Universiteit
Antwerpen

Faculteit Wetenschappen
Departement Fysica

Extension of information geometry for modelling non-statistical systems

Uitbreiding van de informatiemeetkunde voor het modelleren van niet-statistische systemen

Proefschrift voorgelegd tot het behalen van de graad van

doctor in de wetenschappen

aan de Universiteit Antwerpen, te verdedigen door

Ben Anthonis

Promotoren
Prof. dr. Jan Naudts
Prof. dr. Jacques Tempere

Antwerpen, 2014

Dankwoord

Hoewel het mijn naam is die in grote letters op de kaft van dit proefschrift pronkt, is ook dit doctoraat een onderneming geweest waar heel wat mensen rechtstreeks of onrechtstreeks mee te maken hadden. Daarom wil ik hen hier bedanken.

De eerste in deze lange lijst is mijn promotor, prof. Jan Naudts. Deze verhandeling is reeds de derde die ik schrijf onder zijn begeleiding. Het is dan ook van hem dat ik het meest geleerd heb over de wiskundige natuurkunde. Hoewel onze meningen over hoe een bepaald vraagstuk moest worden aangepakt meer dan eens volledig verschilden, heeft hij mij toch het vertrouwen en de vrijheid geschonken om mijn eigen ding te doen. Ook stond hij me telkens toe om tijd te investeren in het onderwijs, of het nu als onderwijsassistent was of als ombudspersoon. Een dergelijke promotor is niet alle doctorandi gegeven.

Mijn co-promotor, prof. Jacques Tempere, was niet even nauw betrokken bij dit onderzoek als mijn promotor, maar toch was ook zijn steun van onmiskenbaar belang. Zo was hij het die me de kans gaf om aan een doctoraat te beginnen wanneer de groep Wiskundige Natuurkunde hier de fondsen niet voor had. De aanstekelijkheid van zijn enthousiasme bleek niet beperkt te zijn tot de fysica. Hij vormt dan ook het levende bewijs dat de universiteit geen ivoren toren is en nog steeds in staat is om haar studenten meer te leren dan alleen maar vakspecifieke kennis.

For other interesting discussions I would like to thank prof. Hiroki Suyari, prof. Atsumi Ohara, prof. Hiroshi Matsuzoe and prof. Tatsuaki Wada, who have honoured prof. Naudts and me with work visits. Also the members of my jury, prof. Giovanni Pistone, prof. Christian Maes, prof. Sandra Van Aert and prof. Etienne Goovaerts, whose remarks were very useful in preparing the final version of this dissertation.

De collega-studenten en -doctoraatsstudenten die ik de afgelopen jaren heb gehad vormen intussen een redelijk grote groep. Toch zijn velen onder hen vrienden geworden. In de hoop niemand te vergeten zijn dit: Amy, Anne, Bert, Bert, Bob, Damiaan, Dries, Enya, Erik, Giovanni, Jeroen, Katleen, Katrijn, Maarten, Mathias, Nick, Nick, Selma, Stijn, Sven, Thomas, Tobias, Tyché, Wim, Winny en Wouter. Geert en Inge, wiens masterthesis ik heb mogen begeleiden, horen mee in deze categorie thuis. Ook kunnen prof. Michiel Wouters, prof. Dirk Callebaut en Onur Umucalılar hier niet ongenoemd gelaten worden. Drie vermeldingen zijn weggelegd voor vrienden die als erecollega's zouden kunnen gelden: Asbjørn, Brent en Yannick.

Vanzelfsprekend zijn er nog andere mensen met wie ik heb samengewerkt de afgelopen jaren en daarom verdienen om in dit dankwoord voor te komen.

Dit zijn: prof. Dirk Lamoen en prof. Jan Sijbers, voor wie ik oefeningenlessen heb mogen verzorgen; Hilde en Gwendolyn, voor het samenwerken daarbij; prof. Paul Scheunders, prof. Joke Hadermann en prof. Sabine Van Doorslaer van wiens onderwijs-, examen- en zelfevaluatie-commissies ik deel mocht uitmaken; Werner Peeters, voor de interessante discussies over wiskunde, onderwijs en de combinatie van beiden; Brigitte Brusselmans en Pieter Caris, voor al de keren dat ik op hen kon rekenen als ombudspersoon; Annick Van Son, voor de college- en examenroosters die ze telkens voor de fysicastudenten heeft gemaakt; Lieff Serrien en Hilde Evans om mij de weg te wijzen in het doolhof van de universitaire administratie.
Dan zijn er ook nog collega's met wie ik niet heb samengewerkt in de strikte betekenis van het woord maar die toch niet onbelangrijk zijn: (opnieuw) Lieff en alle andere mensen die hebben geholpen aan de organisatie van de jaarlijkse departementsreceptie; Quinten, die de verantwoordelijkheid daarvoor heeft overgenomen; (opnieuw) Nick Verhelst, die zijn tweede plaatsje hier kan opeisen door het verderzetten van mijn taak als ombudspersoon; Eveline Jacques, Els Grieten, Dirk Valkenborg en prof. Bart Goethals, die de anders zo saaie trein- en busritten telkens een pak plezieriger maakten; Sonja en Virginie, voor de goedemorgen waarmee ik de afgelopen vier jaar mijn dag mocht beginnen en voor de belangrijke maar soms ondankbare taak van het proper houden van onze kantoren.
De allergrootste groep van mensen die ik hier aan bod wil laten komen is die van de honderden studenten die "mijn" oefeningenlessen hebben bevolkt. Ik heb me geweldig geamuseerd met het lesgeven aan jullie. Een speciale vermelding behoort hier toe aan Karen, die ik sinds de afgelopen herfst tot mijn beste vrienden mag rekenen en wie me telkens een *pep talk* gaf wanneer de laatste loodjes begonnen te wegen.
Wie zeker niet vergeten mag worden in dit dankwoord is mijn familie. In de eerste plaats houdt dit mijn ouders in: Willy en Gerda—moeke en voke voor mij—voor alle steun en aanmoediging, voor de interesse in mijn onderzoek, voor het geduld dat zij met mij hadden toen ik nog maar eens overliep van enthousiasme en natuurlijk ook voor alle dingen die ouders doen voor hun kinderen en waarvoor die laatsten hen nooit genoeg kunnen bedanken. Ook zonder Va Stanne en mijn nonkel (die ik niet zo noem) Dominiek, zou dit dankwoord niet volledig kunnen zijn.
Ten slotte is er nog één iemand die hier niet vermeld is. Iemand die misschien nog harder naar het eindresultaat zou hebben uitgekeken dan ikzelf en die dolgraag aanwezig zou geweest zijn op de verdediging van dit werk. Het heeft echter niet mogen zijn. Daarom, Moe Simone, draag ik deze thesis op aan jou.

Nederlandstalige samenvatting

In deze doctoraatsverhandeling wordt een abstract meetkundig formalisme ontwikkeld voor het modelleren van data die zeer algemeen mogen zijn. Dit wiskundig kader werd het "data set model formalisme" genoemd en is geïnspireerd op de informatiemeetkunde. Het modelleren—of fitten—gebeurt met behulp van een divergentiefunctie: een veralgemeende afstandsmaat waarvan de relatieve entropie waarschijnlijk het bekendste voorbeeld is. Door te eisen dat deze modellen een differentiaalmeetkundige variëteit vormen kunnen zij worden uitgerust met een meetkundige structuur dewelke volgt uit de divergentiefunctie. Het belang van deze structuur is dat zij toestaat de belangrijkste eigenschappen van het modelleringsproces kwantitatief te beschrijven. Centraal hierin staat de zogenaamde Hessiaanse structuur, die het mogelijk maakt de Riemanniaanse metriek van de modelvariëteit af te leiden uit een familie van scalaire functies. Dit vereist de keuze van een vlakke affiene connectie, waarvoor tevens een constructie uit de divergentiefunctie gegeven wordt.

Het data set model formalisme biedt een aantal voordelen ten opzichte van bestaande modelleringstechnieken. Het belangrijkste daarvan is de grote wiskundige flexibiliteit, die te danken is aan het gebruik van differentiaalmeetkunde. Vanuit theoretisch oogpunt is vooral de mogelijkheid om een breed scala aan verschillende modelleringsproblemen in eenzelfde kader onder te brengen interessant. Zo omvat het formalisme de informatiemeetkunde maar eveneens regressiemodellen en elementen van de kwantumstatistiek. Ook voor praktische toepassingen is dit werk interessant. De reden hiervoor is dat de ontwikkelde technieken toestaan gegeven data via meetkundige methodes te modelleren, zelfs wanneer de modellen kwalitatief verschillen van de data. Dit opent perspectieven voor toepassingen binnen het vakgebied van machinaal leren (*machine learning*) en voor het ontwikkelen van een uitbreiding van de informatietheorie voor kwantumsystemen.

Daar het formalisme een rechtstreekse veralgemening van de informatiemeetkunde inhoudt, kan het gebruik er van ook in die discipline van nut zijn. De belangrijkste innovatie is hierbij dat het nieuwe formalisme toestaat de volledige ruimte van kansverdelingen over een verzameling als model te nemen. Dit kan in de informatiemeetkunde niet op een zinvolle manier gebeuren aangezien de kansmaten daar noodzakelijkerwijs de rol van data vervullen. Deze verzameling tevens als model gebruiken zou het modelleringsproces bijgevolg redundant maken. Bijkomend wordt in dit proefschrift een eenvoudige techniek afgeleid voor het beantwoorden van de vraag of een statistisch model tot de exponentiële familie behoort. De eenvoud van deze werkwijze doet vermoeden dat zij reeds eerder door andere onderzoekers ontdekt is. Desal-

niettemin ontbreken alle verwijzingen hiernaar in de informatiemeetkundige referentiewerken. Dit contrasteert met praktische nut dat zij biedt, zodat een uiteenzetting in dit werk gerechtvaardigd lijkt.
De verhandeling vangt aan met een uiteenzetting van de basideeën van de differentiaalmeetkunde. De daar besproken begrippen vormen het wiskundig kader waaraan de rest van deze tekst wordt opgehangen. Eveneens werd een kort historisch overzicht van de belangrijkste ontwikkelingen in de informatiemeetkunde opgenomen. Deze twee delen vormen samen de inleiding (Hoofdstukken 2 en 3). De stand van het eigenlijke onderzoek wordt beschreven in Hoofdstuk 4. Het data set model formalisme wordt daarin in detail behandeld. Ook de meetkundige structuur wordt gedefinieerd en de studie van haar verband met het modelleringsproces vindt daar plaats. Het vijfde en laatste eigenlijke hoofdstuk richt zich tot een aantal toepassingen, dewelke vooral illustratief van aard zijn.

CONTENTS

1. OPENING CHAPTER

1.1 Situating the dissertation

An often recurring class of problems in science takes the form of describing data by a simplified, parametrised model. Possibly the best-known approach in physics to this kind of modelling is through the method of maximum entropy [1–4]. Its most common use is probably in the context of the thermodynamic canonical ensemble. To apply this method one defines a suitable entropy function and one or more measurable quantities (often called Hamiltonians or extensive variables) on the space of all probability distributions over the possible states of the system. Those distributions which predict the same values for each of the Hamiltonians are grouped together in subsets, each of which are represented by their element exhibiting the highest entropy—the so-called model distributions. To identify the model distribution within such a subset is an optimisation problem with constraints, and the Lagrange multipliers appearing there serve also as labels for these distributions. The task of modelling the experimental data is thus reduced to determining the values of the parameters consistent with the values of the extensive variables that are measured.

While this dissertation does not contain a criticism of the maximum entropy method, a very different and less widely known approach to modelling is studied herein. More specifically, this work is concerned with the development of a framework meant to include a broad range of modelling problems and which is based entirely on geometric foundations: the data set model formalism. The geometry is derived from general divergence functions which quantify how well data is described by a given element of the model. An example of such a divergence function, found in statistics and statistical physics, is the relative entropy or Kullback-Leibler divergence [4, 5].

The motivation to develop such an abstract geometrical theory comes from a criticism made against the mainstream formulation of information theory. This discipline is important to modelling problems as it is intimately related to the question of how to make meaningful statements regarding the quality of the modelling procedure. In particular the fact that information theory is based on probability theory has received critical attention from some authors. This school of thought appears to have started in the work of Ingarden and Urbanik. They proposed an approach to information theory based on Boolean rings and explained their motivation for doing so as [6]

> *. . . information seems intuitively a much simpler and more elementary notion than that of probability . . . [it] represents a more primary step of knowledge than that of cognition of probability.*

Another proponent of this idea is none other than Andrej Kolmogorov, who asserted [7]

> *Information theory must precede probability theory, and not be based on it.*

It must be remarked that the paper in which this statement can be found is devoted to arguing, amongst other things, that the basis of information theory must be combinatorial in nature. While a geometric theory does not have a close relation to combinatorics, the radically different viewpoint it offers, especially when sufficiently abstract, could present new and useful insights into the essence of information even if Kolmogorov should be definitively vindicated.

Such insights of information without probability might also prove to be valuable in physics. An outspoken advocate of the importance of information theory there was John Archibald Wheeler. He stated that [8]

> *All things physical are information theoretic in origin . . . and information gives rise to physics.*

One author to have made an attempt—be it a controversial one—to express Wheeler's idea is Roy Frieden, who tries to base all of physics on an information theoretical foundation [9]. However, it is especially in quantum theory that Wheeler's viewpoint has found many adherents. This is testified by the numerous investigations into the question of whether or not it is possible to frame quantum theory in purely information-theoretic terms, see for example [10–14]. The characterisation of information used by these researchers needs to take into account the laws governing microscopic physics, including the laws of probability. Due to the presence of incompatible observables however, quantum theory requires the notion of conditional probability to be discarded. This implies that quantum information theory must be appreciably different from it classical counterpart, despite being drafted in similar mathematical terms [15, 16]. One potential solution to this problem could lie precisely in the development of a sufficiently abstract formalism, such as the one introduced here, simultaneously generalising both theories of information.

Another argument from quantum physics is related to recent experimental advances in that field, such as those for which Serge Haroche and David Weinland were awarded the 2012 Nobel Prize in Physics. The possibility to perform so-called weak measurements [17, 18] has received much attention, not only in the experimental physics community but also from researchers interested in the foundations of quantum theory [19, 20] and in quantum information theory [21]. In such a measurement, information about the state of the system can be obtained without collapsing the wave function. Recent work even claims successful direct observations of the wave function itself through a combination of weak measurements and ordinary projective measurements [22, 23]—a feat considered impossible by the Copenhagen interpretation of quantum mechanics [24]. An adaptation of quantum information theory in order to accommodate these findings may therefore be required. The availability of an abstract and thereby flexible framework of information would likely be regarded as a boon for those researchers working out the details of such a transition.

Even when no changes to quantum information theory are shown to be needed, the contents of this dissertation may still prove to be useful to that field of study. Newly obtained results—at the time of writing still unpublished—making use of the data set model formalism indicate it may be possible to simultaneously simplify and generalise existing results such as those of Petz on positive-operator valued measures [16].

The data set model formalism is for a large part a generalisation of the results of information geometry. This discipline—which is to be introduced more elaborately in a later chapter—is a differential geometric framework for probability theory and statistical models. The generalisation performed in this research consists mainly in removing the limitation to these topics, leading to a very general picture of the modelling process. While such an approach could have the disadvantage of leaving an impression of abstraction and technicality on the reader, it is also believed to be advantageous in the long term. Indeed, it is precisely by exercising this increased level of abstraction that it is hoped that the geometric essence of information will be laid bare without referring to any context-specific properties.

