$\hat{y} = \hat{f}(\mathbf{x}) = \arg\max_{c=1}^{C} p(y = c|\mathbf{x}, D) \tag{1.1}$$

This corresponds to a **MAP estimate**(MAP stands for**maximum a posteriori**).

1.2.1.2 Applications

1.3 Unsupervised learning

Descriptive or unsupervised learning approach is sometimes called **knowledge discovery**.We will formalize out task as one of **density estimation**,that is we want to build models of the form $p(\mathbf{x_i}|\theta)$,instead of $p(y_i|\mathbf{x_i},\theta)$.

1.3.1 Discovering clusters

Let $z_i \in \{1,...,K\}$ represent the cluster to which data point i is assigned.(z_i is an exmaple of **hidden or latent** variable).

1.3.2 Discovering latent factors

Although the data may appear high dimensional,there may only be a small number of degrees of variability,corresponding to **latent factors**.The most common approach to dimensionality reduction is called **principal components analysis or PCA**.

1.3.3 Discovering graph structure

1.3.4 Matrix completion

1.3.4.1 Image inpainting

1.3.4.2 Collaborative filtering

1.3.4.3 Market basket analysis

1.4 Three elements of a machine learning model

Model = Representation + Evaluation + Optimization[1]

1.4.1 Representation

In supervised learning, a model must be represented as a conditional probability distribution $P(y|\boldsymbol{x})$(usually we call it classifier) or a decision function $f(x)$. The set of classifiers(or decision functions) is called the hypothesis space of the model. Choosing a representation for a model is tantamount to choosing the hypothesis space that it can possibly learn.

[1] Domingos, P. A few useful things to know about machine learning. Commun. ACM. 55(10):7887 (2012).

1.4.2 Evaluation

In the hypothesis space, an evaluation function (also called objective function or risk function) is needed to distinguish good classifiers(or decision functions) from bad ones.

1.4.2.1 Loss function and risk function

Definition 1.1. In order to measure how well a function fits the training data, a **loss function** $L : Y \times Y \to R \geq 0$ is defined. For training example (x_i, y_i), the loss of predicting the value $\widehat{y}$ is $L(y_i, \widehat{y})$.

The following is some common loss functions:

1. 0-1 loss function
   $L(Y, f(X)) = \mathbb{I}(Y, f(X)) = \begin{cases} 1, & Y = f(X) \\ 0, & Y \neq f(X) \end{cases}$
2. Quadratic(squared) loss function $L(Y, f(X)) = \frac{1}{2}(Y - f(X))^2$
3. Absolute loss function $L(Y, f(X)) = |Y - f(X)|$
4. Exponential loss function $L(Y, f(X) = exp(-\hat{y}_i f(\mathbf{x_i}))$
5. Logarithmic loss function
   $L(Y, P(Y|X)) = -\log P(Y|X)$

Name	Loss	Derivative	f^*	Algorithm
Squared error	$\frac{1}{2}(y_i - f(\mathbf{x_i}))^2$	$y_i - f(\mathbf{x_i})$	$\mathbb{E}[y\|\mathbf{x_i}]$	L2Boosting
Absolute error	$\|y_i - f(\mathbf{x_i})\|$	$sgn(y_i - f(\mathbf{x_i}))$	$median(y\|\mathbf{x_i})$	Gradient boosting
Exponential loss	$exp(-\hat{y}_i f(\mathbf{x_i}))$	$-\hat{y}_i exp(-\hat{y}_i f(\mathbf{x_i}))$	$\frac{1}{2} log \frac{\pi_i}{1-\pi_i}$	AdaBoost
Logloss	$log(1 + e^{-\hat{y}_i f_i})$	$y_i - \pi_i$	$\frac{1}{2} log \frac{\pi_i}{1-\pi_i}$	LogitBoost

Definition 1.2. The risk of function f is defined as the expected loss of f:

$$R_{\exp}(f) = E\left[L(Y, f(X))\right] = \int L(y, f(x)) P(x, y) \mathrm{d}x\mathrm{d}y \tag{1.2}$$

which is also called expected loss or **risk function**.

Definition 1.3. The risk function $R_{\exp}(f)$ can be estimated from the training data as

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) \tag{1.3}$$

which is also called empirical loss or **empirical risk**.

You can define your own loss function, but if you're a novice, you're probably better off using one from the literature. There are conditions that loss functions should meet[2]:

1. They should approximate the actual loss you're trying to minimize. As was said in the other answer, the standard loss functions for classification is zero-one-loss (misclassification rate) and the ones used for training classifiers are approximations of that loss.
2. The loss function should work with your intended optimization algorithm. That's why zero-one-loss is not used directly: it doesn't work with gradient-based optimization methods since it doesn't have a well-defined gradient (or even a subgradient, like the hinge loss for SVMs has).
   The main algorithm that optimizes the zero-one-loss directly is the old perceptron algorithm(chapter §**??**).

[2] http://t.cn/zTrDxLO

1.4.2.2 ERM and SRM

Definition 1.4. ERM(Empirical risk minimization)

$$\min_{f\in\mathcal{F}} R_{\text{emp}}(f) = \min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \tag{1.4}$$

Definition 1.5. Structural risk

$$R_{\text{smp}}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.5}$$

Definition 1.6. SRM(Structural risk minimization)

$$\min_{f\in\mathcal{F}} R_{\text{srm}}(f) = \min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.6}$$

1.4.3 Optimization

Finally, we need a **training algorithm**(also called **learning algorithm**) to search among the classifiers in the the hypothesis space for the highest-scoring one. The choice of optimization technique is key to the **efficiency** of the model.

1.5 Some basic concepts

1.5.1 Parametric vs non-parametric models

1.5.2 A simple non-parametric classifier: K-nearest neighbours

1.5.2.1 Representation

$$y = f(\boldsymbol{x}) = \arg\min_{c} \sum_{\boldsymbol{x}_i\in N_k(\boldsymbol{x})} \mathbb{I}(y_i = c) \tag{1.7}$$

where $N_k(\boldsymbol{x})$ is the set of k points that are closest to point $\boldsymbol{x}$.

Usually use **k-d tree** to accelerate the process of finding k nearest points.

1.5.2.2 Evaluation

No training is needed.

1.5.2.3 Optimization

No training is needed.

1.5.3 Overfitting

1.5.4 Cross validation

Definition 1.7. Cross validation, sometimes called *rotation estimation*, is a *model validation* technique for assessing how the results of a statistical analysis will generalize to an independent data set[3].

Common types of cross-validation:

1. K-fold cross-validation. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k 1 subsamples are used as training data.
2. 2-fold cross-validation. Also, called simple cross-validation or holdout method. This is the simplest variation of k-fold cross-validation, k=2.
3. Leave-one-out cross-validation(*LOOCV*). k=M, the number of original samples.

1.5.5 Model selection

When we have a variety of models of different complexity (e.g., linear or logistic regression models with different degree polynomials, or KNN classifiers with different values ofK), how should we pick the right one? A natural approach is to compute the **misclassification rate** on the training set for each method.

[3] http://en.wikipedia.org/wiki/Cross-validation_(statistics)

Chapter 2
Probability

2.1 Frequentists vs. Bayesians

There are two different interpretations of probability. One is called the **frequentist** interpretation. In this view, probabilities represent long run frequencies of events. For example, the above statement means that, if we flip the coin many times, we expect it to land heads about half the time.

The other interpretation is called the **Bayesian** interpretation of probability. In this view, probability is used to quantify our **uncertainty** about something; hence it is fundamentally related to information rather than repeated trials (Jaynes 2003). In the Bayesian view, the above statement means we believe the coin is equally likely to land heads or tails on the next toss

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies. For example, we might want to compute the probability that the polar ice cap will melt by 2020 CE. This event will happen zero or one times, but cannot happen repeatedly. Nevertheless,we thought to be able to quantify our uncertainty about this event. To give another machine learning oriented example, we might have observed a blip on our radar screen, and want to compute the probability distribution over the location of the corresponding target (be it a bird, plane, or missile). In all these cases, the idea of repeated trials does not make sense, but the Bayesian interpretation is valid and indeed quite natural. We shall therefore adopt the Bayesian interpretation in this book. Fortunately, the basic rules of probability theory are the same, no matter which interpretation is adopted.

2.2 Basic concepts of probability theory

2.2.1 Discrete random variables

The expression $p(A)$ denotes the probability that event A is true.We require that $0 \leq p(A) \leq 1$,where 0 means the event definitely will not happen,and $p(A) = 1$ means the event definitely will happen.$p(\hat{A})$ denotes the probability of the event not A;this is defined to be $p(\hat{A}) = 1 - p(A)$.

We denote a random event by defining a **random variable** X. **Descrete random variable**: X ,which can take on any value from a finite or countably infinite set .We denote the probability of the event that $X = x$ by $p(X = x)$,or just $p(x)$ for short.Here $p()$ is called a **probability mass function** or **pmf**.The pmfs are defined one **state space**.$\mathbb{I}$ denotes the binary **indicator funcition**.