Some additionally obtained results may be of use to researchers interested in statistics. In particular, the construction of the data set model geometry offers a convenient method to establish whether or not a parametrised family of probability distributions belongs to the exponential family. In this case, the geometric structure also facilitates the search for the canonical parameters of the family. While the simplicity of this method suggests that it is not an

original finding, reference thereto seems to be missing from the information geometry literature.
Furthermore, a varied range of potential applications is expected in the longer term. This supposition is based on a number of advantageous innovations. Perhaps the most prominent of these is that the parametrised family of models is no longer required to be a subset of the data. It is even allowed for models to be mathematical objects qualitatively different from those data. This is to be contrasted with both the maximum entropy method and with information geometry. For this reason, a possibly less obvious but still rather promising field for applications is that of machine learning. This is a very broad area of ongoing research related to artificial intelligence, data mining and other methods for information processing which have dramatically acquired importance over the last few years. A list of example problems includes curve fitting or the estimation of probability density but also applications less—or less obviously—related to parameter estimation such as fingerprint recognition, Google's page rank algorithm, automatised translation and many more. (See for example [25, 26] for an introduction to this very rich discipline.)

1.2 Structure of the dissertation

This opening chapter has summarised the results of the research upon which this dissertation reports and it has sketched the broader context in which these are to be seen. For clarity, a brief explanation of the conventions and notations used in the rest of the text follows shortly hereafter.
The following two chapters form the introduction. Chapter 2 gives a short overview of the basic concepts of differential geometry which are part of this dissertation's lexicon: manifolds, vectors, differential forms, the Riemannian metric, as well as affine connections and their curvature. These notions will provide the mathematical language in which the rest of the dissertation is framed. Chapter 3 introduces information geometry along the lines of its historical development. This chapter also contains a brief introduction to divergence functions as they appear in literature and a short overview of applications of differential geometry in thermodynamics. The main chapter of this dissertation is found fourth and therein the theoretical aspects of the data set model formalism are elucidated. The basic elements and assumptions are explained, a geometrical structure for the models is erected and the essential properties of this structure are studied. Chapters 2–4 share a similar internal structure as to make the analogies between them more clear. Selected examples and applications of the formalism are found in the fifth

chapter for the purpose of illustration. The end of this dissertation is formed by the conclusion and an outlook of possible future research in this context, as well as a curriculum vitae of the author and the bibliography.

1.3 Conventions and notations

There are some conventions and notations which are used throughout this dissertation. A fairly large part of these are expected to be already known or intuitively clear to the reader. Nevertheless, a short summary of these is presented here.

As this dissertation makes elaborate use of differential geometry, one of the most useful conventions is that of Einstein summation. This convention states that when an index appears twice within the same term of an equation, this index is implied to be summed over all of its values. In such pairs, the index will always appear once as an upper index and once as a lower index. Unpaired indices must necessarily appear in every term of an equation. It is conventional for this to mean that the equation holds for all possible values of the unpaired index. For example, it is possible to write down the affine transformation of the vector $\vec{v}$ into $\vec{w}$ by the matrix A and the vector $\vec{b}$ in terms of components as

$$w^i = a^i{}_j v^j + b^i$$

rather than the equivalent traditional, but much more cumbersome, notation

$$w^i = \sum_{j=1}^{n} a^i{}_j v^j + b^i \quad \forall i \in \{1 \ldots n\}.$$

The convention for unpaired indices—the expression in which they appear holds for all values the index may take—is extended to all quantities and the symbols representing them. Whenever an expression makes use of a symbol which is not specified uniquely, the expression is assumed to hold for every possible concrete object the symbol could represent given any restrictions that may apply. This may sound like a complicated and technical way of introducing quantities. Should the reader be left with this feeling, he or she is reminded that this same convention has already been invoked above when introducing the matrix A and the vectors $\vec{v}$ and $\vec{b}$. A similar convention is well-known from expressions for functions but there it is traditionally obscured by speaking about "variables". However, the same result is obtained by agreeing that—for example—the expression

$$f(x) = x \ln(x)$$

holds for all possible real numbers x in the domain of the function f. This convention is useful to avoid what can sometimes become long enumerations of definitions mostly obvious from context, as these have a habit of making mathematical texts more tedious and less pleasant to read than really necessary.

Another situation where a more concise notation is used than the one the reader may be familiar with is partial derivatives with respect to parameters or coordinates. From Chapter 4 onwards, very general divergence functions, denoted by the letter D, will be used. These are functions taking a pair (x, m_θ) of arguments and mapping the pair to a real number $D(x||m_\theta)$. Since the second argument will be an element of a manifold and can be endowed with coordinates, it is possible to differentiate these functions with respect to the coordinates. The traditional notation for such a derivative, evaluated in the point with coordinates θ, would look like

$$\frac{\partial}{\partial \xi^k} D(x||m_\xi)\bigg|_{\xi=\theta}.$$

Since such expressions will appear frequently, it is preferable to use a more concise notation. Therefore, the quantity above will also be represented by the expression resembling the one for functions of the parameters only,

$$\partial_k D(x||m_\theta).$$

For readers who find this notation confusing, it may be useful to keep in mind the analogy with the commonly used expression $f'(a)$, which denotes evaluation of the derivative of a function f in the point a, avoiding the use of an auxiliary variable. The traditional notation will still be used where ambiguity could arise from using the shorter alternative.

When specific notations are introduced on the fly, this will be denoted by using the symbol $\overset{\text{not.}}{=}$ as the equality sign—not to be confused with the inequality $\neq$. A similar symbol $\overset{\text{def.}}{=}$ is used to indicate that the equality is in fact also the definition of the quantity being introduced.

A final convention has to do with measures over sets. Sometimes it is necessary to sum or integrate over such a set or a subset thereof. This will appear in a number of expressions originating in statistics, for example expressions for expectation values. For this the integration sign shall be used, even when the measurable set over which the integration takes place may be discrete. Unless explicitly stated otherwise, $\mathrm{d}x$ represents the measure in the integral—whereas in the literature it would be common to use a notation like $\mathrm{d}\mu(x)$ instead.

2. ELEMENTARY DIFFERENTIAL GEOMETRY

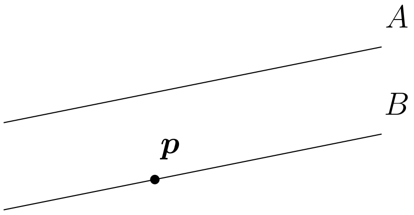
This chapter is the first of two introductory chapters. As such, it will provide a quick introduction into differential geometry. This branch of mathematics will provide the mathematical framework for this dissertation. Before commencing the introduction of this subject in earnest, it should be noted that differential geometry is an exceptionally interesting and rich field. As such no brief introduction could possibly do it justice. Therefore this discussion is limited to those concepts strictly necessary for the understanding of this dissertation. This list of topics, which will also serve as a backbone for the following two chapters, includes manifolds, coordinate functions, the metric tensor and the affine connection.

Readers interested in discovering more of this most elegant field are referred to introductory works such as [27] and [28], upon which this introduction has loosely been based and which serve as the main reference material. The book by Pressley is dedicated to the "extrinsic" variant of differential geometry, which means it studies curves and surfaces embedded in a larger space. Frankel's book on the other hand, devotes most of its attention to the "intrinsic" theory, which can be formulated independently of a containing space and which is also closer to the differential geometry applied in this dissertation. Readers with an interest in information geometry, which will be discussed in the next chapter, may also find the first chapters of Amari's books [29] and [30] interesting introductions, although they are far less broad in scope than more dedicated works.

2.1 The origins of differential geometry

Differential geometry is a field of study opened up by Carl Friedrich Gauss originally concerned with curves and surfaces embedded in Euclidean space. Its scope was greatly expanded after the contributions of the Hungarian mathematician János Bolyai and the Russian mathematician Nikolai Ivanovich Lobachevsky, who discovered the solution to a long standing problem [31]. Over the course of two millennia, scores of mathematicians were plagued by the nature of Euclid's so-called Parallel Postulate [32], which in the Playfair formulation reads[1]

> *In a plane, given a line A and a point $\boldsymbol{p}$ not on it, at most one line B parallel to the given line A can be drawn through $\boldsymbol{p}$.*

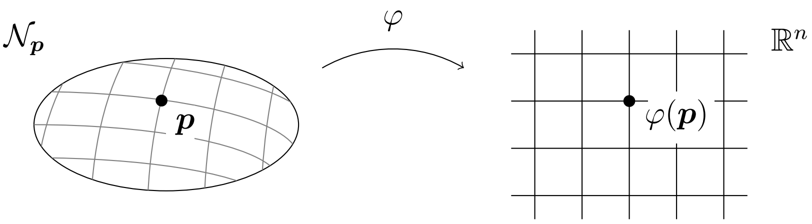


It was generally believed that this axiom is a consequence of the other axioms of Euclidean geometry, even though no one was able to demonstrate this. Bolyai and Lobachevsky independently provided the verdict by showing the existence of mathematically consistent geometries which did not satisfy Euclid's Parallel Postulate. To achieve this, they altered the axiom as to demand that more than one straight line through $\boldsymbol{p}$ can be constructed which is parallel to A. The resulting geometries are known as hyperbolic geometries. Two decades later, the German mathematician Georg Bernhard Riemann developed what is now known as Riemannian geometry [33]. This is probably the best-known class of non-Euclidean geometries in the physics community as the mathematical framework of Einstein's general theory of relativity is based upon Riemann's ideas. Through the collaboration of mathematicians such as Élie Cartan, Henri Poincaré and others, the study of Riemannian geometry eventually led to the mathematical branch now known as differential geometry, a very general framework for describing the geometry of spaces.

2.2 Manifolds and (co)tangent spaces

The stages upon which differential geometry takes place, that is the spaces of which the geometry is studied, are manifolds. Manifolds will be denoted by the double letter $\mathbb{M}$. These are sets in which every element $\boldsymbol{p}$, usually called a point, has a neighbourhood $\mathcal{N}_{\boldsymbol{p}}$ which allows for the definition of coordinates for the points in that neighbourhood. A coordinate function φ is a homeomorphism $\varphi : \mathcal{N}_{\boldsymbol{p}} \to \mathbb{R}^n$, that is it associates with every point an ordered set of n real numbers.

[1] It should be noted that this formulation is only correct in the presence of the four preceding axioms of Euclidean geometry. In the current context, however, this formulation suffices and it has the advantage of simplicity.



These numbers $\{\varphi^i(\boldsymbol{p})\}$ are called the coordinates of the point $\boldsymbol{p}$. A homeomorphism is a bijective map which is continuous and which has a continuous inverse [34]. It is important for φ to be a homeomorphism and not just any map. The invertibility means that φ endows only a single point of its domain with a specific set of coordinates. The continuity of the coordinate map means that points close together in $\mathbb{M}$ are associated with coordinates which are also close together in $\mathbb{R}^n$. This ensures that the topological structure of $\mathbb{M}$ is faithfully reflected in the coordinates. The natural number n is called the dimension of the manifold. For practical reasons the dimension is assumed to be finite, although many interesting properties and applications do exist in situations where the dimension is infinite, see for instance [35,36]. Many mathematical sets encountered in physics are manifolds, even though they are not always identified as such. The space in which classical physics takes place is a manifold and the coordinate maps provide the familiar coordinates of points. The same thing is true for space-time as it appears in the theory of relativity. Most of the terminology in the theory of manifolds is actually derived from the analogy with physical space. Other examples of manifolds include, although care should be taken in the sense that not all of these can be covered by a single coordinate function, the configuration space of classical mechanics and the state space of statistical physics, as well as all finite-dimensional vector spaces and—by extension—Hilbert spaces, which are widely known as an essential element of modern quantum theory.

Two very important structures which can be defined on manifolds and used in describing the geometry of a space are the metric tensor and affine connections. In order to elucidate these concepts, it is necessary to first introduce a few more elementary concepts such as vectors and differential forms.

After the manifold, the most basic object of differential geometry is the vector. Perhaps the physically most intuitive interpretation is derived from velocity vectors. Consider an observer who measures the value of a function f everywhere on her path and who also keeps time in order to tabulate the measured value as a function of time. This observer can then compute the rate of change in the value of f she observed with respect to the time passed. It is possible to define the velocity vector of this observer as a functional which returns precisely this rate of change when applied to the function f.

More rigorously, when a coordinate function φ is fixed, then a vector $\vec{v}$ at a point $\boldsymbol{p}$ is a linear operator on real-valued functions $f : \mathbb{M} \to \mathbb{R}$ satisfying

$$f \mapsto \vec{v}(f) = v^i \frac{\partial}{\partial \varphi^i}(f \circ \varphi^{-1})\Big|_{\varphi(\boldsymbol{p})} \tag{2.1}$$

for all functions f for which these derivatives exist. The numbers v^i are called the components of $\vec{v}$. In the case of velocity vectors these components equal the derivatives of the coordinate functions with respect to time.
Even though the definition (2.1) depends on the choice of a coordinate function φ, it can be shown that the vector itself is indeed invariant under suitable coordinate transformations. In particular, if a second set of coordinates $\{\zeta^a\}$ is employed, these must be a diffeomorphic function, that is a smooth function with a smooth inverse, of the original coordinates $\{\varphi^i\}$. In such case, the chain rule of calculus implies the derivatives relate to each other as

$$\frac{\partial}{\partial \zeta^a} = \frac{\partial \varphi^i}{\partial \zeta^a}\frac{\partial}{\partial \varphi^i}, \tag{2.2}$$

where the first factor on the right hand side represents the components of the Jacobian matrix of the transformation. If the coordinate independence of the vector $\vec{v} = v^i \partial_i = v^a \partial_a$ is to be respected, this means the components must transform according to the inverse transformation, that is

$$v^a = \frac{\partial \zeta^a}{\partial \varphi^i} v^i.$$

That this is satisfied is also a consequence of the chain rule, as it is equivalent to the expression

$$\frac{\mathrm{d}\zeta^a}{\mathrm{d}t} = \frac{\partial \zeta^a}{\partial \varphi^i}\frac{\mathrm{d}\varphi^i}{\mathrm{d}t}.$$

As was hinted at before, the most common example of a vector is the velocity vector, both in classical mechanics and in the theory of relativity. Acceleration is also described as a vector. Many other quantities are described as vectors in physics, such as linear and angular momentum, force, torque… Most of these, however, including the four listed, appear more naturally in modern theories as differential forms [28]. Those objects and their operations can be framed as vectors but to do this requires a few pieces of extra mathematical machinery, not all of which are contained in the scope of this introduction. It is important to remark that the particular definition of vectors used here is more restricted than the algebraic definition as simply an element of a vector space. This particular kind of vectors are properly called tangent vectors. Another type of vectors in the algebraic sense to be discussed in this dissertation are the differential forms with which this section will be concluded. Other, more general, vector structures will not play a major role in this dissertation and will therefore not be given further consideration in this introduction.

It is possible to verify that all tangent vectors with the same point of application $\boldsymbol{p}$ form a vector space when endowed with componentwise addition and scalar multiplication. This follows from the fact that linear combinations of such vectors are still of the form (2.1). This vector space is called the tangent space at the point $\boldsymbol{p}$, or $\mathrm{T}_{\boldsymbol{p}}\mathbb{M}$ for short. Together, all vectors over a manifold constitute the tangent bundle[2], abbreviated $\mathrm{T}\mathbb{M}$. The tangent bundle can also be seen as a manifold—the configuration space of mechanics is probably the best known tangent bundle—but it is not a vector space as addition is only defined for pairs of vectors which share a point of application. The operations obtained by applying partial derivative operators $\{\partial_i\}$ and evaluating the result at a point $\boldsymbol{p}$ form the basis vectors of the tangent space $\mathrm{T}_{\boldsymbol{p}}\mathbb{M}$.

A particularly interesting type of subsets of the tangent bundle are vector fields. A vector field can be likened to a function $F : \mathbb{M} \to \mathrm{T}\mathbb{M}$ but with the additional restriction that the function value of a point $\boldsymbol{p}$ must be an element of $\mathrm{T}_{\boldsymbol{p}}\mathbb{M}$ and not of another tangent space. The traditional but more technical formulation states that a map $F : \mathbb{M} \to \mathrm{T}\mathbb{M}$ is a vector field when $\pi \circ F$, where π is the bundle projection map associating with every vector its point of application, is the identity mapping on the domain of F. The basis vectors employed thus far, the differential operators $\{\partial_i\}$, naturally form $n = \dim(\mathbb{M})$ linearly independent vector fields. Such a collection of vector fields which form a basis for each tangent space (in their domain) is called a frame. It is not required that a frame consists out of differential operators with respect to coordinate functions but this dissertation will exclusively employ such so-called coordinate frames as they are probably the most familiar to readers.

As the tangent spaces $\mathrm{T}_{\boldsymbol{p}}\mathbb{M}$ are vector spaces, it is meaningful to speak of the dual vector space: the space of all linear functionals on $\mathrm{T}_{\boldsymbol{p}}\mathbb{M}$ with values in the real numbers. Where tangent vectors are probably most familiar to readers as column vectors, their duals are traditionally denoted—and better known—as row vectors. The dual space of $\mathrm{T}_{\boldsymbol{p}}\mathbb{M}$ is called the cotangent space at $\boldsymbol{p}$ and is denoted in a concise way as $\mathrm{T}^*_{\boldsymbol{p}}\mathbb{M}$. Together these cotangent spaces make up the cotangent bundle $\mathrm{T}^*\mathbb{M}$. Smooth fields over this bundle are called (differential) 1-forms. Phase space as it appears in Hamiltonian

[2] Some sources such as [37] reserve this name for the bundle projection map.

mechanics is probably the most familiar example of a cotangent bundle, even though it is not usually introduced as such in physics textbooks.

Just as a canonical basis $\{\partial_i\}$ of partial derivative operators exists for $\mathrm{T}_{\boldsymbol{p}}\mathbb{M}$ given a coordinate function φ, a canonical basis $\{\sigma^i\}$ exists for $\mathrm{T}^*_{\boldsymbol{p}}\mathbb{M}$ given this same coordinate function. The functionals σ^j making up this dual basis satisfy

$$\sigma^j(\partial_i) \overset{\text{def.}}{=} \delta^j_i. \tag{2.3}$$

A most elegant way of introducing the canonical dual basis is through the exterior derivative operator, for which the symbol "d" is used. This operator maps its arguments, which are differentiable functions, into differential 1-forms. More in particular the exterior derivative $\mathrm{d}f$ of a scalar function f acts on a vector $\vec{v}$ as

$$(\mathrm{d}f)(\vec{v}) \overset{\text{def.}}{=} \vec{v}(f).$$

By choosing $f = \varphi^i$, the i^{th} component of the coordinate function, it follows that

$$\begin{aligned}(\mathrm{d}\varphi^i)(\partial_j)\Big|_{\boldsymbol{p}} &= \partial_j(\varphi^i \circ \varphi^{-1})\Big|_{\varphi(\boldsymbol{p})} \\ &= \delta^i_j.\end{aligned}$$

This, combined with the linearity of these operators, implies $\sigma^i = \mathrm{d}\varphi^i$ for the duals of the basis $\{\partial_j\}$. It is often useful to choose the 1-forms $\{\mathrm{d}\varphi^i\}$ as the basis for the cotangent spaces, even though it is not strictly necessary to do so. One advantage of this choice is that the exterior derivative of a function f can easily be expressed as

$$\mathrm{d}f = (\partial_i f)\ \mathrm{d}\varphi^i. \tag{2.4}$$

This expression also shows there is a connection between the exterior derivative of a function and its gradient. The numbers $\partial_i f$ are equal to the components of the gradient of f in Cartesian coordinates but not in an arbitrary curvilinear coordinate system. Expression (2.4) however, holds in all coordinate systems, which offers obvious practical advantages.

The 1-forms introduced allow for the construction of a very rich algebraic structure. They can be endowed with a product, denoted as $\wedge$, which is fully antisymmetric. More specifically, given two 1-forms α and β, the 2-form $\alpha\wedge\beta$ is a bilinear mapping defined by

$$(\alpha \wedge \beta)(\vec{v}, \vec{w}) = \alpha(\vec{v})\beta(\vec{w}) - \alpha(\vec{w})\beta(\vec{v}).$$

This operation has many uses in differential geometry, including the definition of the exterior derivative of 1-forms. Given a 1-form $\alpha = \alpha_i \mathrm{d}\varphi^i$, where $\alpha_i = \alpha(\partial_i)$, its exterior derivative equals

$$\mathrm{d}\alpha \overset{\text{def.}}{=} (\mathrm{d}\alpha_i) \wedge \mathrm{d}\varphi^i = (\partial_j \alpha_i)\ \mathrm{d}\varphi^j \wedge \mathrm{d}\varphi^i.$$

From the skew symmetry of $\mathrm{d}\varphi^i \wedge \mathrm{d}\varphi^j$, it automatically follows that $\mathrm{d}(\mathrm{d}f) = 0$ for all real-valued functions f. A very useful property is the (partial) converse to this, which is called Poincaré's lemma. It states that, on a manifold which can be contracted into a single point by a continuous transformation, the vanishing of a form's exterior derivative, i. e. $\mathrm{d}\alpha = 0$, is a sufficient condition for the existence of a function f such that $\alpha = \mathrm{d}f$. Though not all manifolds in this thesis are indeed contractible, it is nevertheless often possible to apply Poincaré's lemma locally even when it does not hold globally. Such a function f is called a potential for the 1-form α.

This structure of differential forms can be extended to define higher forms and their exterior derivatives as well. This is done in a completely analogous fashion. The use of forms in this dissertation will remain fairly limited to accommodate readers more familiar with the index-heavy tensorial notation traditionally used in the physics literature. As a consequence forms or arbitrary degree will not be treated in this introduction. Nevertheless, some results involving higher forms than those of first degree will be required in some situations. One example of such a result is the analogue of the Leibniz rule for the wedge product of two 1-forms α and β:

$$\mathrm{d}(\alpha \wedge \beta) = (\mathrm{d}\alpha) \wedge \beta - \alpha \wedge (\mathrm{d}\beta).$$

Also Poincaré's lemma will be used for higher forms. Its generalisation is straightforward: on a contractible manifold any differential p-form of which the exterior derivative vanishes can be written as the exterior derivative of some differential $(p-1)$-form.

2.3 The Riemannian metric

So far this introduction to differential geometry has not yet included any mention of distance or angles, arguably both important concepts in any study of geometry. In order to introduce these quantities, one must choose an inner product for each of the tangent spaces. Such a mapping is called a metric tensor and in particular, since it is defined on the tangent spaces, a Riemannian metric. This mapping is denoted by g and so it can be written that[3]

$$g(\cdot,\cdot)\big|_{\boldsymbol{p}} : \mathrm{T}_{\boldsymbol{p}}\mathbb{M} \times \mathrm{T}_{\boldsymbol{p}}\mathbb{M} \to \mathbb{R}.$$

An inner product must be symmetric, which means that $g(\vec{v},\vec{w}) = g(\vec{w},\vec{v})$ for all vectors $\vec{v}$ and $\vec{w}$ in the same tangent space. Another requirement is the positive definiteness, which means $g(\vec{v},\vec{v}) \geqslant 0$ for all vectors $\vec{v}$ and that the value can only be equal to zero when all the vector components equal zero. Often the value of the inner product will be written in its coordinate representation $g(\vec{v},\vec{w}) = g_{ij}v^i w^j$, where the components are given by

$$g_{ij} = g(\partial_i, \partial_j). \tag{2.5}$$

The fact that this notation is possible follows from the property of bilinearity, which means linearity holds in both arguments separately.
Lengths of vectors in $\mathrm{T}_{\boldsymbol{p}}\mathbb{M}$ and angles between two vectors in this space can be computed using the metric through the respective familiar relations

$$v = \sqrt{g(\vec{v},\vec{v})|_{\boldsymbol{p}}} \quad \text{and} \quad \cos(\phi) = \frac{g(\vec{v},\vec{w})|_{\boldsymbol{p}}}{vw}.$$

Readers will no doubt be familiar with the dot product on $\mathbb{R}^3$ but this does not necessarily paint a representative image of Riemannian metrics. This is because a large part part of the physics literature describes space also as $\mathbb{R}^3$, which itself can also be seen as a vector space. Elementary textbooks traditionally make use of this extra structure when introducing vector calculus. In particular, identifying all tangent spaces with each other conceals that the dot product is actually defined on each tangent space separately. The best known examples of non-trivial Riemannian metrics are probably the solutions to the field equations which lie at the heart of Einstein's general theory of relativity. The Fisher information matrix [38] appearing in statistics, for example in the Cramér-Rao inequality [39], is used as a metric tensor in the field of information geometry. This will be discussed in more detail in the next chapter.

2.4 Connections and curvature

A similarly important structure on a manifold is the affine connection. The connection may seem a more technical notion than the metric but it has many

[3] Where context avoids confusion, the index $\boldsymbol{p}$ from this notation will not be mentioned explicitly. Alternatively, this definition can be expanded to map pairs of vector fields into real-valued functions over $\mathbb{M}$.

practical applications nonetheless. The most common way to introduce a connection on a tangent bundle is through a covariant derivative operation, denoted by ∇. Although strictly speaking an abuse of notation, this symbol is often used to refer to the connection as well. Just as derivatives of functions play a role in many branches of mathematics and science, a similar notion is often very useful for vector fields. Consider the problem of defining the directional derivative of a vector field $\vec{X}$ in the direction of the vector $\vec{v} \in \mathrm{T}_{\boldsymbol{p}}\mathbb{M}$. It is possible to construct such a derivative and the resulting vector is denoted by $\nabla_{\vec{v}}\vec{X}$. However, some difficulty is involved in working out this notion so it is instructive to first take a look at the easy aspects and leave the hard part for last.

A first property that the covariant derivative must satisfy is linearity in its first argument—just like the more familiar directional derivative of scalar functions. This means

$$\nabla_{\vec{v}}\vec{X} = v^i \nabla_{\partial_i}\vec{X} \overset{\text{not.}}{=} v^i \nabla_i \vec{X}.$$

Furthermore, the covariant derivative should be linear for addition in the second argument and satisfy the Leibniz rule

$$\nabla_{\vec{v}}(f\vec{X}) = \vec{v}(f)\vec{X} + f\nabla_{\vec{v}}\vec{X}, \tag{2.6}$$

where f is an arbitrary differentiable function. This property—combined with the expansion of the vector field $\vec{X} = X^i\partial_i$—states that

$$\nabla_{\vec{v}}\vec{X} = \vec{v}(X^i)\partial_i + X^i\nabla_{\vec{v}}\partial_i. \tag{2.7}$$

This shows it suffices to determine the covariant derivatives of the basis vectors—here of course seen as fields. Because of the first linearity property, it is even enough to know at every point $\boldsymbol{p}$ the coefficients of the expansion

$$\nabla_i\partial_j = \omega^k{}_{ij}\partial_k. \tag{2.8}$$

These numbers $\omega^k{}_{ij}$ are called the coefficients of the connection at $\boldsymbol{p}$. There is no obvious way to determine these numbers in all contexts, something which is also the case with the metric tensor. In fact, the whole point of Einstein's general theory of relativity is that the geometric properties of spacetime follow from physical laws rather than from mathematical principles! The connections playing an important role in information geometry are also defined from context-specific arguments. This is elucidated in the next chapter.

It is interesting to note that the second term in the right hand side of equation (2.7) is essentially a linear transformation of the vector $\vec{X}$ (at $\boldsymbol{p}$). This is even more apparent when leaving the first argument of the covariant derivative unmentioned:

$$\nabla\vec{X} = \partial_i \otimes [\mathrm{d}X^i + {\omega^i}_j X^j],$$

where the expression between square brackets is an element of $\mathrm{T}^*\mathbb{M}$ and the connection 1-forms

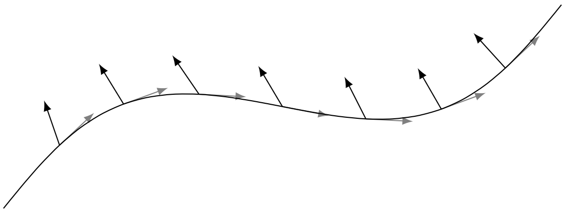
$$ {\omega^i}_j \overset{\text{def.}}{=} {\omega^i}_{kj}\mathrm{d}\varphi^k \tag{2.9}$$

were implicitly introduced. This notation of connection forms turns out to be very convenient in some formulas. Against the general trend in this dissertation of limiting the use of forms, this notation will appear in some places. For example the proof of Riemann's theorem following shortly would be much more cumbersome when making use of the more usual notations introduced in equations (2.7) and (2.8).

Apart from yielding a definition for the derivative of a vector field, a connection can also be used to define whether or not two vectors at different points are parallel. In order to do this, the concept of parallel transport must first be defined. Given a curve $C : [t_i, t_f] \to \mathbb{M}$, it is said a vector field $\vec{Y}$, defined (at least) on the curve C with tangent vector field $\vec{X} = \frac{\mathrm{d}}{\mathrm{d}t}$, is parallel along C or has undergone parallel transport along C if

$$\begin{aligned} 0 &= \nabla_{\vec{X}}\vec{Y} \\ &= \partial_j \otimes X^i[\partial_i Y^j + {\omega^j}_{ik}Y^k] \\ &= \partial_j \otimes \left[\frac{\mathrm{d}Y^j}{\mathrm{d}t} + {\omega^j}_{ik}\frac{\mathrm{d}\varphi^i}{\mathrm{d}t}Y^k\right]. \end{aligned}$$

This gives a way to transport vectors from one tangent space to another: solving these n ordinary first order differential equations yields the components of the vector field everywhere along the curve (parametrised by t). In the drawing this is demonstrated for a toy connection. The tangent vector field $\vec{X}$ to the curve (in grey) is also shown.



In general it should be noted that transporting a vector over different curves with identical end points need not yield the same result. Equivalently, parallel transport of a vector around a closed loop will in general not make the final vector coincide with the first one. The difference between the original vector and a vector transported around a closed loop can be quantified using the curvature of the connection. If the curvature vanishes, the connection is said to be flat and this is equivalent to saying that parallel transport is independent of the chosen curve.

Most introductory texts introduce curvature through the commutativity of covariant derivatives, which is equivalent with the vanishing of the curvature. While this is correct, such derivation usually does not include the reasoning leading up to the most interesting result, which is the existence of covariant constant vector fields. A vector field $\vec{X}$ is said to be covariant constant if parallel transport of the value of the vector field at a point $\boldsymbol{p}$ to a point $\boldsymbol{q}$ is equal to the value of the field at $\boldsymbol{q}$. This definition requires the parallel transport to be independent of the path and so the vanishing of the connection is a necessary condition for covariant constant vector fields. The next two pages will deal with showing that this is also a sufficient condition by proving the slightly stronger result known as Riemann's theorem.

The best-known property of manifolds endowed with a flat connection is that they allow for global Cartesian coordinates. However, the connection may arbitrary. Rather it must be the metric connection of Levi-Civita, which has coefficients given by

$$\omega^k{}_{ij} = \Gamma^k{}_{ij} \overset{\text{def.}}{=} \frac{1}{2} g^{ks} \left(\partial_i g_{sj} + \partial_j g_{is} - \partial_s g_{ij} \right), \tag{2.10}$$

where the numbers g^{ij} are the components of the matrix inverse of the metric tensor. It will be shown that when this particular connection exhibits vanishing curvature, there exists on the manifold a local set of coordinates in which the metric tensor everywhere takes the Euclidean form

$$g(\vec{v}, \vec{w}) = \sum_{j=1}^{n} v^j w^j. \tag{2.11}$$

The property that the metric tensor can be written in this form when the metric connection (2.10) is flat, is Riemann's theorem. Note that this is not a trivial statement. The metric can take this form at any point by choosing an appropriate set of coordinates called normal coordinates. However, nothing guarantees that this form will be valid also outside this one point while still using the same set of coordinates. It is only when the curvature vanishes that coordinates exist in which this property holds everywhere.