Continuous random variable: the value of X is real-valued.

2.2.2 Fundamental rules

In this section,we review the basic rule of probability.

2.2.2.1 Probability of a union of two events

Given two events,A and B,we define the probability of A or B as follows:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \tag{2.1}$$
$$= p(A) + p(B) \tag{2.2}$$

if A and B are mutually independent

2.2.2.2 Joint probabilities

We define the probability of the joint event A and B as follows:

$$p(A,B) = p(A \cap B) = p(A|B)p(B) \tag{2.3}$$

This is sometimes called the **product rule**

2.2.2.3 Conditional probability

Define the **conditional probability** of event A,given that event B is true,as follows:

$$p(A|B) = \frac{p(A,B)}{p(B)}, if p(B) > 0 \tag{2.4}$$

2.2.2.4 CDF

$$F(x) \triangleq P(X \leq x) = \begin{cases} \sum_{u \leq x} p(u) & \text{, discrete} \\ \int_{-\infty}^{x} f(u)\mathrm{d}u & \text{, continuous} \end{cases} \tag{2.5}$$

2.2.2.5 PMF and PDF

For descrete random variable, We denote the probability of the event that $X = x$ by $P(X = x)$, or just $p(x)$ for short. Here $p(x)$ is called a **probability mass function** or **PMF**.A probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value[4]. This satisfies the properties $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$.

For continuous variable, in the equation $F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u$, the function $f(x)$ is called a **probability density function** or **PDF**. A probability density function is a function that describes the relative likelihood for this random variable to take on a given value[5].This satisfies the properties $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

2.2.3 Mutivariate random variables

2.2.3.1 Joint CDF

We denote joint CDF by $F(x,y) \triangleq P(X \leq x \cap Y \leq y) = P(X \leq x, Y \leq y)$.

$$F(x,y) \triangleq P(X \leq x, Y \leq y) = \begin{cases} \sum_{u \leq x, v \leq y} p(u,v) \\ \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\mathrm{d}u\mathrm{d}v \end{cases} \tag{2.6}$$

product rule:

$$p(X,Y) = P(X|Y)P(Y) \tag{2.7}$$

Chain rule:

$$p(X_{1:N}) = p(X_1)p(X_3|X_2,X_1)...p(X_N|X_{1:N-1}) \tag{2.8}$$

[4] http://en.wikipedia.org/wiki/Probability_mass_function

[5] http://en.wikipedia.org/wiki/Probability_density_function

2.2.3.2 Marginal distribution

Marginal CDF:

$$F_X(x) \triangleq F(x,+\infty) = \begin{cases} \sum_{x_i \le x} P(X=x_i) = \sum_{x_i \le x} \sum_{j=1}^{+\infty} P(X=x_i, Y=y_j) \\ \int_{-\infty}^{x} f_X(u)du = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} f(u,v)\mathrm{d}u\mathrm{d}v \end{cases} \tag{2.9}$$

$$F_Y(y) \triangleq F(+\infty,y) = \begin{cases} \sum_{y_j \le y} p(Y=y_j) = \sum_{i=1}^{+\infty} \sum_{y_j \le y} P(X=x_i, Y=y_j) \\ \int_{-\infty}^{y} f_Y(v)dv = \int_{-\infty}^{+\infty} \int_{-\infty}^{y} f(u,v)\mathrm{d}u\mathrm{d}v \end{cases} \tag{2.10}$$

Marginal PMF and PDF:

$$\begin{cases} P(X=x_i) = \sum_{j=1}^{+\infty} P(X=x_i, Y=y_j) & \text{, descrete} \\ f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y & \text{, continuous} \end{cases} \tag{2.11}$$

$$\begin{cases} p(Y=y_j) = \sum_{i=1}^{+\infty} P(X=x_i, Y=y_j) & \text{, descrete} \\ f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x & \text{, continuous} \end{cases} \tag{2.12}$$

2.2.3.3 Conditional distribution

Conditional PMF:

$$p(X=x_i|Y=y_j) = \frac{p(X=x_i, Y=y_j)}{p(Y=y_j)} \text{ if } p(Y)>0 \tag{2.13}$$

The pmf $p(X|Y)$ is called **conditional probability**.

Conditional PDF:

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \tag{2.14}$$

2.2.4 Bayes rule

$$\begin{aligned} p(Y=y|X=x) &= \frac{p(X=x,Y=y)}{p(X=x)} \\ &= \frac{p(X=x|Y=y)p(Y=y)}{\sum_{y'} p(X=x|Y=y')p(Y=y')} \end{aligned} \tag{2.15}$$

sum rule

$$p(X) = \sum_{Y} p(X,Y) \tag{2.16}$$

product rule

$$p(X,Y) = p(Y|X)p(X) \tag{2.17}$$

Bayes' theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{2.18}$$

Denominator in Bayes' theorem

$$p(X) = \sum_{Y} p(X|Y)p(Y) \tag{2.19}$$

probability densities

$$p(x \in (a,b)) = \int_a^b p(x)dx \tag{2.20}$$

The probability density function p(x) must satisfy the two conditions

$$\begin{cases} p(x) \geq 0 \\ \int_{-\infty}^{\infty} p(x)dx = 1 \end{cases} \tag{2.21}$$

Combinations of discrete and continuous variables.

$$p(x) = \int_p (x,y)dy \tag{2.22}$$

$$p(x,y) = p(y|x)p(x) \tag{2.23}$$

Bayesian probabilities So far, we have viewed probabilities in terms of the frequencies of random,repeatable events,which we shall refer to as the classical or frequentist interpretation of probability.Now we turn to the more general Bayesian view,in which probabilities provide a quantification of uncertainty. We can adopt a similar approach when making inferences about quantities such as the parameters **w** in the polynomial curve fitting.We capture our assumptions about **w**,before observing the data,in the form of a prior probability distribution $p(\mathbf{w})$.The effect of the observed data $D = t_1, ..., t_N$ is expressed through the conditional probability $p(D|w)$

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)} \tag{2.24}$$

The quantity $p(D|w)$ on the right-hand side of Bayes' theorem is evaluated for the observed data set D and can be viewed as a function of the parameter vector **w**,in which case it is called the likelihood function.

$$posteroir \propto likelihood \times prior \tag{2.25}$$

Integrating both side with respect to **w**

$$p(D) = \int p(D|\mathbf{w})p(\mathbf{w})d\mathbf{w} \tag{2.26}$$

A widely used frequentise estimator is maximum likelihood,in which **w** is set to the value that maximizes the likehood function $p(D|w)$.

2.2.5 Independence and conditional independence

We say X and Y are unconditionally independent or marginally independent, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals, i.e.,

$$X \perp Y = P(X,Y) = P(X)P(Y) \tag{2.27}$$

We say X and Y are conditionally independent(CI) given Z if the conditional joint can be written as a product of conditional marginals:

$$X \perp Y|Z = P(X,Y|Z) = P(X|Z)P(Y|Z) \tag{2.28}$$

2.2.6 Quantiles

Since the cdf F is a monotonically increasing function, it has an inverse; let us denote this by F^{-1}. If F is the cdf of X , then $F^{-1}(\alpha)$ is the value of x_α such that $P(X \leq x_\alpha) = \alpha$; this is called the α quantile of F. The value $F^{-1}(0.5)$ is the

median of the distribution, with half of the probability mass on the left, and half on the right. The values $F^{-1}(0.25)$ and $F^{1}(0.75)$are the lower and upper **quartiles**.

2.2.7 Mean and variance

The most familiar property of a distribution is its **mean**,or **expected value**, denoted by μ. For discrete rvs, it is defined as $\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} xp(x)$, and for continuous rvs, it is defined as $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} xp(x)\mathrm{d}x$. If this integral is not finite, the mean is not defined (we will see some examples of this later).

The **variance** is a measure of the spread of a distribution, denoted by σ^2. This is defined as follows:

$$\begin{aligned} var[X] &= \mathbb{E}[(X-\mu)^2] \qquad (2.29)\\ &= \int (x-\mu)^2 p(x)\mathrm{d}x \\ &= \int x^2 p(x)\mathrm{d}x + \mu^2 \int p(x)\mathrm{d}x - 2\mu \int xp(x)\mathrm{d}x \\ &= \mathbb{E}[X^2] - \mu^2 \qquad (2.30)\end{aligned}$$

from which we derive the useful result

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2 \qquad (2.31)$$

The **standard deviation** is defined as

$$std[X] \triangleq \sqrt{var[X]} \qquad (2.32)$$

This is useful since it has the same units as X itself.