First, it will be shown that for any connection the parallel displacement of vector fields is independent of the path followed if and only if its curvature vanishes. Only then will it be shown how this gives rise to the Euclidean

form of the metric tensor.
A vector field $\vec{X}$ is said to be parallel along a curve C parametrised by t if its covariant derivative vanishes along that curve or, equivalently,

$$\begin{aligned} 0 &= \partial_i \otimes \left[\frac{\mathrm{d}X^i}{\mathrm{d}t} + \omega^i{}_{jk} \frac{\mathrm{d}\varphi^j}{\mathrm{d}t} X^k \right] \\ &= \partial_i \otimes \left[\frac{\partial X^i}{\partial \varphi^j} \frac{\mathrm{d}\varphi^j}{\mathrm{d}t} + \omega^i{}_{jk} \frac{\mathrm{d}\varphi^j}{\mathrm{d}t} X^k \right] \\ &= \partial_i \otimes \left[\mathrm{d}X^i + \omega^i{}_k X^k \right] \left(\frac{\mathrm{d}\varphi^j}{\mathrm{d}t} \partial_j \right) . \end{aligned} \tag{2.12}$$

From the condition (2.12) for the vanishing of the covariant derivative it can be seen that parallel transport yields the same result independent of the curve C if and only if

$$\mathrm{d}X^i + \omega^i{}_k X^k = 0 \tag{2.13}$$

since this left hand side is the curve-independent part of the condition. Frobenius' theorem states that a sufficient condition for the equations (2.13) to have a solution is then for the 2-forms

$$\mathrm{d}(\mathrm{d}X^i + \omega^i{}_k X^k)$$

to vanish [28]. It is possible to rewrite, using (2.13) again in the second equality

$$\begin{aligned} \mathrm{d}(\mathrm{d}X^i + \omega^i{}_k X^k) &= \mathrm{d}^2 X^i + (\mathrm{d}\omega^i{}_k) X^k + \mathrm{d}X^k \wedge \omega^i{}_k \\ &= (\mathrm{d}\omega^i{}_j) X^j + (-\omega^k{}_j X^j) \wedge \omega^i{}_k \\ &= (\mathrm{d}\omega^i{}_j + \omega^i{}_k \wedge \omega^k{}_j) X^j . \end{aligned}$$

This means parallel transport of any vector field $\vec{X}$ is path-independent when the curvature 2-forms

$$\Omega^l{}_k \overset{\text{def.}}{=} \mathrm{d}\omega^l{}_k + \omega^l{}_s \wedge \omega^s{}_k \tag{2.14}$$

vanish identically, which concludes the proof of the first statement made above. In a more index-heavy notation, this can be expressed as

$$0 = \Omega^l{}_{kij} = \partial_i \omega^l{}_{jk} - \partial_j \omega^l{}_{ik} + \omega^l{}_{is} \omega^s{}_{jk} - \omega^l{}_{js} \omega^s{}_{ik} .$$

When the curvature vanishes, parallel transport is path independent and so covariant constant vector fields can be constructed by parallel transporting

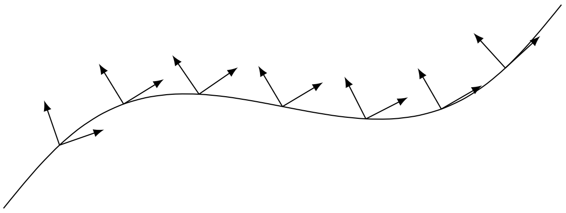
a vector at a single point to other points of $\mathbb{M}$.
The proof of Riemann's theorem is continued by showing that these covariant constant vector fields are partial derivatives with respect to some coordinate functions θ^i. This is useful since in these coordinates, the connection coefficients

$$\omega^k{}_{ij} = \mathrm{d}\theta^k(\nabla_i \partial_j)$$

vanish identically. Such coordinates are known as the affine coordinates of the connection. Although the proof is not in itself very difficult, it requires some new concepts and properties, the introduction of which would burden the reader with more text than can be justified from their use in this dissertation. Interested readers are thus referred to the source material mentioned earlier for the details of this part of the proof.
It should be stressed that the results up to now hold for all flat affine connections. Riemann's theorem deals with the specific case of the Levi-Civita connection (2.10). The interesting property of this connection is that it satisfies

$$\vec{X}(g(\vec{Y},\vec{Z})) = g(\nabla_{\vec{X}}\vec{Y},\vec{Z}) + g(\vec{Y},\nabla_{\vec{X}}\vec{Z}) \tag{2.15}$$

and so it preserves inner products between vector fields under parallel transport. But this means that when one chooses an orthonormal basis at one point of $\mathbb{M}$, parallel transport through the Levi-Civita connection will result in vector fields which are orthonormal everywhere. The drawing illustrates this for transport along a curve in the plane.



By the above, the vector fields obtained in this way are partial derivatives with respect to some coordinates $\{\theta^i\}$. In those coordinates, the metric will thus take its Euclidean form (2.11). This concludes the proof of Riemann's theorem.
Perhaps the best known application of connections consists in the construction of geodesics. Most people are familiar with geodesics as the shortest paths connecting two given points but this is not generally true as this is a particular property exclusive to the Levi-Civita connection. In general a geodesic is a curve to which the tangent vectors form a covariant constant

vector field. Since such vector fields are said to be parallel to themselves, the geodesics are the natural generalisation of straight lines.

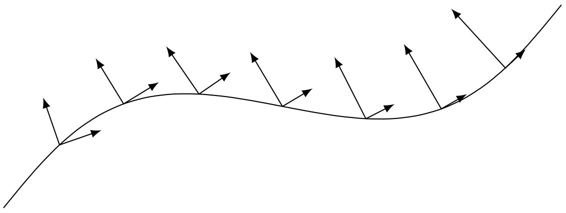
The curvature as measured by the 2-forms ${\Omega^l}_k$ is the intrinsic curvature and, if applicable, is independent of the way the manifold is embedded in a larger manifold. As such it must be contrasted with extrinsic or embedding curvature, which depends on the way the manifold is embedded in a larger space and in particular on how the normal vector fields on the manifold behave as one moves over the manifold. An entire theory of embedding curvature can be set up and a short but clear summary can be found in Chapter 2 of Amari's book [29]. Also Pressley's book [27] concerns itself with embedding curvature but it is limited to three dimensions and it uses the traditional notation, making it perhaps more difficult to see how to set up such a theory in an arbitrary number of dimensions. The difference between intrinsic and extrinsic curvature will become an important point later in this introduction, in particular in Chapter 3.

To draw an intuitive picture of the difference between the two kinds of curvature, some examples might be useful to keep in mind. A curve, which is a one-dimensional manifold, cannot have intrinsic curvature. This is because all 2-forms, including the curvature 2-forms ${\Omega^l}_k$, vanish on one-dimensional manifolds by the antisymmetry of the wedge product. A curve can have extrinsic curvature, however, and this is the case when a tangent vector field must change direction as one moves over the curve. This extrinsic curvature is measured by the inverse of the radius of curvature as it is known from more elementary texts. A sphere, on the other hand, is intrinsically curved. This means that when the sphere is embedded in a flat space like $\mathbb{R}^n$, the normal vectors to the sphere must necessarily change direction as one moves over the sphere and so the embedding curvature will not vanish globally, even though it may do so locally should the sphere be deformed appropriately.

In this dissertation several affine connections will appear and these will not all be metric connections. However, it is still possible to obtain interesting properties of affine connections beyond Riemann's famous result. In particular, it is possible to define the notion of dual connections through a relation generalising (2.15). When two connections ∇ and ∇^* satisfy

$$\vec{X}g(\vec{Y},\vec{Z}) = g(\nabla_{\vec{X}}\vec{Y},\vec{Z}) + g(\vec{Y},\nabla^*_{\vec{X}}\vec{Z}), \tag{2.16}$$

again for all vector fields $\vec{X}$, $\vec{Y}$ and $\vec{Z}$, it is said these connections are dual with respect to the metric tensor g. Every connection ∇ has a dual connection as expression (2.16) can be used as a definition of ∇^*. Parallel transport of one vector field through a connection ∇ and of the other vector field through the dual connection ∇^* will always preserve the inner product between the vector fields. The drawing illustrates this for a pair of orthogonal vector fields in the plane transported over a curve, each through a different connection. Note that neither of these connections preserves the length of the vectors under parallel transport. Since the original vectors are orthogonal, their transported counterparts are as well. In general the angle between the vectors need not be preserved either.



Due to the symmetry of the metric, it holds that $(\nabla^*)^* = \nabla$ and it is possible to show that a pair of dual connections have curvatures which cannot vanish independently of each other [29]. Such pairs of dual connections play an important role in information geometry but they are not so common in most applications of differential geometry. Therefore most introductory texts on differential geometry do not cover this topic.
Before finally concluding this introduction to differential geometry, it is important to devote attention to a last application of covariant derivatives. The previous seven pages considered the covariant derivative of a vector. At least as important is the notion of the covariant derivative of a 1-form and the resulting definition of the Hessian.Remember that the Leibniz rule for a covariant derivative, equation (2.6), implicitly made use of the property that the covariant derivative of a function is the same as the regular derivative, that is

$$\nabla_{\vec{v}} f = \vec{v}(f).$$

One way to construct a differentiable scalar function is to apply a 1-form to a vector field—both of which must have differentiable components. This leads to the scalar function $f = \alpha(\vec{X}) = \alpha_i X^i$. This form of the expression hints that a differentiation of f can be performed using a variation on the Leibniz rule,

$$\nabla_{\vec{v}}(f) = \nabla_{\vec{v}}(\alpha(\vec{X})) = (\nabla_{\vec{v}}\alpha)_i X^i + \alpha_i(\nabla_{\vec{v}}\vec{X})^i.$$

The presence of a lower index on $\nabla_{\vec{v}}\alpha$ shows that this quantity itself is also a 1-form, just as the covariant derivative of a vector field is again a vector field. Rearranging the above line and working out the resulting expression yields

$$\begin{aligned}(\nabla_{\vec{v}}\alpha)(\vec{X}) &= \nabla_{\vec{v}}(\alpha(\vec{X})) - \alpha(\nabla_{\vec{v}}\vec{X}) \\ &= v^i\partial_i(\alpha_j X^j) - \alpha_k(v^i\partial_i X^j + \omega^k{}_{ij}v^i X^j) \\ &= v^i[\partial_i\alpha_j - \omega^k{}_{ij}\alpha_k]X^j.\end{aligned}$$

Since this must hold for any vector field $\vec{X}$, the covariant derivative of a 1-form is given by

$$\nabla_{\vec{v}}\alpha = v^i[\partial_i\alpha_j - \omega^k{}_{ij}\alpha_k]\mathrm{d}\varphi^j.$$

The covariant derivative has the same function for 1-forms as for vectors: it performs a directional derivative in such a way that the result is independent of the chosen basis for the (co)tangent spaces. This is particularly useful when speaking about the Hessian in the context of manifolds. The Hessian is the generalisation of the second derivative of a scalar function. However, computing higher derivatives must take into account that not only the function is liable to change from point to point. Also the direction in which a partial derivative quantifies the change of a function need not remain the same. This change of direction is exactly what the connection coefficients express. Therefore, the Hessian of a function f is defined as

$$\mathrm{Hess}(f)(\vec{u},\vec{v}) \overset{\text{def.}}{=} (\nabla_{\vec{u}}\mathrm{d}f)(\vec{v}).$$

Note that though $\mathrm{Hess}(f)$ is a function taking two vectors as arguments and mapping them into a real number, it is not a 2-form. If f is continuously differentiable then $\mathrm{Hess}(f)$ is a symmetric mapping—whereas a 2-form would be antisymmetric. To stress the symmetry of the Hessian, a more commonly used notation for this object is

$$\begin{aligned}\mathrm{Hess}(f)(\partial_i,\partial_j) &= \nabla_i\nabla_j f \\ &= \partial_i\partial_j f - \omega^k{}_{ij}\partial_k f. \qquad (2.17)\end{aligned}$$

It is this notation that will be used throughout the other chapters.

As is stated in the first paragraphs of this chapter, a complete introduction to differential geometry would include a large amount of material not present in this introduction. In particular many theorems and properties are mentioned but not treated explicitly. Some of those theorems, such as the result on integrability that bears Frobenius' name and which is used as a step in the above proof of Riemann's theorem, are also important in other parts of this dissertation. Where such theorems are invoked they shall be mentioned when possible. For their proofs and the required background, the reader is again referred to introductory material.

3. INFORMATION GEOMETRY

The largest source of inspiration for the work in this dissertation is the mathematical field of information geometry. In this part of the introduction, a brief overview of this discipline will be presented. The text of this chapter is based upon Amari's books [29,30], supplemented with elements of landmark articles in this field of research.

3.1 Parametrised statistical models

Information geometry is the study of statistics through the methods of differential geometry. This can happen either through the use of known results from differential geometry to expand the existing knowledge of statistics or by employing geometric methods to facilitate computations in statistical problems. Some examples of such applications are found in [35,40]. Of particular interest for this dissertation is the problem of parameter estimation. This problem occurs whenever one has quantitative data, which can be obtained from an experiment performed on a statistical sample, and one assumes that the true underlying distribution is an element of a parametrised set or a family of distributions. This assumption usually follows from a priori available information on the mechanism generating the data [38].
The most familiar example of this problem in physics is probably encountered when one has measured the energies of particles in a gas and wishes to determine from this the temperature of the gas, assuming that the probability of a particle to have an energy E when the gas has temperature T is given by the Boltzmann-Gibbs distribution

$$p_T(E) = \frac{1}{\mathcal{Z}} \exp\left\{-\frac{E}{k_B T}\right\}; \qquad \mathcal{Z} = \int_0^\infty \exp\left\{-\frac{E}{k_B T}\right\} \rho(E)\mathrm{d}E,$$

where ρ represents the density of states. Another familiar example is that of the normal distribution

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

where μ represents the mean of the distribution and σ is the standard deviation or second central moment which is used in many applications. In the rest of this introduction, the two-parameter Gaussian densities shall be used to illustrate the concepts discussed. The interested reader can find a more detailed discussion of multivariate Gaussian distributions and their information geometrical behaviour in [41].

In the above examples, the parameters T, μ and σ map homeomorphically onto the distributions they label. Hence, they can also be considered as coordinates for the family of distributions, which in turn can be viewed as a topological manifold. A manifold of statistical distributions (or equivalently: statistical measures) is called a statistical manifold, a term first introduced by Lauritzen [42].

3.2 Tangent spaces

It is common in information geometry to look at a particular, less abstract representation for the tangent spaces than the one used in texts on differential geometry. However, this less abstract representation directly inherits all structure from the general setting. The intuitive choice is to take as the basis vectors of the tangent space $\mathrm{T}_\theta\mathbb{M}$ not the partial derivative operators but rather the derivatives of a particular function of the probability density functions. The simplest choice is

$$\partial_i p_\theta(x) = \frac{\partial}{\partial\theta^i} p_\theta(x). \tag{3.1}$$

In the case of the Gaussian density functions, these derivatives become

$$\begin{aligned}\partial_\mu p_{\mu,\sigma}(x) &= p_{\mu,\sigma}(x)\left(\frac{x-\mu}{\sigma^2}\right)\\ \partial_\sigma p_{\mu,\sigma}(x) &= p_{\mu,\sigma}(x)\left(-\frac{1}{\sigma}+\frac{(x-\mu)^2}{\sigma^3}\right).\end{aligned}$$

Objects such as these are stochastic variables since they depend on x, which is an element of a measurable set. However, for reasons of convenience the practitioners of information geometry often choose a different representation of the tangent vectors. They choose as basis vectors the derivatives of a power-like function of the density function, called the α-representation of the tangent space. These objects are given by the expressions, for some $\alpha \in \mathbb{R}$,

$$\partial_i \ell_\alpha(x) = \begin{cases} \dfrac{2}{1-\alpha}\partial_i\left(p_\theta(x)^{\frac{1-\alpha}{2}}-1\right), & \alpha \neq 1,\\ \partial_i \ln p_\theta(x), & \alpha = 1.\end{cases}$$

The previously mentioned basis vectors (3.1) are obtained as a special case by choosing $\alpha = -1$. The different representations of basis vectors may seem unconventional at first sight. However, they are the conventional basis vectors to tangent spaces of other manifolds, those consisting of the stochastic variables $x \mapsto \ell_\alpha(x)$. These manifolds are very similar to the statistical manifold and so they can be studied in lieu of it. Because of the similarity, it is also possible to study one manifold while using the tangent spaces of another. It is interesting to note that the functions ℓ_α correspond to the q-logarithmic functions as introduced by Tsallis [43] through the relation $\alpha = 1 - 2q$ [30,44].

Where the most natural representation of the tangent spaces corresponds to the case $\alpha = -1$, another very convenient representation is found when $\alpha = 1$ is chosen. Not only does this representation have a clear relation to the log likelihood $\theta \mapsto \ln p_\theta(x)$, the tangent vectors in the 1-representation all have vanishing expectation value, as

$$\mathbb{E}_\theta[\partial_i \ln p_\theta] = \int_X p_\theta(x) \frac{\partial}{\partial \theta^i} \ln p_\theta(x) \mathrm{d}x = \frac{\partial}{\partial \theta^i} \int_X p_\theta(x) \mathrm{d}x = 0.$$

The tangent spaces in this representation are thus subsets of the set of stochastic variables whose expectation value vanishes in a sample described by p_θ. Exactly which subset this is depends on θ as well as on the particular statistical manifold under consideration.

This latest property illustrates an important perspective of information geometry: a parametrised family of distributions over some measurable set X is often explicitly thought of as a manifold embedded in the larger, convex set of all distributions over X. From that perspective it is can be seen that the tangent planes will depend on the choice of submanifold.[1] I will not offer a detailed discussion of the other representations in this introduction as their treatment is more elegant for sets of measures which need not be normalised (see for example Chapter 2.6 of [30] for an introduction to this topic).

For ease of reference, the α=1-representation is called the "exponential representation" due to its relation with the logarithm and the α=-1-representation is referred to as the "mixture representation".

[1] It is important to keep in mind this picture is only viable when the set X has a finite number of elements. For measurable sets X with an infinite number of elements, the simplex of all probability distributions over X cannot be considered a manifold in the strict sense of the definition as it is given in the previous chapter.

3.3 The Riemannian metric

In order to make the topological manifold $\mathbb{M}$ into a Riemannian manifold, a choice must be made for the inner product of the tangent spaces. Information geometry almost exclusively makes use of the Fisher information metric [38] for this purpose. This means the inner product of two tangent vectors $\vec{V}$ and $\vec{W}$ is defined to be equal to the covariance (in the statistical sense) of the stochastic variables V and W which correspond to the vectors, that is

$$g(\vec{V},\vec{W})|_\theta = \mathbb{E}_\theta[VW] - \mathbb{E}_\theta[V]\mathbb{E}_\theta[W]. \tag{3.2}$$

The most commonly used expression for the components of this tensor are found in the exponential representation, where they are given by

$$\begin{aligned} g_{ij}(\theta) &= \mathbb{E}_\theta[(\partial_i \ln p_\theta)(\partial_j \ln p_\theta)] \\ &= \int_X p_\theta(x)\left(\frac{\partial}{\partial\theta^i}\ln p_\theta(x)\right)\left(\frac{\partial}{\partial\theta^j}\ln p_\theta(x)\right)\mathrm{d}x. \end{aligned} \tag{3.3}$$

As an example, for a manifold of Gaussian distributions, the Fisher information matrix has components

$$g_{\mu\mu}(\mu,\sigma) = \frac{1}{\sigma^2}, \quad g_{\mu\sigma}(\mu,\sigma) = 0 \quad \text{and} \quad g_{\sigma\sigma}(\mu,\sigma) = \frac{2}{\sigma^2}.$$

In his 1925 article [38], Fisher first discussed desirable properties of estimators; functions of the data which are designed such that their values can be used to estimate the parameters of distributions describing the underlying data. He argues estimators must attain a fixed value when the sample size of the data is increased (“consistency”) and their variance multiplied by sample size must be minimal (“efficiency”). This efficiency is important as it lead Fisher to the introduction of the intrinsic precision of a probability density for the purpose of estimating a parameter θ. In particular, for a distribution p_θ over a set X he defines the quantity

$$g(\theta) = \int_X p_\theta(x)\left(\frac{\partial}{\partial\theta}\ln p_\theta(x)\right)^2 \mathrm{d}x,$$

which is known as the Fisher information and of which (3.3) is the higher-dimensional generalisation. Fisher himself describes this quantity as

> *the amount of information in a single observation belonging to such a distribution*,

by which he means the amount of information contained in this observation about the parameter(s) of that distribution [38]. The importance of the Fisher information to statistics, and to the problem of parameter estimation in particular, is contained in the Cramér-Rao theorem [39]. Unbiased estimators for a parameter θ^k are stochastic variables $\hat{\theta}^k$ for which $\mathbb{E}_\theta[\hat{\theta}^k] = \theta^k$. The Cramér-Rao theorem states that after making N observations of these estimators $\hat{\theta}^k$ in a population distributed according to p_θ, it holds that

$$\left\{\mathbb{E}_\theta[\hat{\theta}^i\hat{\theta}^j] - \mathbb{E}_\theta[\hat{\theta}^i]\mathbb{E}_\theta[\hat{\theta}^j]\right\} - \frac{1}{N}g^{ij}(\theta) \geqslant 0,$$

where the inequality means that the expression on the left hand side represents the components of a positive definite matrix. The Fisher information matrix thus expresses the minimal variance these estimators may attain. Furthermore the theorem shows there exists an estimator which attains this lower bound on its covariance as the number of observations tends to infinity. Such an estimator is said to be maximally efficient.

In the same paper [39] where he introduced the theorem which bears his name, Rao was the first to endow the statistical manifold with the Fisher information matrix as its Riemannian metric. This is possible as the covariance matrix (3.3) is positive definite and its components behave as those of a rank two tensor under a change of parameters—the general case of which follows from combining equations (2.2) and (2.5)—

$$\begin{aligned} g_{ij}(\theta) &= \frac{\partial\zeta^a}{\partial\theta^i}\frac{\partial\zeta^b}{\partial\theta^j}\int_X p_\zeta(x)\left(\frac{\partial}{\partial\zeta^a}\ln p_\zeta(x)\right)\left(\frac{\partial}{\partial\zeta^b}\ln p_\zeta(x)\right)\mathrm{d}x \\ &= \frac{\partial\zeta^a}{\partial\theta^i}\frac{\partial\zeta^b}{\partial\theta^j}g_{ab}(\zeta). \end{aligned}$$

An application of this transformation property is found in Bayesian probability. Since the Fisher information metric transforms as a rank two tensor, its volume form

$$\mathrm{vol}(\theta) = \sqrt{\det(g(\theta))}\;\mathrm{d}\theta^1 \wedge \ldots \wedge \mathrm{d}\theta^n$$

is unchanged under coordinate transformations. For this reason, Jeffreys suggested to use $\sqrt{\det g}$ as a non-informative (though possibly improper) prior on the parameter space [45].

As a last property of the Fisher information matrix to be discussed here is that it also appears in the second (and thereby lowest) order term of the expansion of the Kullback-Leibler divergence

$$D(p||p_\theta) = \int_X p(x)\ln\left(\frac{p(x)}{p_\theta(x)}\right)\mathrm{d}x, \tag{3.4}$$

which in the physical literature is better known as the relative entropy [4,5]. More precisely, for $\delta\theta$ sufficiently small it is possible to expand this divergence function as a function of the parameters θ to obtain

$$\begin{aligned}
&D_{KL}(p_\theta||p_{\theta+\delta\theta}) \\
&= -\int_X p_\theta(x)\ln\left(\frac{p_{\theta+\delta\theta}(x)}{p_\theta(x)}\right)\mathrm{d}x \\
&= -\int_X p_\theta(x)\ln\left(1+\frac{\delta p_\theta(x)}{p_\theta(x)}\right)\mathrm{d}x \\
&= -\int_X p_\theta(x)\left(\frac{\delta p_\theta(x)}{p_\theta(x)}-\frac{1}{2}\left(\frac{\delta p_\theta(x)}{p_\theta(x)}\right)^2+\mathcal{O}((\delta\theta)^3)\right)\mathrm{d}x \\
&= \frac{1}{2}\int_X \frac{1}{p_\theta(x)}\left(\frac{\partial}{\partial\theta^j}p_\theta(x)\right)\left(\frac{\partial}{\partial\theta^k}p_\theta(x)\right)\mathrm{d}x\ \delta\theta^j\delta\theta^k+\mathcal{O}((\delta\theta)^3).
\end{aligned}$$

The Fisher information thus expresses also the infinitesimal distance between nearby points on a manifold of probability distributions as expressed by the Kullback-Leibler divergence. This validates Rao's choice to endow statistical manifolds with the Fisher information matrix as their Riemannian metric.

3.4 The affine connections

Another important differential geometrical quantity is the affine connection with which a manifold can be endowed. The first attempt at investigating this structure for statistical manifolds was made by Rao. He studied the metric connection derived from the Fisher information metric and computed geodesic distances [39]. However, a statistical interpretation of this metric connection was not immediately obvious [29]. A breakthrough in the study of affine connections on statistical manifolds came in the 1970s with the work of Chentsov, Efron and Amari.

In his notably technical book [46] and the preceding articles, Chentsov showed that the space of multinomial distributions admits only a single statistically invariant Riemannian metric—the Fisher information metric—and a unique family of statistically invariant affine connections. The statistical invariance means that the geometric quantities defined through the metric tensor and the affine connections remain unchanged when the underlying probability space is mapped into another one through a Markov process. Chentsov's work, though published in Russian in 1972, was only translated into English ten years later. As a consequence it remained unknown outside the Soviet Union for some time after its initial publication.

The study of these affine connections outside the Soviet Union started in

1975, when Bradley Efron published the paper "Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency)" [47]. In this article, Efron studies one-parameter families of distributions, seen as curves through the space of all distributions over some measurable set. He implicitly defines a connection this space such that one-parameter exponential families coincide with geodesics, which are curves for which the tangent vector field is covariant constant along the curve. A number of commentary texts have been published together with the article. The last two of these, written by Dawid and Reeds, are probably the most important.

Exponential families are parametrised sets of distributions over a (measurable) set X for which there exist functions $\mathcal{H}_k : X \to \mathbb{R}$ and a function $\Phi : \mathbb{R}^n \to \mathbb{R}$ such that it is possible to write

$$p_\theta(x) = \exp\left\{-\Phi(\theta) - \sum_{k=1}^{n} \theta^k \mathcal{H}_k(x)\right\} \tag{3.5}$$

with respect to some given measure [4, 47]. The parameters θ are called the canonical parameters of the exponential family. The function Φ is called the Massieu function [48]. Since the distribution p_θ must be normalised, the function Φ satisfies

$$\Phi(\theta) = \ln \int_X \exp\left\{-\sum_{k=1}^{n} \theta^k \mathcal{H}_k(x)\right\} \mathrm{d}x.$$

In what follows it is always assumed that the values of the parameters θ are such that $\Phi(\theta) < \infty$.

The Gaussian distributions serving as examples in this chapter belong to the exponential family with parameters

$$\theta^1 = \frac{1}{2\sigma^2} \quad \text{and} \quad \theta^2 = -\frac{\mu}{\sigma^2}$$

and Hamiltonians

$$\mathcal{H}_1(x) = x^2 \quad \text{and} \quad \mathcal{H}_2(x) = x.$$

This can be seen from rewriting the expression for the Gaussian density into an expression into the form (3.5).

The focus of Efron's paper is on one-dimensional subsets S of exponential families, that is subsets for which it holds that

$$S = \{p_\theta(x) | \theta = F(\eta), \eta \in N \subset \mathbb{R}\}.$$

As Reeds points out in his commentary [49], Efron implicitly makes a number of assumptions about the higher-dimensional exponential family. In particular, he takes this set to be a Euclidean space, endowed with a constant metric tensor equal to the value of the Fisher information metric at θ_0, the true value of the parameter (or parameters) that one seeks to estimate. This assumption allows Efron to consider the one-dimensional submanifold S as a curve through Euclidean space and to compute the extrinsic geometrical curvature thereof. This curvature is defined as the length of the vector

$$\left.\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}\right|_{\boldsymbol{p}}, \tag{3.6}$$

where $\vec{T}$ is the tangent of unit length to the set S and s is the arc length of the curve, measured from a point which can be chosen arbitrarily. He then defines the "statistical curvature" of S to be equal in value to this geometrical curvature. When the curvature (3.6) does not vanish, the one-dimensional family is said to be curved in the statistical sense as well as in the geometrical sense. Such a family is therefore called a curved exponential family.

The computation of geometrical curvature through formula (3.6) requires no less than three separate uses of the Riemannian metric (once for the definition of the arc length and twice to compute the length of a vector) as well as a covariant derivative (hidden as a regular derivative with respect to the arc length). This means Efron's assumptions of dealing with a Euclidian space and of the metric tensor having constant components in the coordinate system of canonical parameters are important.

The true interesting point of Efron's work lies not in his definition of the statistical curvature as a quantity itself but in the realisation that the square of this curvature plays a crucial role in the properties of the Fisher information metric [47]. This establishes a connection between the properties of a statistical estimation problem and the geometry of the set of distributions that is used to model the data. The relation of Efron's work with affine connections is that (non-curved) one-dimensional exponential families coincide with geodesics through the larger space in which they are embedded. It was Dawid who identified the connection implicitly used by Efron in the definition of the geodesics [50]. This connection would later be known as the "exponential connection" and is a member of the family identified by Chentsov. Dawid suggested also the "mixture connection" in his reply and remarks that both connections are torsionless as well as flat. It should be stressed that these are quantitatively different connections and not different representations of the same connection—despite what the naming similar to that of the exponential and mixture representations of the tangent spaces

may suggest. Furthermore, Dawid briefly touches Chentsov's family of connections but he does not study their properties in his reply.
Another thing remarked by Reeds in his commentary is that the curvature used by Efron to define statistical curvature is the extrinsic or embedding curvature. This is a subtle but important point as it demonstrates the necessity of considering the exponential family to be embedded in a larger, Euclidean space. In information geometry, this space is the set of all distributions if it is finite dimensional. In the work presented in this dissertation, however, an analogous larger space cannot be specified in general, even though specific examples usually do allow for it. Reeds also initiates the work on a higher-dimensional generalisation of Efron's work. Many authors cite L. T. Madsen as having completed this task in her doctoral dissertation and the subsequent 1979 paper [51]. Unfortunately, this article seems to be very difficult to obtain, perhaps since it has not appeared in a peer reviewed journal but rather as a research report for the Danish Medical Society. For this reason, it is difficult to tell exactly which of the advances were made by dr. Madsen.
The work of Amari advanced that of Chentsov, Efron, Dawid and Madsen by introducing a differential geometric framework for the construction of higher-order asymptotic theory of statistical inference (see for instance [30,52,53]) in which the family of connections introduced by Chentsov plays an important role.
These affine connections are usually presented in a relatively technical way but they have very simple interpretations. This was pointed out by Dawid already in this reply to Efron's paper [50] but this piece of knowledge is—unfortunately—often ignored in the rest of the introductory literature on the subject. Remember that affine connections serve to define covariant derivatives and thereby parallel transport. In the exponential representation, the elements of a tangent space $\mathrm{T}_\theta\mathbb{M}$ are stochastic variables V such that

$$\mathbb{E}_\theta[V] = \int_X p_\theta(x)V(x)\mathrm{d}x = 0.$$

However, there is no guarantee that the expectation value of $V \in \mathrm{T}_\theta\mathbb{M}$ will also vanish when evaluated at another distribution, that is $\mathbb{E}_\xi[V] = 0$ cannot be guaranteed when $\xi \neq \theta$. Consequently, parallel transport need not preserve the vanishing of the above expectation value. The family of affine connections of information geometry solves this problem by defining parallel transport in such a way that the expectation value of a stochastic variable remains zero when it undergoes parallel transport. In particular, the exponential connection ($\alpha = 1$) defines the operation $\Pi^{(1)} : \mathrm{T}_\theta\mathbb{M} \to \mathrm{T}_\xi\mathbb{M}$ by explicitly subtracting the expectation value in the end point:

$$\Pi^{(1)}V = V - \mathbb{E}_\xi[V].$$

The mixture connection ($\alpha = -1$) on the other hand multiplies the statistic with the appropriate Radon-Nykodym derivative (see for instance [46] or an introductory work on the matter), that is $\Pi^{(-1)} : \mathrm{T}_\theta\mathbb{M} \to \mathrm{T}_\xi\mathbb{M}$ works as

$$\Pi^{(-1)}V = V\frac{p_\theta}{p_\xi}.$$

This means the expectation value equals

$$\begin{aligned}\mathbb{E}_\xi[\Pi^{(-1)}V] &= \int_X V(x)\frac{p_\theta(x)}{p_\xi(x)}p_\xi(x)\mathrm{d}x \\ &= \int_X V(x)p_\theta(x)\mathrm{d}x \\ &= \mathbb{E}_\theta[V] \\ &= 0.\end{aligned}$$

The other affine connections of the α-family are simply linear combinations of the exponential and mixture connections, in particular

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}.$$

A similar interpretation exists for the other representations of the tangent space but this requires the introduction of the notion of escort probability distributions [30]. This is closely related to the work of Tsallis and an extensive introduction of escort probabilities can be found in [4] but to treat this explicitly would fall outside the scope of this introduction.