Expectations and covariances The average value of some function $f(x)$ under a probability distribution p(x) is called the expectation of f(x) and will be denoted by

$$\mathbb{E}[f] = \sum_x p(x)f(x)\mathbb{E}[f] = \int p(x)f(x)dx \qquad (2.33)$$

approximation

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n) \qquad (2.34)$$

conditional expectation with respect to a conditional distribution

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x) \qquad (2.35)$$

variance of f(x) is defined by

$$var[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2 var[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \qquad (2.36)$$

convariance

$$cov[x,y] = \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \qquad (2.37)$$

In the case of two vectors of random variables **x** and **y**

$$cov[\mathbf{x},\mathbf{y}] = \mathbb{E}_{x,y}[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T]\}]cov[\mathbf{x},\mathbf{y}] = \mathbb{E}_{x,y}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T] \qquad (2.38)$$

2.3 Some common discrete distributions

In this section, we review some commonly used parametric distributions defined on discrete state spaces, both finite and countably infinite.

2.3.1 The Bernoulli and binomial distributions

Definition 2.1. Now suppose we toss a coin only once. Let $X \in \{0,1\}$ be a binary random variable, with probability of success or heads of θ. We say that X has a **Bernoulli distribution**. This is written as $X \sim \text{Ber}(\theta)$, where the pmf is defined as

$$\text{Ber}(x|\theta) \triangleq \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \tag{2.39}$$

Definition 2.2. Suppose we toss a coin n times. Let $X \in \{0,1,\cdots,n\}$ be the number of heads. If the probability of heads is θ, then we say X has a **binomial distribution**, written as $X \sim \text{Bin}(n,\theta)$. The pmf is given by

$$\text{Bin}(k|n,\theta) \triangleq \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{2.40}$$

2.3.2 The multinoulli and multinomial distributions

Definition 2.3. The Bernoulli distribution can be used to model the outcome of one coin tosses. To model the outcome of tossing a K-sided dice, let $\boldsymbol{x} = (\mathbb{I}(x=1),\cdots,\mathbb{I}(x=K)) \in \{0,1\}^K$ be a random vector(this is called **dummy encoding** or **one-hot encoding**), then we say X has a **multinoulli distribution**(or **categorical distribution**), written as $X \sim \text{Cat}(\theta)$. The pmf is given by:

$$p(\boldsymbol{x}) \triangleq \prod_{k=1}^{K}\theta_k^{\mathbb{I}(x_k=1)} \tag{2.41}$$

Definition 2.4. Suppose we toss a K-sided dice n times. Let $\boldsymbol{x} = (x_1,x_2,\cdots,x_K) \in \{0,1,\cdots,n\}^K$ be a random vector, where x_j is the number of times side j of the dice occurs, then we say X has a **multinomial distribution**, written as $X \sim \text{Mu}(n,\boldsymbol{\theta})$. The pmf is given by

$$p(\boldsymbol{x}) \triangleq \binom{n}{x_1\cdots x_k}\prod_{k=1}^{K}\theta_k^{x_k} \tag{2.42}$$

where $\binom{n}{x_1\cdots x_k} \triangleq \dfrac{n!}{x_1!x_2!\cdots x_K!}$

Bernoulli distribution is just a special case of a Binomial distribution with $n=1$, and so is multinoulli distribution as to multinomial distribution. See Table 2.1 for a summary.

Table 2.1: Summary of the multinomial and related distributions.

Name	K	n	X
Bernoulli	1	1	$x \in \{0,1\}$
Binomial	1	-	$\boldsymbol{x} \in \{0,1,\cdots,n\}$
Multinoulli	-	1	$\boldsymbol{x} \in \{0,1\}^K, \sum_{k=1}^K x_k = 1$
Multinomial	-	-	$\boldsymbol{x} \in \{0,1,\cdots,n\}^K, \sum_{k=1}^K x_k = n$

2.3.3 The Poisson distribution

Definition 2.5. We say that $X \in \{0,1,2,\cdots\}$ has a **Poisson distribution** with parameter $\lambda > 0$, written as $X \sim \text{Poi}(\lambda)$, if its pmf is

$$p(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{2.43}$$

The first term is just the normalization constant, required to ensure the distribution sums to 1.

The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents.

Table 2.2: Summary of Bernoulli, binomial multinoulli and multinomial distributions.

Name	Written as	X	$p(x)$(or $p(\boldsymbol{x})$)	$\mathbb{E}[X]$	var$[X]$
Bernoulli	$X \sim \text{Ber}(\theta)$	$x \in \{0,1\}$	$\theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$	θ	$\theta(1-\theta)$
Binomial	$X \sim \text{Bin}(n,\theta)$	$x \in \{0,1,\cdots,n\}$	$\binom{n}{k}\theta^k(1-\theta)^{n-k}$	$n\theta$	$n\theta(1-\theta)$
Multinoulli	$X \sim \text{Cat}(\boldsymbol{\theta})$	$\boldsymbol{x} \in \{0,1\}^K, \sum_{k=1}^K x_k = 1$	$\prod_{k=1}^K \theta_j^{\mathbb{I}(x_j=1)}$	-	-
Multinomial	$X \sim \text{Mu}(n,\boldsymbol{\theta})$	$\boldsymbol{x} \in \{0,1,\cdots,n\}^K, \sum_{k=1}^K x_k = n$	$\binom{n}{x_1\cdots x_k}\prod_{k=1}^K \theta_j^{x_j}$	-	-
Poisson	$X \sim \text{Poi}(\lambda)$	$x \in \{0,1,2,\cdots\}$	$e^{-\lambda}\frac{\lambda^x}{x!}$	λ	λ

2.3.4 The empirical distribution

The **empirical distribution function**[6], or **empirical cdf**, is the cumulative distribution function associated with the empirical measure of the sample. Let $\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$ be a sample set, it is defined as

$$F_n(x) \triangleq \frac{1}{N}\sum_{i=1}^N \mathbb{I}(x_i \leq x) \tag{2.44}$$

2.4 Some common continuous distributions

In this section we present some commonly used univariate (one-dimensional) continuous probability distributions.

2.4.1 Gaussian (normal) distribution

If $X \sim N(0,1)$,we say X follows a **standard normal** distribution.

The Gaussian distribution is the most widely used distribution in statistics. There are several reasons for this.

1. First, it has two parameters which are easy to interpret, and which capture some of the most basic properties of a distribution, namely its mean and variance.
2. Second,the central limit theorem (Section TODO) tells us that sums of independent random variables have an approximately Gaussian distribution, making it a good choice for modeling residual errors or noise.

[6] http://en.wikipedia.org/wiki/Empirical_distribution_function

Table 2.3: Summary of Gaussian distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	var$[X]$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	μ	μ	σ^2

3. Third, the Gaussian distribution makes the least number of assumptions (has maximum entropy), subject to the constraint of having a specified mean and variance, as we show in Section TODO; this makes it a good default choice in many cases.
4. Finally, it has a simple mathematical form, which results in easy to implement, but often highly effective, methods, as we will see.

See (Jaynes 2003, ch 7) for a more extensive discussion of why Gaussians are so widely used.

2.4.2 Student's t-distribution

Table 2.4: Summary of Student's t-distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	var$[X]$
$X \sim \mathcal{T}(\mu, \sigma^2, \nu)$	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\nu}\right)^2\right]$	μ	μ	$\frac{\nu\sigma^2}{\nu-2}$

where $\Gamma(x)$ is the gamma function:

$$\Gamma(x) \triangleq \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t \tag{2.45}$$

μ is the mean, $\sigma^2 > 0$ is the scale parameter, and $\nu > 0$ is called the **degrees of freedom**. See Figure 2.1 for some plots.

The variance is only defined if $\nu > 2$. The mean is only defined if $\nu > 1$.

As an illustration of the robustness of the Student distribution, consider Figure 2.2. We see that the Gaussian is affected a lot, whereas the Student distribution hardly changes. This is because the Student has heavier tails, at least for small ν(see Figure 2.1).

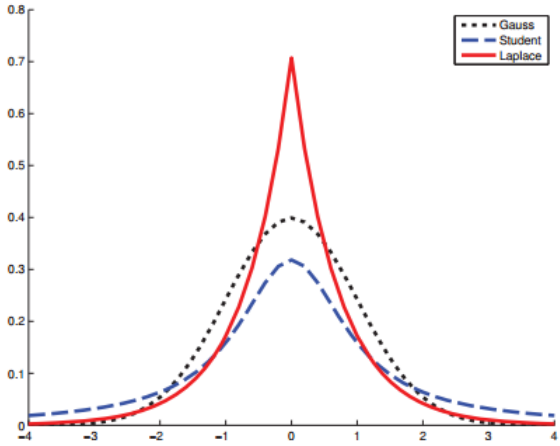
If $\nu = 1$, this distribution is known as the **Cauchy** or **Lorentz** distribution. This is notable for having such heavy tails that the integral that defines the mean does not converge.

To ensure finite variance, we require $\nu > 2$. It is common to use $\nu = 4$, which gives good performance in a range of problems (Lange et al. 1989). For $\nu \gg 5$, the Student distribution rapidly approaches a Gaussian distribution and loses its robustness properties.