Even though the exponential and mixture connections give rise to a path-independent definition of parallel transport and are thus flat, the rest of the α-family has a constant but non-vanishing scalar curvature (also known as the Ricci scalar of the connection), which is given by [29]

$$R \overset{\text{def.}}{=} g^{ij}\Omega^k{}_{ikj} = \frac{1-\alpha^2}{4}.$$

The metric connection associated with the Fisher information metric is obtained in the case $\alpha = 0$ and so it is the most (positively) curved member of this family. As is elucidated in the previous chapter, the metric connection is the unique torsionless affine connection satisfying

$$\vec{X}(g(\vec{Y},\vec{Z})) = g(\nabla_{\vec{X}}\vec{Y},\vec{Z}) + g(\vec{Y},\nabla_{\vec{X}}\vec{Z}). \qquad \text{(2.15 revisited)}$$

This property only holds when $\alpha = 0$ and in general it can be shown that for any $\alpha \in \mathbb{R}$

$$\vec{X}(g(\vec{Y}, \vec{Z})) = g(\nabla^{(\alpha)}\vec{X}, \vec{Y}) + g(\vec{X}, \nabla^{(-\alpha)}\vec{Y}), \tag{3.7}$$

where g represents the Fisher information metric. This shows that the α- and $-\alpha$-connections are dual with respect to this metric. Dual connections play an important role in the more advanced topics of information geometry. One application is found in a generalisation of the Pythagorean property, which holds in a triangle where one leg is a geodesic for the exponential connection and the other leg is a geodesic for the mixture connection [30].

3.5 Divergence functions

Due to the importance in information geometry of the relative entropy (3.4), a central role in this dissertation shall be played by divergence functions, or as they are sometimes also called, contrast functions. They shall serve as the elementary structure from which all geometric objects shall be constructed, just as all the Riemannian geometry of statistical manifolds elucidated above can be derived from the divergence function of Kullback and Leibler.

The Kullback-Leibler distance is a function quantifying in a certain sense the difference between statistical distributions over a measurable set X. As was mentioned earlier in this chapter, it is given by the expression

$$D(q||p) = \int_X q(x) \ln\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x.$$

It should be noted that this integral is always defined, even though its value may be infinite, as the measures from which p and q are derived are finite [5]. The relative entropy plays an important role not only in statistics but also in information theory. In fact, it was introduced by Kullback and Leibler as an abstraction of Shannon's entropy

$$S(p) = -\int_X p(x) \ln p(x) \mathrm{d}x, \tag{3.8}$$

which in turn was introduced for purposes of information theory in [54], even though the expression also appears in the work of Boltzmann and Gibbs. The quantity

$$\ln\left(\frac{q(x)}{p(x)}\right)$$

is a measure for the information obtained in the result of a measurement x for discrimination between the hypotheses "x is distributed according to the distribution q" and "x is distributed according to the distribution p". The Kullback-Leibler divergence is therefore equal to the mean information for discrimination [5] and it is in this sense that it can be said to quantify the difference between its arguments.

A commonly used set of contrast functions are the f-divergences of Csiszár [55]. They are of the form

$$D_f(q||p) = \int_X p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x,$$

where f is a convex function for which $f(1) = 0$. The Csiszár f-divergences are the largest class of statistically invariant divergence functions. Another often made choice is the set of Bregman divergences [56]. The most general definition of these divergences is not limited to statistical distributions but when restricted thereto, it is possible to write them in the form

$$D_F(q||p) = \int_X \int_{p(x)}^{q(x)} [F'(u) - F'(p(x))] \mathrm{d}u \mathrm{d}x.$$

where F is a strictly convex function. These are also known as U-divergences, after their definition by Eguchi [57]. It can be shown that the Kullback-Leibler divergence is the only contrast function which belongs to both the classes of Csiszár and Bregman divergences [29].

In order to gain more insight in general divergence functions, it is interesting to compare them to the more familiar metrics—not to be mistaken with metric tensors. A metric on a set S is a function

$$d : S \times S \to \mathbb{R}^+$$

which is zero everywhere on the diagonal of $S \times S$ and strictly positive elsewhere, is symmetric ($d(x, y) = d(y, x)$) and satisfies the triangular inequality

$$d(x, y) + d(y, z) \geqslant d(x, z).$$

Divergence functions on S on the other hand are only required to satisfy the first condition. An example of a symmetric divergence is the *squared* Euclidean distance in $\mathbb{R}^3$, which satisfies the cosine rule instead of the triangular inequality:

$$d(x, y)^2 + d(y, z)^2 = d(x, z)^2 + 2d(x, y)d(y, z)\cos(\phi),$$

with ϕ the angle between the legs of the triangle meeting at the point y.
Once an appropriate divergence function D over a manifold $\mathbb{M}$ has been chosen, it can be used to define a differential geometric structure upon $\mathbb{M}$, see for example [58, 59]. In order for this to be possible, the divergence must be sufficiently many times continuously differentiable with respect to the coordinates of its arguments and this at least in a neighbourhood of the diagonal of $\mathbb{M} \times \mathbb{M}$. A lot of interesting work in this context has been due to Eguchi and collaborators [57, 60, 61] but also due to Amari and collaborators [52, 53, 59]. In their research, they have investigated the geometric structure of statistical manifolds as well as more general manifolds endowed with divergence functions and applications thereof.
The first geometric structure to be introduced is, as usual, the metric tensor. Since divergence functions over a manifold $\mathbb{M}$ are zero everywhere on the diagonal of $\mathbb{M} \times \mathbb{M}$ and only there, it must automatically hold that the lowest order term in its Taylor expansion is the second order term, that is

$$D(m_\theta||m_{\theta+\delta\theta}) = \frac{1}{2}\frac{\partial^2}{\partial\xi^i\partial\xi^j}D(m_\theta||m_\xi)\bigg|_{\xi=\theta}\delta\theta^i\delta\theta^j + \mathcal{O}((\delta\theta)^3).$$

The coefficient of this lowest order term—without the factor $\frac{1}{2}$—is the metric tensor induced by the divergence:

$$g_{ij}(\theta) = \frac{\partial^2}{\partial\xi^i\partial\xi^j}D(m_\theta||m_\xi)\bigg|_{\xi=\theta}. \tag{3.9}$$

This is a positive definite quantity as it is the matrix of second derivatives of a function in a local minimum. Despite this, it behaves properly under a coordinate transformation, as can be seen also from the alternative expression

$$g_{ij}(\theta) = -\frac{\partial^2}{\partial\xi^i\partial\theta^j}D(m_\theta||m_\xi)\bigg|_{\xi=\theta},$$

which can be shown to hold since the first derivatives of the divergence vanish identically on the diagonal.
Affine connections can be constructed from divergence functions just as well. In particular, it is possible to consider the pair of mutually dual connections ∇ and ∇^* defined through the expressions [59, 60]

$$g_{ks}(\theta){\omega^s}_{ij}(\theta) = -\frac{\partial^3}{\partial\theta^i\partial\theta^j\partial\xi^k}D(m_\theta||m_\xi)\bigg|_{\xi=\theta}$$

and

$$g_{ks}(\theta){\varpi^s}_{ij}(\theta) = -\frac{\partial^3}{\partial\xi^i\partial\xi^j\partial\theta^k}D(m_\theta||m_\xi)\bigg|_{\xi=\theta}. \tag{3.10}$$

For more background regarding the geometry induced by divergence functions, the reader is referred to the work of Amari and Eguchi cited in the text above.

3.6 Applications in thermodynamics

The material introduced so far in this chapter is almost exclusively based upon the mathematical literature. However, it deserves mentioning that some of these geometrical aspects have also been adopted by the community of researchers in thermodynamics. Since the necessary background for the following chapters has already been introduced, this overview shall be kept fairly short. Nevertheless, the historical development of this particular field of research shows some interesting parallels with the goal of this dissertation. Indeed, both are instances where geometry is applied in an attempt to unearth the mathematical foundations of a physical theory of which the basics are less rigorous.

The first connection between differential geometry and thermodynamics was made by Constantin Carathéodory, who sought to establish an axiomatic basis for thermodynamics. He chose to express this in terms of differential geometry [62,63]. In these papers he could phrase thermodynamics on sound mathematical principles, rather than on the more usual references to imaginary devices such as Carnot engines or to concepts such as the flow of heat.[2] For this, he works on the topological manifold of thermodynamic states of the system. His rendering of the Second Law states [28,64]

> *In every neighbourhood of every equilibrium state x, there are states y that are not accessible from x via quasi-static adiabatic paths.*

This formulation is weaker than Kelvin's better known one, which states that no cyclical process can exists which turns heat into its mechanical equivalent of work. Starting from this axiom, Carathéodory could derive thermodynamics as it was known in his time. His results therefore extend those of Helmholtz, who had already noticed that a definition of temperature or entropy does not require cyclical processes or ideal gasses [64].

The next step in building a geometric basis for thermodynamics and statistical mechanics comes from Tisza [65] and Griffiths and Wheeler [66], whose contributions may be deemed important more for steering geometric thermodynamics away from the difficult formalism of Carathéodory than for

[2] It is interesting to note that Carathéodory introduces heat as a derived rather than as a fundamental quantity, which is the approach of conventional thermodynamics.

their actual results. A few years later, Weinhold published a series of papers regarding the metric geometry of equilibrium thermodynamics [67]. His main contribution, at least in the context of this discussion, is to endow the equation of state surface appearing in the Gibbs formulation of equilibrium thermodynamics with a Riemannian metric. This is a completely different approach from that of Carathéodory, which is based primarily on differential 1-forms over the set of states. Weinhold's metric has components given by

$$g_{ij} = \frac{\partial^2 U}{\partial N^i \partial N^j}$$

where U is the internal energy and the set of N^i contains the independent conserved extensive quantities of the system (such as volume and particle numbers) and the entropy. This is a matrix containing quantities related to standard thermodynamic linear response functions such as compressibility and specific heat. Due to the convexity of the internal energy in single phase regions, the metric g is positive definite. Application of the Cauchy-Schwarz and Bessel inequalities allowed Weinhold to derive many of the standard thermodynamical inequalities. He did not, however, compute distances between points of the equilibrium surface [68].

Four years later, Ruppeiner published an interpretation of the metric structure and introduced an intrinsic rather than an extrinsic geometry [69]. He does this not by endowing the equilibrium surface with a metric but rather by introducing the metric on an abstract manifold of equilibrium states. Apart from this difference, his metric tensor ($g^{(R)}$) is related to the one introduced by Weinhold ($g^{(W)}$) through a conformal transformation:

$$g^{(R)} = Tg^{(W)}.$$

The interpretation Ruppeiner gives to this metric tensor is very simple. It expresses the distance between neighbouring states in the sense that the more likely a fluctuation bringing one equilibrium state into another is to occur, the closer they are in the Ruppeiner metric. A related result is that the geodesic distance is related to the diffusion of the system through state space by fluctuations.

Furthermore, Ruppeiner argues that curvature exhibited by the metric connection is due to interactions in the fluid, as the interactionless ideal fluid gives rise to a geometry exhibiting no curvature. He goes on to expand upon this idea by arguing that the curvature is proportional to the cube of the correlation length of the system. He does this through a line of reasoning based on dimensional analysis and scaling relations. This leads to the identification of universal constants, a result supported by experimental evidence (see [69] for the details).

The metric tensors introduced by Weinhold and Ruppeiner give rise to a measure of distance sometimes known under the name of thermodynamic length (see for instance [70] and references therein). Nevertheless, there is also criticism on the view that it is possible to see this metric tensor as suitable to define a meaningful distance between points on the equation-of-state surface. This was elucidated by Gilmore [71]. He showed that also using the positive definite quantity introduced by Weinhold—which is the second fundamental form of this surface—as the metric tensor leads to constraints on the third derivatives of the internal energy of a system. Such constraints, however, are nowhere to be found in thermodynamics.

Another interesting result obtained by the application of geometrical considerations is found in other work by Gilmore [72]. Basing himself on fluctuation theory just like Ruppeiner and applying the Cramér-Rao bound (which is nothing but a consequence of the Cauchy-Schwarz inequality from a geometrical point of view), he was able to obtain uncertainty relations for thermodynamical quantities in a system undergoing fluctuations. More precisely, if a system is in equilibrium with a reservoir with well-defined intensive variables, then variations in the measured values of the system's extensive variables will lead to variations in the estimation of the reservoir's corresponding intensive variables. Gilmore's result states that the product of the variances of the distributions of these quantities must be larger than a constant, for example

$$\Delta U \Delta \frac{1}{T} \geqslant k_B,$$

where the right hand side here represents Boltzmann's constant. While this particular example was already known by Gibbs [73], Gilmore's line of reasoning can be applied to any conjugate pair of intensive or extensive variables. Furthermore, it is possible to show that these relations are equivalent to the stability criteria of equilibrium thermodynamics (see again [72] for the details).

4. THE DATA SET MODEL FORMALISM

This is the central chapter of this dissertation. In it, the development of the data set model formalism is outlined. First, the different elements which must be present in order for the formalism to be applicable are outlined and discussed in detail. Afterwards, the geometry of the data set models will be constructed and the general consequences investigated. Concrete examples and applications are examined in the next chapter.

4.1 The elements of the formalism

4.1.1 The data sets

Consider a set $\mathbb{X}$ of mathematical objects, called data sets, which one would like to model by representing them by the elements of a parametrised set. In principle $\mathbb{X}$ can be any set but the work set out herein will assume $\mathbb{X}$ to be endowed with a topology. This is truly an assumption and not a demand; the formalism can work without such structure. However, it is expected that when there is a topology on the set $\mathbb{X}$, as is often the case, one will desire that this topology is respected by the modelling process. This will be accommodated by the data set model formalism through for instance the demand that the mapping from data to model point is continuous. In order to make the discussion of this aspect meaningful, it is thus assumed that $\mathbb{X}$ has indeed been endowed with a topology.