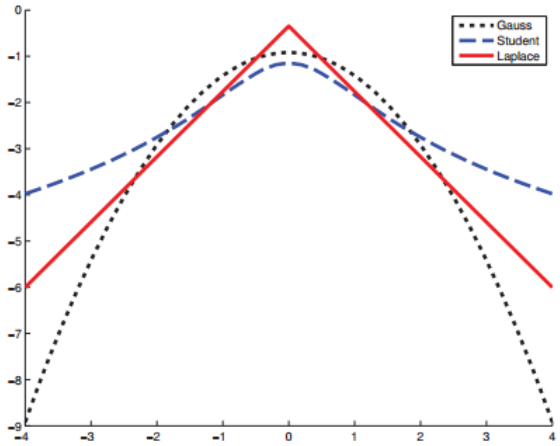
2.4.3 The Laplace distribution

Here μ is a location parameter and $b > 0$ is a scale parameter. See Figure 2.1 for a plot.

Its robustness to outliers is illustrated in Figure 2.2. It also put mores probability density at 0 than the Gaussian. This property is a useful way to encourage sparsity in a model, as we will see in Section TODO.



(a)



(b)

Fig. 2.1: (a) The pdfs for a $\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ and $Lap(0,1/\sqrt{2})$. The mean is 0 and the variance is 1 for both the Gaussian and Laplace. The mean and variance of the Student is undefined when $\nu = 1$.(b) Log of these pdfs. Note that the Student distribution is not log-concave for any parameter value, unlike the Laplace distribution, which is always log-concave (and log-convex...) Nevertheless, both are unimodal.

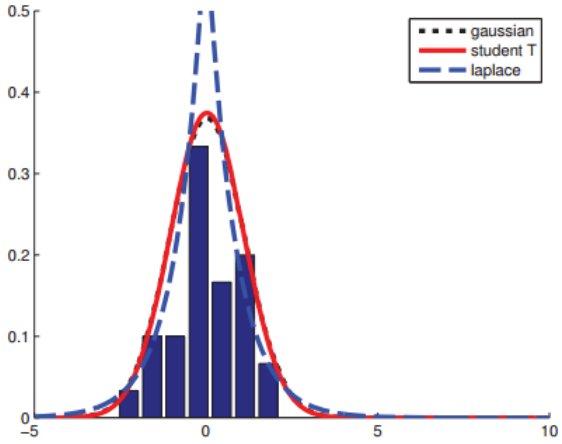
Table 2.5: Summary of Laplace distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	var$[X]$
$X \sim \text{Lap}(\mu, b)$	$\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$	μ	μ	$2b^2$

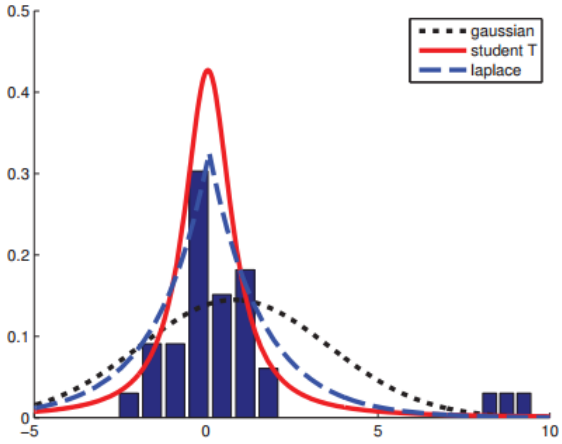
2.4.4 The gamma distribution

Table 2.6: Summary of gamma distribution

Written as	X	$f(x)$	$\mathbb{E}[X]$	mode	var$[X]$
$X \sim \text{Ga}(a,b)$	$x \in \mathbb{R}^+$	$\frac{b^a}{\Gamma(a)}x^{a-1}e^{-xb}$	$\frac{a}{b}$	$\frac{a-1}{b}$	$\frac{a}{b^2}$



(a)



(b)

Fig. 2.2: Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions.

Here $a > 0$ is called the shape parameter and $b > 0$ is called the rate parameter. See Figure 2.3 for some plots.

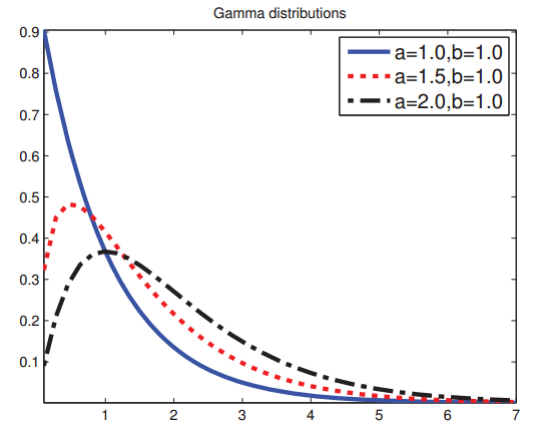
2.4.5 The beta distribution

Table 2.7: Summary of Beta distribution

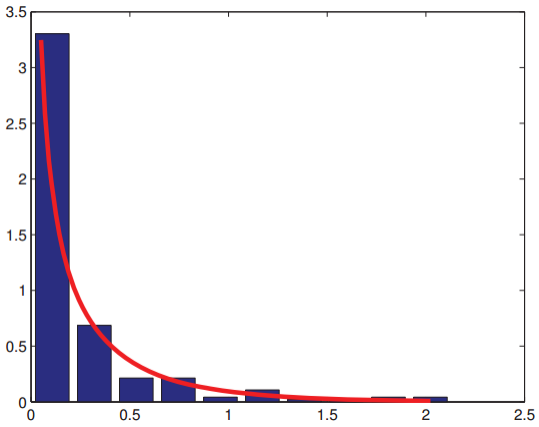
Name	Written as	X	$f(x)$	$\mathbb{E}[X]$	mode	var$[X]$
Beta distribution	$X \sim \text{Beta}(a,b)$	$x \in [0,1]$	$\frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$	$\frac{a}{a+b}$	$\frac{a-1}{a+b-2}$	$\frac{ab}{(a+b)^2(a+b+1)}$

Here $B(a,b)$is the beta function,

$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{2.46}$$



(a)



(b)

Fig. 2.3: Some Ga$(a,b=1)$ distributions. If $a \leq 1$, the mode is at 0, otherwise it is > 0.As we increase the rate b, we reduce the horizontal scale, thus squeezing everything leftwards and upwards. (b) An empirical pdf of some rainfall data, with a fitted Gamma distribution superimposed.

See Figure 2.4 for plots of some beta distributions. We require $a,b > 0$ to ensure the distribution is integrable (i.e., to ensure $B(a,b)$ exists). If $a = b = 1$, we get the uniform distirbution. If a and b are both less than 1, we get a bimodal distribution with spikes at 0 and 1; if a and b are both greater than 1, the distribution is unimodal.

2.4.6 Pareto distribution

Table 2.8: Summary of Pareto distribution

Name	Written as	X	$f(x)$	$\mathbb{E}[X]$	mode	var$[X]$
Pareto distribution	$X \sim \text{Pareto}(k,m)$	$x \geq m$	$km^k x^{-(k+1)} \mathbb{I}(x \geq m)$	$\frac{km}{k-1}$ if $k > 1$	m	$\frac{m^2 k}{(k-1)^2(k-2)}$ if $k > 2$

The **Pareto distribution** is used to model the distribution of quantities that exhibit **long tails**, also called **heavy tails**.

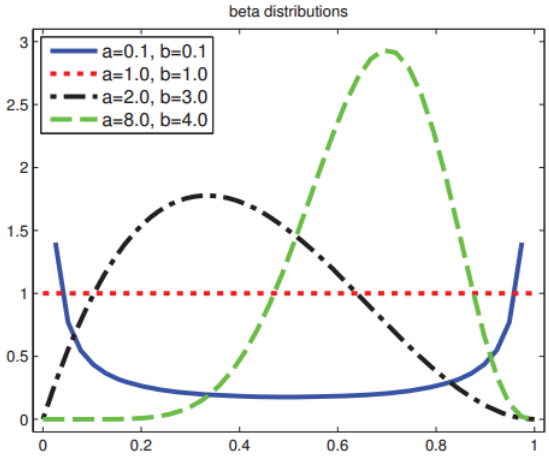
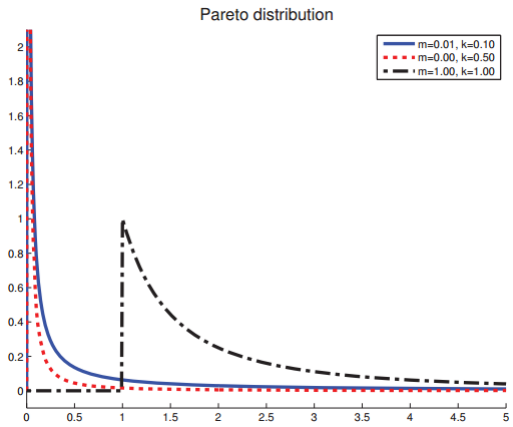
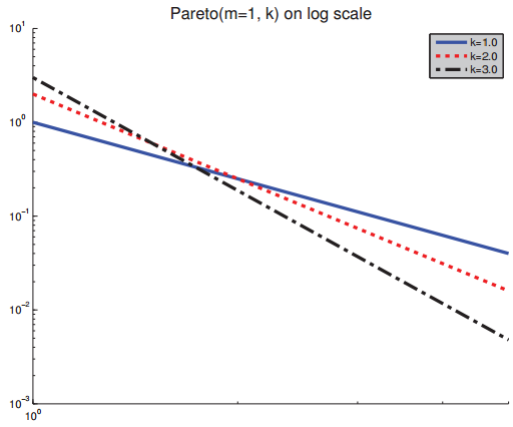


Fig. 2.4: Some beta distributions.