In statistical physics, the data sets are elements of the simplex of distributions over a measurable set—they represent the empirical data obtained from an experiment [4]. Another often encountered example is a collection of measurements for which a functional relation is to be found, as is for example the case in linear regression. In a quantum mechanical setting, the data to be modelled could be data obtained through measurements performed on an ensemble of states. An example from machine learning could be a fingerprint acquired at the scene of a crime and of which the general class has already been determined but which still needs to be characterised to be stored efficiently in a database through which to search in future investigations.

4.1.2 The model points

The data sets in $\mathbb{X}$ are intended to be modelled by an element m of a parametrised set, which will be called a model point. In order for the methods set out in this dissertation to be applicable, this parametrisation must be a homeomorphism. As such, the parameters can also play the role of coordinates and the model points will constitute a topological manifold. The symbol $\mathbb{M}$ will henceforth be used to denote the manifold of models—and no longer an arbitrary manifold. When the parameters are represented by θ, the points of the manifold will often be denoted as m_θ, in analogy with the parametrised distributions p_θ. As it is the the case in information geometry, it is important to keep in mind that the number of parameters must be finite in order to practice differential geometry with the particular mathematical techniques used here. This finite number, which by definition is also the dimension of the manifold $\mathbb{M}$, is represented by the letter n in this chapter. Examples of infinitely dimensional models would be the Hilbert space of some quantum systems such as the harmonic oscillator, models parametrised by response functions… .

Perhaps the most common example of such a model is the family of first order polynomials

$$\{f|f(x) = ax + b,\ (a, b) \in \mathbb{R}^2\}$$

used to fit data points in a plane when practising linear regression. In a quantum setting, parametrised subsets of the Hilbert space provide an obvious choice for the models of the formalism. Earlier published work in this context considered the coherent states of the (one-dimensional) quantum harmonic oscillator, which is a two-dimensional family of the most classical states of that quantum system [74]. In the above example of the fingerprint, once the general class has been determined, the exact ridge pattern could be described by a limited number of continuously varying parameters indicating the positions of characteristic points of the pattern.

4.1.3 The model map

The actual process of modelling, determining which model point is chosen to represent a given data set, happens by means of the model map μ. This map is required to be continuous (when a topology on $\mathbb{X}$ has been chosen) as a small change in the data set should never cause a large change in the model chosen to represent it.

Furthermore, the model map and the divergence function must be compatible. This condition will be elucidated in the upcoming discussion of the divergence. A consequence of this is, however, that the domain of μ may be limited to a subset of $\mathbb{X}$. Therefore it is more correct to identify $\mathbb{X}$ as the domain of μ and to take into account the possibility that this domain is in fact a subset of a larger set. For the sake of succinctness this remark will not be reiterated explicitly in those instances where the distinction is clear from context. A similar remark holds for the image of the model map; the geometric quantities that are introduced can only be defined on the (topological) closure of the image of μ. In order for the data set model to be applicable, this closure must thus be a manifold in its own right.

The model map can be assumed not to be injective, as it would make the process of modelling redundant. The notation $\mu^{-1}(m)$ will therefore be used to indicate the collection of those data sets x for which $\mu(x) = m$. These subsets of $\mathbb{X}$ upon which μ attains a constant value will be called fibres. For reasons of convenience the metaphor of fibre bundles is also used in saying a model point m is the projection of a data set x, by which again it is meant that $\mu(x) = m$. It should be noted that this nomenclature is suggestive rather than rigorous. The set $\mathbb{X}$ can be likened to the total space of a fibre bundle, the manifold $\mathbb{M}$ to the base space and the model map μ to the projection but together these objects cannot be properly called a fibre bundle. The definition of a proper fibre bundle requires the fibres to be copies of the same space [28, 37] and this is not generally the case for the sets $\mu^{-1}(m)$.

In certain contexts, such as a physical experiment, one may not have access to the data sets itself but only to a limited list of observable quantities. When this occurs, the data sets may acquire extra structure. In the simplest case the data sets can only be described by a finite number of real numbers, in which case $\mathbb{X}$ will adopt a manifold structure. Also the divergence function may induce more properties upon the set $\mathbb{X}$. Though this may give rise to specific consequences, these were not investigated explicitly and hence this dissertation does not report upon results concerning such additional structure.

4.1.4 The divergence

A crucial element in the modelling process considered is a divergence function, which is consistently denoted as D. The value $D(x||m)$ is a measure of how well the model point $m \in \mathbb{M}$ describes the data set $x \in \mathbb{X}$ and where a smaller value indicates a better match. As such, it can be thought of as a “badness of fit”. An alternative viewpoint is that $D(x||m)$ represents the cost of describing x by means of m. Such perspective can be found treated in more detail for instance in the work of Topsøe [75–77] but this idea is not actively entertained in this dissertation.

The divergence function must satisfy four conditions, most of which are in some way related to the properties of divergence functions as they were treated in the previous chapter. These conditions are:

1. The domain of D contains $\mathrm{Dom}(\mu) \times \mathrm{Im}(\mu)$. In many practical problems the divergence will imply the model map and its properties, making this condition largely trivial.

2. The divergence is sufficiently many times continuously differentiable with respect to the parameters of its second argument. This demand must hold for each of the different coordinate systems used.

3. Given any $x \in \mathrm{Dom}(\mu)$, the function

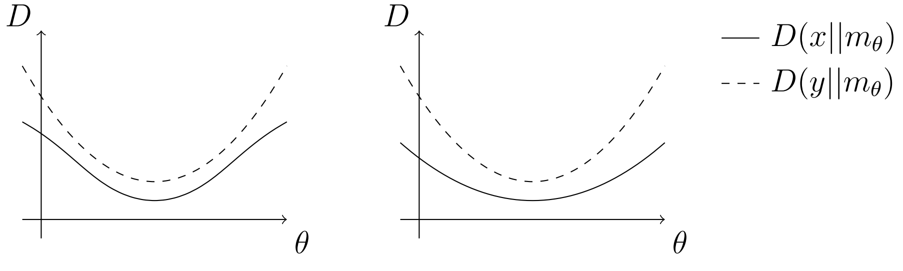
$$m \mapsto D(x||m)$$

has a local minimum at $m = \mu(x)$. This is the condition of compatibility of the model map and the divergence that was mentioned earlier. That $\mu(x)$ provides a unique global minimum is only strictly required for the generalised Pythagorean theorem. Nevertheless, this will be assumed to hold as well.

4. For all $m_\theta \in \mathrm{Im}(\mu)$, the function

$$x \mapsto \partial_i \partial_j D(x||m_\theta)$$

has a constant value on the fibre of m_θ. The drawings illustrate this schematically for two divergence functions on a 1-dimensional $\mathbb{M}$. For both drawings, two data points x and y within the same fibre are chosen. In the leftmost drawing, the curvature of both graphs in the minimum is the same, though this is not the case elsewhere, and the condition is satisfied. In the rightmost drawing, the curvature is not the same in the minimum even though the graphs may seem more similar—they are both parabolas—and the condition is not satisfied.



From the assumption of a global minimum follows that the divergence is bounded from below, at least on $\mathrm{Dom}(\mu) \times \mathrm{Im}(\mu)$. Without loss of generality, this function can then be assumed to attain positive values only. In previous work [74,78] a stronger condition was imposed but this turns out to be largely unnecessary. Furthermore, it is often convenient to impose extra conditions, depending on the presence of a topology for $\mathbb{X}$.

5. When a topology for $\mathbb{X}$ has been specified, it will be assumed that the divergence is also continuous in its argument. (The continuity in the second argument is already implied by Condition 2.)

6. Given any $x \in \mathrm{Dom}(\mu)$, there exists a neighbourhood $\mathcal{N}_x$ of x wherein there exist data sets $x^{(l)}$ such that the numbers

$$\partial_k D(x^{(l)}||\mu(x))$$

make up the components of an invertible matrix.
A stronger version of this condition is that in this neighbourhood, the function $y \mapsto \partial_k D(y||\mu(x))\mathrm{d}\theta^k$ is continuous and has an image homeomorphic to a subset of $\mathbb{R}^n$. Then there exist functions X^l mapping a neighbourhood of 0 ($\in \mathbb{R}$) to $\mathcal{N}_x$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_k D(X^l(\varepsilon)||\mu(x))\bigg|_{\varepsilon=0} = \delta^l_k.$$

This is a construction similar to one used in the study of gradient flows on metric spaces—see for instance [79] for an introduction to that discipline.

At some points in this dissertation, it will be necessary to distinguish between the definition of a divergence function given here and the more restrictive definition that is presented in the introduction. To make the distinction where confusion may arise, the latter kind will sometimes be referred to as a "proper divergence (function)".
Together the sets and maps $\mathbb{X}$, $\mathbb{M}$, μ and D make up what will called the data set model $(\mathbb{X}, \mathbb{M}, \mu, D)$. In what follows, the geometry of these models will be studied and it will be elucidated how this geometry gives rise to a method to easily determine whether or not a given parametrised family of distributions belongs to an exponential family.

4.2 The geometry of data set models

4.2.1 Topology

For the sake of completeness, it is worthwhile to consider for a moment the topologies of the set $\mathbb{X}$ and the manifold $\mathbb{M}$. Many, if not all, important functions in this dissertation have these two sets (or at least subsets thereof) as either their domain or as their co-domain and many of these functions can be demanded to be continuous. Examples thereof include the model map ($\mathbb{X} \to \mathbb{M}$), the divergence function ($\mathbb{X} \times \mathbb{M} \to \mathbb{R}$) and the coordinate functions ($\mathbb{M} \to \mathbb{R}^n$).

That $\mathbb{M}$ is a manifold is actually a property of its topology. In particular, the local homeomorphism of $\mathbb{M}$ with $\mathbb{R}^n$ is equivalent to stating that neighbourhoods of model points in $\mathbb{M}$ are, from a topological point of view, Tychonoff spaces [34].

4.2.2 The Riemannian metric

In order to induce a geometric structure on the manifold of models, a metric tensor must be constructed. This can be done in analogy with the Fisher information metric and the metric tensor derived from divergence functions. In those examples, the components of the metric tensor are second derivatives of the divergence. The geometric interpretation of this is a generalisation of Fisher's ideas as they were discussed in the previous chapter.

Fix a data set x and consider the surface defined by the graph of the map

$$\theta \mapsto D(x||m_\theta)$$

on a neighbourhood of the coordinates of $\mu(x)$. In order to obtain a good measure of information in Fisher's sense, it is required to look at how strongly peaked the graph is around its minimum as this encodes the sensitivity of the divergence for small variations in the chosen model point. This sharpness is nothing else than the extrinsic or embedding curvature of the surface in $\mathbb{R}^{n+1}$ and this can be quantified using the theory of surface geometry. In particular, it is encoded in the second fundamental form and it can be shown that in the minimum, the curvature is quantified by the values

$$\partial_i \partial_j D(x||\mu(x)).$$

The logical choice for the metric tensor would therefore be an expression of the form

$$g_{ij}(\theta) = \partial_i \partial_j D(x||m_\theta) \quad \text{with} \quad \mu(x) = m_\theta. \tag{4.1}$$

This expression is only well-defined, however, if the divergence satisfies Condition 4 set out earlier in this chapter. The tensor (4.1) will be called the generalised Fisher information metric. Although the it is defined as a second derivative, which in general does not behave under coordinate transformations in the way a metric tensor ought to, no problems arise with this definition.

Theorem 1. *The definition (4.1) for the metric tensor is invariant under coordinate transformations.*

Proof. Choose $x \in \mu^{-1}(m_\theta)$ and a transformation from the coordinates $\{\theta^i\}$ to the coordinates $\{\zeta^a\} = \{Z^a(\theta)\}$. Then the components of the metric tensor transform as

$$\begin{aligned} g_{ij}(\theta) &= \partial_i \partial_j D(x||m_\theta) \\ &= \frac{\partial Z^a}{\partial \theta^i}\frac{\partial Z^b}{\partial \theta^j}\partial_a \partial_b D(x||m_{Z(\theta)}) + \frac{\partial^2 Z^a}{\partial \theta^i \partial \theta^j}\partial_a D(x||m_{Z(\theta)}) \\ &= \frac{\partial Z^a}{\partial \theta^i}\frac{\partial Z^b}{\partial \theta^j}\partial_a \partial_b D(x||m_{Z(\theta)}) \\ &= \frac{\partial Z^a}{\partial \theta^i}\frac{\partial Z^b}{\partial \theta^j} g_{ab}(Z(\theta)). \end{aligned}$$

This completes the proof. □

The expression (4.1) achieves this coordinate invariance without changing the second derivative into a Hessian, which would require a connection to be chosen. Since such a choice would necessarily be arbitrary at this point, it is a benefit of this approach that this scenario can be avoided. As such, a metric can be defined under relatively weak conditions; it will be shown shortly that the introduction of the affine connection places a stronger demand on the divergence.

The Kullback-Leibler divergence satisfies Condition 4 when $\mathbb{M}$ is an exponential family of probability distributions, since—with $\{\theta^i\}$ the canonical parameters—

$$\begin{aligned} \partial_i \partial_j D(p||p_\theta) &= -\frac{\partial}{\partial \theta^i}\int_X p(x)\frac{\partial}{\partial \theta^j}\ln p_\theta(x)\ \mathrm{d}x \\ &= \partial_i \partial_j \Phi(\theta). \end{aligned}$$

This is obviously a strong way to satisfy this condition, as the derivatives of the Massieu function Φ are completely independent of the distribution p. However, it is not necessary for the existence of a Riemannian geometry that such a strong condition is imposed. An example of a data set model outside the field of information geometry allowing for such a metric tensor is that of the grand canonical ensemble of bosonic particles, treated in detail in the next chapter. The Gumbel distributions discussed shortly thereafter are an example of a family of distributions which do not satisfy Condition 4 when the same Kullback-Leibler divergence is used.

4.2.3 On the properties of exponential families

After a metric tensor has been constructed for a data set model, the next geometric structure to be treated is the affine connection. In order to guide the development of the data set model formalism, it is useful to first discuss particular properties of exponential families.

As is mentioned in the introduction, Efron introduced the statistical curvature of one-dimensional submanifolds of exponential families. This was generalised to higher dimensions, first by Reeds and then by Madsen and Amari. A recurring property in their work is that the manifold of distributions belonging to an exponential family is flat when it is endowed with the exponential connection. (Amari has also shown similar results for generalised families, see [29].) This is an important element of the argument for the choice of affine connection for data set models.

Exponential families are parametrised sets of distributions that can be written in the canonical form

$$p_\theta(x) = \exp\{-\Phi(\theta) - \theta^k \mathcal{H}_k(x)\}. \tag{4.2}$$

This means that for any distribution p, the second derivative of the Kullback-Leibler divergence evaluated in the pair (p, p_θ) satisfies

$$\partial_i \partial_j D(p||p_\theta) = \partial_i \partial_j \Phi(\theta),$$

still in coordinates coinciding with the canonical parametrisation. Using arbitrary coordinates $\{\zeta^a\}$ and the notation of data set models, this becomes

$$\nabla_a \nabla_b D(x||m_\zeta) = \nabla_a \nabla_b \Phi(\zeta) \tag{4.3}$$

for all $x \in \mathbb{X}$. That the Hessian of the divergence function does not depend its first argument is an important property since it enables the construction of an affine connection for the model manifold in the general formalism. As an additional advantage, it will be shown that the property (4.3) cannot hold for an arbitrary connection. In fact that connection is unique—a claim the proof of which will be provided by the explicit construction of the coefficients in the next subsection.

Its uniqueness is not the only interesting property this connection exhibits however. It turns out that equation (4.3) gives rise to a rich geometry. For this reason, data set models satisfying this equation are said to exhibit a Hessian structure. Another name could be “Legendre structure”. This name is inspired by its use for very similar—though not necessarily identical—ideas in the literature [40, 80, 81]. The discussion of this structure and its properties provides the content of the next subsection. It is only afterwards that the connection will be constructed starting from equation (4.3), thereby also showing its uniqueness.