As $k \to \infty$, the distribution approaches $\delta(x-m)$. See Figure 2.5(a) for some plots. If we plot the distribution on a log-log scale, it forms a straight line, of the form $\log p(x) = a \log x + c$ for some constants a and c. See Figure 2.5(b) for an illustration (this is known as a **power law**).



(a)



(b)

Fig. 2.5: (a) The Pareto distribution $\text{Pareto}(x|m,k)$ for $m = 1$. (b) The pdf on a log-log scale.

2.5 Joint probability distributions

Given a **multivariate random variable** or **random vector** [7] $X \in \mathbb{R}^D$, the **joint probability distribution**[8] is a probability distribution that gives the probability that each of $X_1, X_2, \cdots, X_D$ falls in any particular range or discrete set of values specified for that variable. In the case of only two random variables, this is called a **bivariate distribution**, but the concept generalizes to any number of random variables, giving a **multivariate distribution**.

The joint probability distribution can be expressed either in terms of a **joint cumulative distribution function** or in terms of a **joint probability density function** (in the case of continuous variables) or **joint probability mass function** (in the case of discrete variables).

2.5.1 Covariance and correlation

Definition 2.6. The **covariance** between two rvs X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as

$$\begin{aligned} \text{cov}[X,Y] &\triangleq \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \tag{2.47}$$

Definition 2.7. If X is a D-dimensional random vector, its **covariance matrix** is defined to be the following symmetric, positive definite matrix:

$$\text{cov}[X] \triangleq \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right] \tag{2.48}$$

$$= \begin{pmatrix} \text{var}[X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_D] \\ \text{Cov}[X_2, X_1] & \text{var}[X_2] & \cdots & \text{Cov}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_D, X_1] & \text{Cov}[X_D, X_2] & \cdots & \text{var}[X_D] \end{pmatrix} \tag{2.49}$$

Definition 2.8. The (Pearson) **correlation coefficient** between X and Y is defined as

$$\text{corr}[X,Y] \triangleq \frac{\text{Cov}[X,Y]}{\sqrt{\text{var}[X], \text{var}[Y]}} \tag{2.50}$$

A **correlation matrix** has the form

$$\mathbf{R} \triangleq \begin{pmatrix} \text{corr}[X_1, X_1] & \text{corr}[X_1, X_2] & \cdots & \text{corr}[X_1, X_D] \\ \text{corr}[X_2, X_1] & \text{corr}[X_2, X_2] & \cdots & \text{corr}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}[X_D, X_1] & \text{corr}[X_D, X_2] & \cdots & \text{corr}[X_D, X_D] \end{pmatrix} \tag{2.51}$$

The correlation coefficient can viewed as a degree of linearity between X and Y, see Figure 2.6.

Uncorrelated does not imply independent. For example, let $X \sim U(-1,1)$ and $Y = X^2$. Clearly Y is dependent on X(in fact, Y is uniquely determined by X), yet one can show that $\text{corr}[X,Y] = 0$. Some striking examples of this fact are shown in Figure 2.6. This shows several data sets where there is clear dependence between X and Y, and yet the correlation coefficient is 0. A more general measure of dependence between random variables is mutual information, see Section TODO.

[7] http://en.wikipedia.org/wiki/Multivariate_random_variable

[8] http://en.wikipedia.org/wiki/Joint_probability_distribution

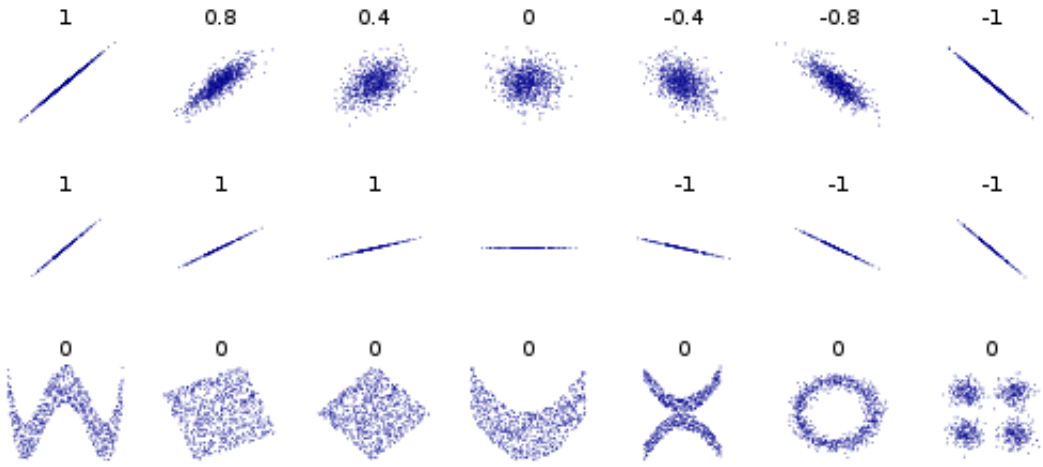


Fig. 2.6: Several sets of (x,y) points, with the Pearson correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.Source:http://en.wikipedia.org/wiki/Correlation

2.5.2 Multivariate Gaussian distribution

The **multivariate Gaussian** or **multivariate normal**(MVN) is the most widely used joint probability density function for continuous variables. We discuss MVNs in detail in Chapter 4; here we just give some definitions and plots.

The pdf of the MVN in D dimensions is defined by the following:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) \triangleq \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{2.52}$$

where $\boldsymbol{\mu} = \mathbb{E}[X] \in \mathbb{R}^D$ is the mean vector, and $\Sigma = \mathrm{Cov}[X]$ is the $D \times D$ covariance matrix. The normalization constant $(2\pi)^{D/2}|\Sigma|^{1/2}$ just ensures that the pdf integrates to 1.

Figure 2.7 plots some MVN densities in 2d for three different kinds of covariance matrices. A full covariance matrix has A $D(D+1)/2$ parameters (we divide by 2 since Σ is symmetric). A diagonal covariance matrix has D parameters, and has 0s in the off-diagonal terms. A spherical or isotropic covariance,$\Sigma = \sigma^2\boldsymbol{I}_D$, has one free parameter.

2.5.3 Multivariate Student's t-distribution

A more robust alternative to the MVN is the multivariate Student's t-distribution, whose pdf is given by

$$\begin{aligned}
&\mathcal{T}(x|\boldsymbol{\mu},\Sigma,\nu) \\
&\triangleq \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})}\frac{|\Sigma|^{-\frac{1}{2}}}{(\nu\pi)^{\frac{D}{2}}}\left[1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{\nu+D}{2}} &(2.53)\\
&= \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})}\frac{|\Sigma|^{-\frac{1}{2}}}{(\nu\pi)^{\frac{D}{2}}}\left[1+(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{\nu+D}{2}} &(2.54)
\end{aligned}$$

where Σ is called the scale matrix (since it is not exactly the covariance matrix) and $\boldsymbol{V} = \nu\Sigma$. This has fatter tails than a Gaussian. The smaller ν is, the fatter the tails. As $\nu \to \infty$, the distribution tends towards a Gaussian. The distribution

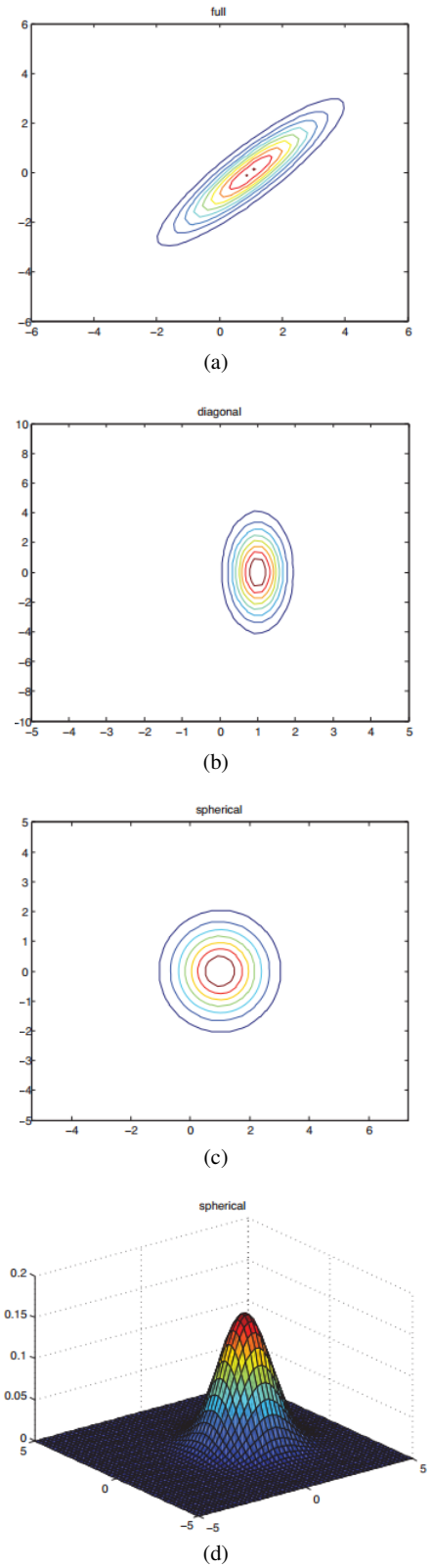


Fig. 2.7: We show the level sets for 2d Gaussians. (a) A full covariance matrix has elliptical contours.(b) A diagonal covariance matrix is an axis aligned ellipse. (c) A spherical covariance matrix has a circular shape. (d) Surface plot for the spherical Gaussian in (c).

has the following properties

$$\text{mean} = \boldsymbol{\mu} \text{ , mode} = \boldsymbol{\mu} \text{ , Cov} = \frac{\nu}{\nu - 2}\Sigma \tag{2.55}$$

2.5.4 Dirichlet distribution

A multivariate generalization of the beta distribution is the **Dirichlet distribution**, which has support over the probability simplex, defined by

$$S_K = \left\{ \boldsymbol{x} : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1 \right\} \tag{2.56}$$

The pdf is defined as follows:

$$\text{Dir}(\boldsymbol{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\boldsymbol{x} \in S_K) \tag{2.57}$$

where $B(\alpha_1, \alpha_2, \cdots, \alpha_K)$ is the natural generalization of the beta function to K variables:

$$B(\alpha) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \text{ where } \alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k \tag{2.58}$$

Figure 2.8 shows some plots of the Dirichlet when $K = 3$, and Figure 2.9 for some sampled probability vectors. We see that α_0 controls the strength of the distribution (how peaked it is), and thekcontrol where the peak occurs. For example, $\text{Dir}(1,1,1)$ is a uniform distribution, $\text{Dir}(2,2,2)$ is a broad distribution centered at $(1/3, 1/3, 1/3)$, and $\text{Dir}(20,20,20)$ is a narrow distribution centered at $(1/3, 1/3, 1/3)$.If $\alpha_k < 1$ for all k, we get spikes at the corner of the simplex.

For future reference, the distribution has these properties

$$\mathbb{E}(x_k) = \frac{\alpha_k}{\alpha_0}, \text{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}, \text{var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \tag{2.59}$$

2.6 Transformations of random variables

If $\boldsymbol{x} \sim P()$ is some random variable, and $\boldsymbol{y} = f(\boldsymbol{x})$, what is the distribution of Y? This is the question we address in this section.

2.6.1 Linear transformations

Suppose $g()$ is a linear function:

$$g(\boldsymbol{x}) = A\boldsymbol{x} + b \tag{2.60}$$

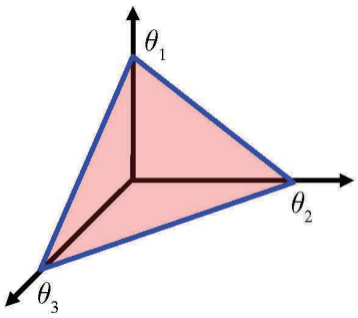
First, for the mean, we have

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[A\boldsymbol{x} + b] = A\mathbb{E}[\boldsymbol{x}] + b \tag{2.61}$$

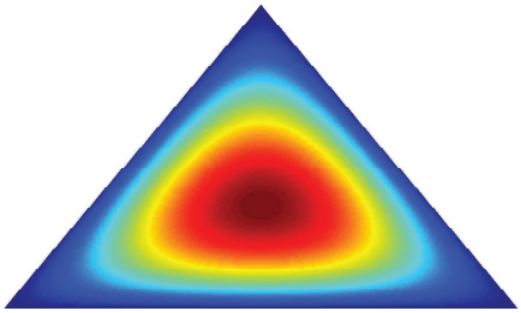
this is called the **linearity of expectation**.

For the covariance, we have

$$\text{Cov}[\boldsymbol{y}] = \text{Cov}[A\boldsymbol{x} + b] = A\Sigma A^T \tag{2.62}$$

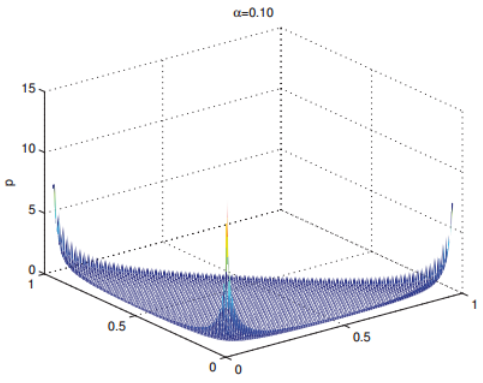


(a)



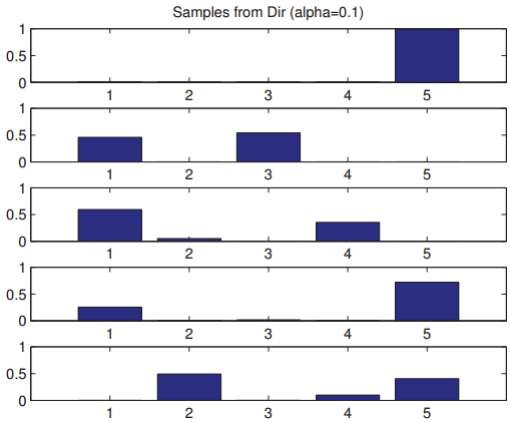
(b)

(c)

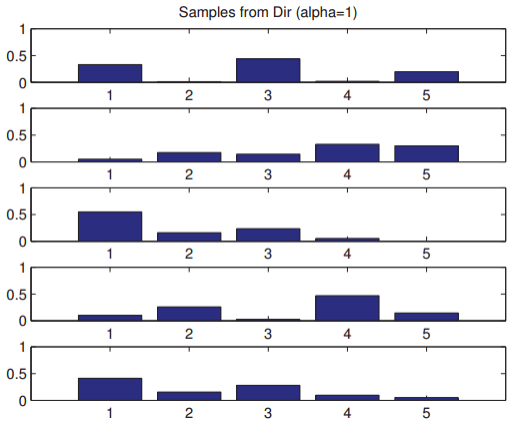


(d)

Fig. 2.8: (a) The Dirichlet distribution when $K = 3$ defines a distribution over the simplex, which can be represented by the triangular surface. Points on this surface satisfy $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{K} \theta_k = 1$. (b) Plot of the Dirichlet density when $\boldsymbol{\alpha} = (2,2,2)$. (c) $\boldsymbol{\alpha} = (20,2,2)$.



(a) $\boldsymbol{\alpha} = (0.1, \cdots, 0.1)$. This results in very sparse distributions, with many 0s.



(b) $\boldsymbol{\alpha} = (1, \cdots, 1)$. This results in more uniform (and dense) distributions.

Fig. 2.9: Samples from a 5-dimensional symmetric Dirichlet distribution for different parameter values.

2.6.2 General transformations

If X is a discrete rv, we can derive the pmf for y by simply summing up the probability mass for all the xs such that $f(x) = y$:

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x) \tag{2.63}$$

If X is continuous, we cannot use Equation 2.63 since $p_X(x)$ is a density, not a pmf, and we cannot sum up densities. Instead, we work with cdfs, and write

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \int_{g(X) \leq y} f_X(x)\mathrm{d}x \tag{2.64}$$

We can derive the pdf of Y by differentiating the cdf:

$$f_Y(y) = f_X(x)|\frac{dx}{dy}| \tag{2.65}$$

This is called **change of variables** formula. We leave the proof of this as an exercise.

For example, suppose $X \sim U(1,1)$, and $Y = X^2$. Then $p_Y(y) = \frac{1}{2} y^{-\frac{1}{2}}$.

2.6.2.1 Multivariate change of variables *

Let f be a function $f : \mathbb{R}^n \to \mathbb{R}^n$, and let $\boldsymbol{y} = f(\boldsymbol{x})$. Then its Jacobian matrix $\boldsymbol{J}$ is given by

$$\boldsymbol{J}_{\boldsymbol{x}\to\boldsymbol{y}} \triangleq \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \triangleq \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \tag{2.66}$$

$|\det(\boldsymbol{J})|$ measures how much a unit cube changes in volume when we apply f.

If f is an invertible mapping, we can define the pdf of the transformed variables using the Jacobian of the inverse mapping $\boldsymbol{y} \to \boldsymbol{x}$:

$$p_y(\boldsymbol{y}) = p_x(\boldsymbol{x})|\det(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}})| = p_x(\boldsymbol{x})|\det(\boldsymbol{J}_{\boldsymbol{y}\to\boldsymbol{x}})| \tag{2.67}$$

2.6.3 Central limit theorem

Given N random variables $X_1, X_2, \cdots, X_N$, each variable is **independent and identically distributed**[9](**iid** for short), and each has the same mean μ and variance σ^2, then

$$\frac{\sum_{i=1}^{n} X_i - N\mu}{\sqrt{N}\sigma} \sim \mathcal{N}(0,1) \tag{2.68}$$

this can also be written as

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1) \quad \text{, where } \bar{X} \triangleq \frac{1}{N}\sum_{i=1}^{n} X_i \tag{2.69}$$

2.7 Monte Carlo approximation

In general, computing the distribution of a function of an rv using the change of variables formula can be difficult. One simple but powerful alternative is as follows. First we generate S samples from the distribution, call them $x_1, \cdots, x_S$. (There are many ways to generate such samples; one popular method, for high dimensional distributions, is called Markov chain Monte Carlo or MCMC; this will be explained in Chapter TODO.) Given the samples, we can approximate the distribution of $f(X)$ by using the empirical distribution of $\{f(x_s)\}_{s=1}^{S}$. This is called a **Monte Carlo approximation**[10], named after a city in Europe known for its plush gambling casinos.

We can use Monte Carlo to approximate the expected value of any function of a random variable. We simply draw samples, and then compute the arithmetic mean of the function applied to the samples. This can be written as follows:

$$\mathbb{E}[g(X)] = \int g(x)p(x)\mathrm{d}x \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s) \tag{2.70}$$

where $x_s \sim p(X)$.

[9] http://en.wikipedia.org/wiki/Independent_identically_distributed

[10] http://en.wikipedia.org/wiki/Monte_Carlo_method

This is called **Monte Carlo integration**[11], and has the advantage over numerical integration (which is based on evaluating the function at a fixed grid of points) that the function is only evaluated in places where there is non-negligible probability.

2.8 Information theory

2.8.1 Entropy

The entropy of a random variable X with distribution p, denoted by $\mathbb{H}(X)$ or sometimes $\mathbb{H}(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) \tag{2.71}$$

Usually we use log base 2, in which case the units are called **bits**(short for binary digits). If we use log base e , the units are called **nats**.

The discrete distribution with maximum entropy is the uniform distribution (see Section XXX for a proof). Hence for a K-ary random variable, the entropy is maximized if $p(x=k) = 1/K$; in this case, $\mathbb{H}(X) = \log_2 K$.

Conversely, the distribution with minimum entropy (which is zero) is any **delta-function** that puts all its mass on one state. Such a distribution has no uncertainty.

2.8.2 KL divergence

One way to measure the dissimilarity of two probability distributions, p and q , is known as the **Kullback-Leibler divergence**(**KL divergence**)or **relative entropy**. This is defined as follows:

$$\mathbb{KL}(P||Q) \triangleq \sum_x p(x)\log_2 \frac{p(x)}{q(x)} \tag{2.72}$$

where the sum gets replaced by an integral for pdfs[12]. The KL divergence is only defined if P and Q both sum to 1 and if $q(x) = 0$ implies $p(x) = 0$ for all x(absolute continuity). If the quantity $0\ln 0$ appears in the formula, it is interpreted as zero because $\lim_{x\to 0} x\ln x$. We can rewrite this as

$$\begin{aligned}\mathbb{KL}(p||q) &\triangleq \sum_x p(x)\log_2 p(x) - \sum_{k=1}^{K} p(x)\log_2 q(x) \\ &= \mathbb{H}(p) - \mathbb{H}(p,q)\end{aligned} \tag{2.73}$$

where $\mathbb{H}(p,q)$ is called the **cross entropy**,

$$\mathbb{H}(p,q) = \sum_x p(x)\log_2 q(x) \tag{2.74}$$

One can show (Cover and Thomas 2006) that the cross entropy is the average number of bits needed to encode data coming from a source with distribution p when we use model q to define our codebook. Hence the regular entropy $\mathbb{H}(p) = \mathbb{H}(p,p)$, defined in section §2.8.1,is the expected number of bits if we use the true model, so the KL divergence

[11] http://en.wikipedia.org/wiki/Monte_Carlo_integration

[12] The KL divergence is not a distance, since it is asymmetric. One symmetric version of the KL divergence is the **Jensen-Shannon divergence**, defined as $JS(p_1,p_2) = 0.5\mathbb{KL}(p_1||q) + 0.5\mathbb{KL}(p_2||q)$,where $q = 0.5p_1 + 0.5p_2$

is the diference between these. In other words, the KL divergence is the average number of *extra* bits needed to encode the data, due to the fact that we used distribution q to encode the data instead of the true distribution p.

The extra number of bits interpretation should make it clear that $\mathbb{KL}(p||q) \geq 0$, and that the KL is only equal to zero if $q = p$. We now give a proof of this important result.

Theorem 2.1. ***(Information inequality)*** $\mathbb{KL}(p||q) \geq 0$ *with equality iff* $p = q$.

One important consequence of this result is that *the discrete distribution with the maximum entropy is the uniform distribution*.

2.8.3 Mutual information

Definition 2.9. Mutual information or **MI**, is defined as follows:

$$\begin{aligned} \mathbb{I}(X;Y) &\triangleq \mathbb{KL}(P(X,Y)||P(X)P(X)) \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{2.75}$$

We have $\mathbb{I}(X;Y) \geq 0$ with equality if $P(X,Y) = P(X)P(Y)$. That is, the MI is zero if the variables are independent.

To gain insight into the meaning of MI, it helps to re-express it in terms of joint and conditional entropies. One can show that the above expression is equivalent to the following:

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) \tag{2.76}$$
$$= \mathbb{H}(Y) - \mathbb{H}(Y|X) \tag{2.77}$$
$$= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \tag{2.78}$$
$$= \mathbb{H}(X,Y) - \mathbb{H}(X|Y) - \mathbb{H}(Y|X) \tag{2.79}$$

where $\mathbb{H}(X)$ and $\mathbb{H}(Y)$ are the **marginal entropies**, $\mathbb{H}(X|Y)$ and $\mathbb{H}(Y|X)$ are the **conditional entropies**, and $\mathbb{H}(X,Y)$ is the **joint entropy** of X and Y, see Fig. 2.10[13].

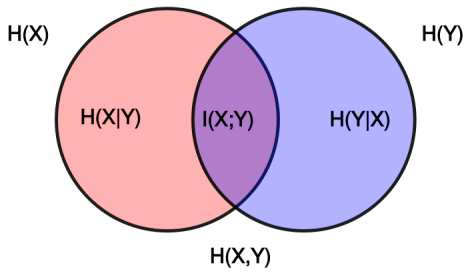


Fig. 2.10: Individual $\mathbb{H}(X), \mathbb{H}(Y)$, joint $\mathbb{H}(X,Y)$, and conditional entropies for a pair of correlated subsystems X, Y with mutual information $\mathbb{I}(X;Y)$.

Intuitively, we can interpret the MI between X and Y as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X.

A quantity which is closely related to MI is the **pointwise mutual information** or **PMI**. For two events (not random variables) x and y, this is defined as

[13] http://en.wikipedia.org/wiki/Mutual_information

$$PMI(x,y) \triangleq \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{2.80}$$

This measures the discrepancy between these events occuring together compared to what would be expected by chance. Clearly the MI of X and Y is just the expected value of the PMI. Interestingly, we can rewrite the PMI as follows:

$$PMI(x,y) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{2.81}$$

This is the amount we learn from updating the prior $p(x)$ into the posterior $p(x|y)$, or equivalently, updating the prior $p(y)$ into the posterior $p(y|x)$.

Appendix A
Optimization methods

A.1 Convexity

Definition A.1. (**Convex set**) We say aset $\mathcal{S}$ is convex if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{S}$, we have

$$\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2 \in \mathcal{S}, \forall \lambda \in [0,1] \tag{A.1}$$

Definition A.2. (**Convex function**) A function $f(\boldsymbol{x})$ is called convex if its **epigraph**(the set of points above the function) defines a convex set. Equivalently, a function $f(\boldsymbol{x})$ is called convex if it is defined on a convex set and if, for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{S}$, and any $\lambda \in [0,1]$, we have

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \tag{A.2}$$

Definition A.3. A function $f(\boldsymbol{x})$ is said to be **strictly convex** if the inequality is strict

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) < \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \tag{A.3}$$

Definition A.4. A function $f(\boldsymbol{x})$ is said to be (strictly) **concave** if $-f(\boldsymbol{x})$ is (strictly) convex.

Theorem A.1. *If $f(x)$ is twice differentiable on $[a,b]$ and $f''(x) \geq 0$ on $[a,b]$ then $f(x)$ is convex on $[a,b]$.*

Proposition A.1. $\log(x)$ *is strictly convex on* $(0,\infty)$.

Intuitively, a (strictly) convex function has a bowl shape, and hence has a unique global minimum x^* corresponding to the bottom of the bowl. Hence its second derivative must be positive everywhere, $\frac{\mathrm{d}^2}{\mathrm{d}x^2} f(x) > 0$. A twice-continuously differentiable, multivariate function f is convex iff its Hessian is positive definite for all $\boldsymbol{x}$. In the machine learning context, the function f often corresponds to the NLL.

Models where the NLL is convex are desirable, since this means we can always find the globally optimal MLE. We will see many examples of this later in the book. However, many models of interest will not have concave likelihoods. In such cases, we will discuss ways to derive locally optimal parameter estimates.

A.2 Gradient descent

A.2.1 Stochastic gradient descent

A.2.2 Batch gradient descent

A.2.3 Line search

The **line search**[1] approach first finds a descent direction along which the objective function f will be reduced and then computes a step size that determines how far $\boldsymbol{x}$ should move along that direction. The descent direction can be computed by various methods, such as gradient descent(Section A.2), Newton's method(Section A.4) and Quasi-Newton method(Section A.5). The step size can be determined either exactly or inexactly.

[1] http://en.wikipedia.org/wiki/Line_search

input : Training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | i = 1 : N\}$
output: A linear model: $y_i = \boldsymbol{\theta}^T \boldsymbol{x}$
$\boldsymbol{w} \leftarrow 0$; $b \leftarrow 0$; $k \leftarrow 0$;
while *no mistakes made within the for loop* **do**
  for $i \leftarrow 1$ **to** N **do**
    if $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \leq 0$ **then**
      $\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta y_i \boldsymbol{x}_i$;
      $b \leftarrow b + \eta y_i$;
      $k \leftarrow k + 1$;
    end
  end
end

Algorithm 1: Stochastic gradient descent

A.2.4 Momentum term

A.3 Lagrange duality

A.3.1 Primal form

Consider the following, which we'll call the **primal** optimization problem:

$$xyz \tag{A.4}$$

A.3.2 Dual form

A.4 Newton's method

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_k) + \boldsymbol{g}_k^T (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_k)^T \boldsymbol{H}_k (\boldsymbol{x} - \boldsymbol{x}_k)$$

$$\text{where } \boldsymbol{g}_k \triangleq \boldsymbol{g}(\boldsymbol{x}_k) = f'(\boldsymbol{x}_k), \boldsymbol{H}_k \triangleq \boldsymbol{H}(\boldsymbol{x}_k),$$

$$\boldsymbol{H}(\boldsymbol{x}) \triangleq \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{D \times D} \quad \text{(Hessian matrix)}$$

$$f'(\boldsymbol{x}) = \boldsymbol{g}_k + \boldsymbol{H}_k (\boldsymbol{x} - \boldsymbol{x}_k) = 0 \Rightarrow \tag{A.5}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{H}_k^{-1} \boldsymbol{g}_k \tag{A.6}$$

Initialize $\boldsymbol{x}_0$
while *(!convergency)* **do**
  Evaluate $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$
  Evaluate $\boldsymbol{H}_k = \nabla^2 f(\boldsymbol{x}_k)$
  $\boldsymbol{d}_k = -\boldsymbol{H}_k^{-1} \boldsymbol{g}_k$
  Use line search to find step size η_k along $\boldsymbol{d}_k$
  $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \eta_k \boldsymbol{d}_k$
end

Algorithm 2: Newtons method for minimizing a strictly convex function

A.5 Quasi-Newton method

From Equation A.5 we can infer out the **quasi-Newton condition** as follows:

$$\begin{aligned} f'(\boldsymbol{x}) - \boldsymbol{g}_k &= \boldsymbol{H}_k(\boldsymbol{x} - \boldsymbol{x}_k) \\ \boldsymbol{g}_{k-1} - \boldsymbol{g}_k &= \boldsymbol{H}_k(\boldsymbol{x}_{k-1} - \boldsymbol{x}_k) \Rightarrow \\ \boldsymbol{g}_k - \boldsymbol{g}_{k-1} &= \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \\ \boldsymbol{g}_{k+1} - \boldsymbol{g}_k &= \boldsymbol{H}_{k+1}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) \quad \text{(quasi-Newton condition)} \end{aligned} \tag{A.7}$$

The idea is to replace $\boldsymbol{H}_k^{-1}$ with a approximation $\boldsymbol{B}_k$, which satisfies the following properties:

1. $\boldsymbol{B}_k$ must be symmetric
2. $\boldsymbol{B}_k$ must satisfies the quasi-Newton condition, i.e., $\boldsymbol{g}_{k+1} - \boldsymbol{g}_k = \boldsymbol{B}_{k+1}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$.
   Let $\boldsymbol{y}_k = \boldsymbol{g}_{k+1} - \boldsymbol{g}_k$, $\boldsymbol{\delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$, then
   $$\boldsymbol{B}_{k+1}\boldsymbol{y}_k = \boldsymbol{\delta}_k \tag{A.8}$$
3. Subject to the above, $\boldsymbol{B}_k$ should be as close as possible to $\boldsymbol{B}_{k-1}$.

Note that we did not require that $\boldsymbol{B}_k$ be positive definite. That is because we can show that it must be positive definite if $\boldsymbol{B}_{k-1}$ is. Therefore, as long as the initial Hessian approximation $\boldsymbol{B}_0$ is positive definite, all $\boldsymbol{B}_k$ are, by induction.

A.5.1 DFP

Updating rule:

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \boldsymbol{P}_k + \boldsymbol{Q}_k \tag{A.9}$$

From Equation A.8 we can get

$$\boldsymbol{B}_{k+1}\boldsymbol{y}_k = \boldsymbol{B}_k\boldsymbol{y}_k + \boldsymbol{P}_k\boldsymbol{y}_k + \boldsymbol{Q}_k\boldsymbol{y}_k = \boldsymbol{\delta}_k$$

To make the equation above establish, just let

$$\begin{aligned} \boldsymbol{P}_k\boldsymbol{y}_k &= \boldsymbol{\delta}_k \\ \boldsymbol{Q}_k\boldsymbol{y}_k &= -\boldsymbol{B}_k\boldsymbol{y}_k \end{aligned}$$

In DFP algorithm, $\boldsymbol{P}_k$ and $\boldsymbol{Q}_k$ are

$$\boldsymbol{P}_k = \frac{\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T}{\boldsymbol{\delta}_k^T\boldsymbol{y}_k} \tag{A.10}$$

$$\boldsymbol{Q}_k = -\frac{\boldsymbol{B}_k\boldsymbol{y}_k\boldsymbol{y}_k^T\boldsymbol{B}_k}{\boldsymbol{y}_k^T\boldsymbol{B}_k\boldsymbol{y}_k} \tag{A.11}$$

A.5.2 BFGS

Use $\boldsymbol{B}_k$ as a approximation to $\boldsymbol{H}_k$, then the quasi-Newton condition becomes

$$\boldsymbol{B}_{k+1}\boldsymbol{\delta}_k = \boldsymbol{y}_k \tag{A.12}$$

The updating rule is similar to DFP, but $\boldsymbol{P}_k$ and $\boldsymbol{Q}_k$ are different. Let

$$\boldsymbol{P}_k \boldsymbol{\delta}_k = \boldsymbol{y}_k$$
$$\boldsymbol{Q}_k \boldsymbol{\delta}_k = -\boldsymbol{B}_k \boldsymbol{\delta}_k$$

Then

$$\boldsymbol{P}_k = \frac{\boldsymbol{y}_k \boldsymbol{y}_k^T}{\boldsymbol{y}_k^T \boldsymbol{\delta}_k} \tag{A.13}$$

$$\boldsymbol{Q}_k = -\frac{\boldsymbol{B}_k \boldsymbol{\delta}_k \boldsymbol{\delta}_k^T \boldsymbol{B}_k}{\boldsymbol{\delta}_k^T \boldsymbol{B}_k \boldsymbol{\delta}_k} \tag{A.14}$$

A.5.3 Broyden

Broyden's algorithm is a linear combination of DFP and BFGS.

Glossary

Use the template *glossary.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your glossary in the Springer layout.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.