4.2.4 The Hessian structure

The property defining the Hessian structure for a data set model $(\mathbb{X}, \mathbb{M}, \mu, D)$ is the relation (4.3),

$$\nabla_a \nabla_b D(x||m_\zeta) = \nabla_a \nabla_b \Phi(\zeta). \qquad \text{(4.3 revisited)}$$

However, it is a sufficient condition that the left hand side of this expression is independent of the data set x. That there then always exists a function Φ satisfying equation (4.3) is presented as a Theorem below. Note that this property implies a well-defined metric tensor on $\mathbb{M}$; it is a stronger version of Condition 4.

A most useful observation is that the Hessian of the divergence is equal to the metric tensor, which can be shown through a short computation. Indeed, choose $x \in \mu^{-1}(m_\zeta)$, then

$$\begin{aligned}\nabla_a \nabla_b D(x||m_\zeta) &= \partial_a \partial_b D(x||m_\zeta) - \omega^k{}_{ij}(\zeta) \partial_k D(x||m_\zeta)\\ &= \partial_a \partial_b D(x||m_\zeta)\\ &= g_{ab}(\zeta).\end{aligned}$$

Even though this argument only holds for data sets which have m_ζ as their projection on $\mathbb{M}$, the condition (4.3) implies that

$$\nabla_a \nabla_b D(x||m_\zeta) = g_{ab}(\zeta) \qquad \forall x \tag{4.4}$$

Another consequence of equation (4.3) is that it imposes conditions on the properties of the connection ∇. The first of these is that ∇ is torsionless or, expressed through its coefficients

$$\omega^c{}_{ab}(\zeta) = \omega^c{}_{ba}(\zeta), \tag{4.5}$$

which follows from the definition of the Hessian (2.17) and the symmetry of both the matrix of second derivatives and of the components of the metric tensor. The other two properties can be determined by computing the partial derivative of equation (4.4). Indeed, consider

$$\begin{aligned}\partial_a g_{bc} &= \partial_a(\partial_b\partial_c D(x||m_\zeta) - {\omega^d}_{bc}\partial_d D(x||m_\zeta)) \\ &= \partial_a\partial_b\partial_c D(x||m_\zeta) - (\partial_a {\omega^d}_{bc})\partial_d D(x||m_\zeta) - {\omega^d}_{cb}\partial_a\partial_d D(x||m_\zeta) \\ &= \partial_a\partial_b\partial_c D(x||m_\zeta) - (\partial_a {\omega^d}_{bc})\partial_d D(x||m_\zeta) \\ &\quad - {\omega^d}_{bc}[g_{ad} + {\omega^e}_{ad}\partial_e D(x||m_\zeta)].\end{aligned}$$

Subtracting this equation from itself with the indices a and b interchanged and rearranging the terms yields

$$\begin{aligned}0 = &[\partial_a {\omega^d}_{bc} - \partial_b {\omega^d}_{ac} + {\omega^d}_{ae}{\omega^e}_{bc} - {\omega^d}_{be}{\omega^e}_{ac}]\partial_d D(x||m_\zeta) \\ &+ [\partial_a g_{bc} - \partial_b g_{ac} + g_{ad}{\omega^d}_{bc} - g_{bd}{\omega^d}_{ac}].\end{aligned}$$

Since this must hold for all data sets x, the two expressions between square brackets must vanish independently and as such the two other conditions are obtained. The best known of these,

$$\partial_a {\omega^d}_{bc} - \partial_b {\omega^d}_{ac} + {\omega^d}_{ae}{\omega^e}_{bc} - {\omega^d}_{be}{\omega^e}_{ac} = 0, \tag{4.6}$$

has already been identified in the introduction as expressing the connection to be flat and its consequences have been discussed there. The other condition,

$$\partial_a g_{bc} - \partial_b g_{ac} + g_{ad}{\omega^d}_{bc} - g_{bd}{\omega^d}_{ac} = 0, \tag{4.7}$$

is of the same form as the Codazzi-Peterson equation, which originated as a condition on the second fundamental form of a two-dimensional surface embedded in $\mathbb{R}^3$ [28, 82, 83]. Since the generalised Fisher information metric is obtained as the second fundamental form of a surface—be it a different surface for every value of θ—this property is not wholly unexpected. There is, however, also another interpretation. Using equation (2.16) to obtain the coefficients ${\varpi^c}_{ab}$ of the connection ∇^* dual to ∇, (4.7) can be rewritten into

$$g_{cd}({\varpi^d}_{ab} - {\varpi^d}_{ba}) = 0,$$

which means the dual connection is also torsionless. The three conditions (4.5), (4.6) and (4.7) play an important role in the rest of this dissertation: they will directly or indirectly give rise to the rich properties of the Hessian structure and they will be used as a way to verify whether or not a parametrised family of distributions belongs to the exponential family. Their first use, however, is in showing the existence of a generalised Massieu function for data set models—a proof promised to the reader in the beginning of this subsection.

Theorem 2. *Given a data set model $(\mathbb{X}, \mathbb{M}, \mu, D)$, for which there exists a connection ∇ such that*

$$\nabla_a \nabla_b D(x||m_\zeta)$$

is independent of x, there exists a function $\Phi : \mathbb{R}^n \to \mathbb{R}$ such that its Hessian (with respect to ∇) equals the Hessian of the divergence function.

Proof. In the preceding discussion, the x-independence of $\nabla_a \nabla_b D(x||m_\zeta)$ was already shown to lead to the equality of this Hessian to the metric tensor. As such, it is only required to show that there exists a solution Φ to the system of differential equations

$$\partial_a \partial_b \Phi(\zeta) - \omega^c{}_{ab}(\zeta) \partial_c \Phi(\zeta) = g_{ab}(\zeta).$$

As an intermediate step, consider the differential equations

$$\partial_a \alpha_b(\zeta) - \omega^c{}_{ab}(\zeta) \alpha_c(\zeta) = g_{ab}(\zeta). \tag{4.8}$$

This is a system of the Mayer-Lie type and it is integrable if and only if equations (4.6) and (4.7) are satisfied.[1] Since these conditions were shown to be a consequence of this theorem's premise, a solution $\alpha : \mathbb{R}^n \to \mathbb{R}^n$ to the system (4.8) is thus found to exist.
The only step required to complete the proof is to show that $\alpha_b = \partial_b \Phi$ for some function Φ. Since the metric tensor is symmetric and the connection ∇ is torsionless by assumption, it follows from (4.8) that

$$\partial_a \alpha_b(\zeta) - \partial_b \alpha_a(\zeta) = 0.$$

Invoking Poincaré's lemma shows the existence of a potential for $\alpha \overset{\text{def.}}{=} \alpha_c \mathrm{d}\zeta^c$, which is exactly what is needed to complete the proof. □

As a consequence of this theorem, it immediately follows that

$$g_{ab}(\zeta) = \nabla_a \nabla_b D(x||m_\zeta) = \nabla_a \nabla_b \Phi.$$

A metric that can be written as the Hessian of a generalised Massieu function is called a Hessian metric (see also [83]). Since the metric is a positive definite tensor, this also shows that Φ is a strictly convex function when expressed in the affine coordinates of ∇. (In the remainder of this subsection, the notation θ^i will consistently be used for these affine coordinates.)

[1] The detailed proof of this statement is rather long and technical. The interested reader is therefore referred to Frankel's excellent book [28], which contains the argument in full.

Without any additional constraints, the Massieu function thus has a well-defined Legendre-Fenchel transform S, defined by

$$S(U) = \inf_{\theta}\{\Phi(\theta) + \theta^k U_k\}$$

and where the convention common in the physics literature has been used.[2] The function S can be called the generalised thermodynamic entropy in analogy with statistical mechanics.
From the equation (4.4) and Theorem 2 follows that the divergence of a data set model exhibiting a Hessian structure takes a very simple form when expressed in the affine coordinates of the connection. Indeed, in such coordinates it is possible to write

$$\partial_i \partial_j D(x||m_\theta) = g_{ij}(\theta) = \partial_i \partial_j \Phi(\theta).$$

This leads to the simple expression

$$D(x||m_\theta) = \Phi(\theta) + \theta^k q_k(x) - \sigma(x), \tag{4.9}$$

where the values of the functions q_k and σ depend on the particular choice for Φ, which is determined only up to an affine term in the θ-coordinates. However, the functions q_k must have constant values on the different fibres. This is the case since data sets x contained in the fibre of m_θ must satisfy

$$\partial_k D(x||m_\theta) = 0 \quad \text{and so} \quad \partial_k \Phi(\theta) = -q_k(x).$$

Introduce the function $u : \mathbb{R}^n \to \mathbb{R}^n$ through

$$u_k(\theta) = q_k(x) \quad \text{where} \quad \mu(x) = m_\theta.$$

Since the functions q_k are constant on the fibres, the function u is well-defined. It is also a homeomorphism as it is defined through the derivative of a strictly convex function. This function u is the key to interpreting the precise meaning of the generalised Fisher information metric. After all,

$$g_{ij}(\theta) = \partial_i \partial_j \Phi(\theta) = -\partial_i u_j(\theta). \tag{4.10}$$

Because of the definition of the function u, the Fisher information thus quantifies how sensitive the values of the functions q_k are to a change in data set, or more precisely how sensitive these values are to a change in the fibre containing the data set. Perhaps an easier way to think about this is that the

[2] Another possible convention would use $\sup_\theta\{\theta^k U_k - \Phi(\theta)\}$ to obtain a second convex function, whereas S as introduced in the main body of the text is concave.

inverse of this tensor expresses how sensitive the parameters of a data set's projection is to a small change in the observed values of the q_k. This generalises the meaning of the Fisher information metric in statistics, where it has this same interpretation for exponential families, with the expectation values of the Hamiltonians taking the role of the q_k [4]. It is important to remember that this is a purely geometric statement, concerned only with the sensitivity of the parameters to a change in the values of the functions q_k. It does not support statements about the accuracy—how close to or how far from the true value of the parameters a given choice of model point is.

An obvious question to ask in any theory pertaining to statistics is whether or not a generalisation of the Cramér-Rao bound [39] can be formulated. At this point an elegant answer to that question can be provided. It is indeed possible to produce an inequality reminiscent of this important result, although its interpretation as a bound on the variances of estimators does not generally hold. Instead, it is possible to place a bound on the derivatives of the function u, which reduces to the Cramér-Rao bound in the case of exponential families of probability distributions.

Since the metric tensor is positive definite, it must satisfy the inequality of Cauchy-Schwarz,

$$(v^i g_{ij} w^j)^2 \leqslant (v^i g_{ij} v^j)(w^i g_{ij} w^j),$$

for all vectors v and w. Using the equalities (4.10) to replace the components of the metric, it is found after some rearranging that

$$-v^i v^j \partial_i u_j \geqslant \frac{(v^i w^j \partial_i \partial_j \Phi)^2}{w^i w^j g_{ij}}. \tag{4.11}$$

The mathematical form of this inequality is very similar to a general expression of the Cramér-Rao bound for multivariate exponential families and even for deformed exponential families [4]. Indeed, in the former case it holds that

$$\begin{aligned} -\partial_i u_j(\theta) &= -\partial_i \mathbb{E}_\theta[\mathcal{H}_j] \\ &= \mathbb{E}_\theta[\mathcal{H}_i \mathcal{H}_j] - \mathbb{E}_\theta[\mathcal{H}_i]\mathbb{E}_\theta[\mathcal{H}_j], \end{aligned}$$

where $\mathbb{E}_\theta[\mathcal{H}_i]$ is the expectation value of the Hamiltonian $\mathcal{H}_i$ with respect to p_θ. Substituting this expression back into (4.11) yields the promised result.

4.2.5 Computing the connection coefficients

The construction of the connection coefficients for a data set model exhibiting a Hessian structure will proceed from equation (4.4). This formulation is

easier to use in practice than (4.3) since one has access to the metric tensor in principle as soon as the data set model has been specified, whereas the generalised Massieu function Φ is harder to find.
The sixth condition on the divergence implies the existence of data sets $x^{(c)}$ such that

$$\partial_d D(x^{(c)}||m_\zeta)$$

form the components of an invertible matrix A. By employing these data sets it is possible to isolate the connection coefficients from equation (4.4). Indeed, choose $x \in \mu^{-1}(m_\theta)$, then the Hessian structure implies

$$\begin{aligned}0 &= \nabla_a\nabla_b D(x^{(c)}||m_\zeta) - \nabla_a\nabla_b D(x||m_\zeta)\\ &= \partial_a\partial_b D(x^{(c)}||m_\zeta) - \omega^d{}_{ab}(\zeta)\partial_d D(x^{(c)}||m_\zeta) - g_{ab}(\zeta)\\ &= \partial_a\partial_b D(x^{(c)}||m_\zeta) - \omega^d{}_{ab}(\zeta)A_d{}^c(\zeta) - g_{ab}(\zeta).\end{aligned}$$

From this it follows easily that

$$\omega^c{}_{ab}(\zeta) = \sum_{d=1}^{n} (A^{-1}(\zeta))^c{}_d[\partial_a\partial_b D(x^{(d)}||m_\zeta) - g_{ab}(\zeta)].$$

In many practical applications a more careful choice of $x^{(c)}$ can be made such that A is a diagonal matrix. The expression for the connection coefficients can then be written in the simpler form—without using Einstein summation—

$$\omega^c{}_{ab}(\zeta) = \frac{\partial_a\partial_b D(x^{(c)}||m_\zeta) - g_{ab}(\zeta)}{\partial_c D(x^{(c)}||m_\zeta)}. \tag{4.12}$$

In order to be able to make a link with the definition of connections in the existing literature, however, requires use of the curves $X^{(a)}$ appearing as a consequence of the stronger form of Condition 6. Note that these curves will in general have an implicit parameter dependency—as their definition requires the choice of a model point $\mu(X^{(a)}(0))$—even though the notation does not indicate this. Starting again from equation (4.4) and deriving with respect to the parameter of the curve $X^{(a)}$ shows

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\nabla_a\nabla_b D(X^c(\varepsilon)||m_\zeta)\Big|_{\varepsilon=0}\\ &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_a\partial_b D(X^c(\varepsilon)||m_\zeta)\Big|_{\varepsilon=0} - \omega^d{}_{ab}(\zeta)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_d D(X^c(\varepsilon)||m_\zeta)\Big|_{\varepsilon=0}\\ &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_a\partial_b D(X^c(\varepsilon)||m_\zeta)\Big|_{\varepsilon=0} - \omega^c{}_{ab}(\zeta).\end{aligned} \tag{4.13}$$

This definition is more technical than it needs to be for practical purposes. In particular, it is again sufficient that the expression

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_b D(X^c(\varepsilon)||m_\theta)\bigg|_{\varepsilon=0}$$

is an invertible matrix, instead of specifically the identity, so the connection coefficients can be determined uniquely.

It is instructive to remark that the connection introduced by applying the the definitions (4.12) or (4.13) to a statistical model endowed with the relative entropy as its divergence can be identified as the exponential connection of Efron and Reeds. To be precise, it is the construction of the connection that generalises the construction of the exponential connection in information geometry. Using the same definition for different divergence functions than yields different connections. For example, using the Kullback-Leibler divergence with its arguments interchanged would yield the mixture connection [29,30].

The absence of a derivative in expression (4.12) compared to expression (4.13) may give the impression that the discrete method is easier and less likely to cause mathematical problems. This need not be the case, however. For a general data set model, there is no guarantee that the connection coefficients defined through the expression (4.12) are independent of the choice of the data sets $x^{(c)}$. This will only be the case when the data set model exhibits a Hessian structure.

4.3 A generalised Pythagorean theorem

For some data set models, the analysis of their geometric properties can be simplified considerably if a specific property is satisfied. In particular, this property allows one to reproduce much of the geometry of proper divergence functions as they are discussed in Chapter 3.

A sufficient demand for a model to satisfy Condition 4 as well as for a connection to exist is that for all $(x, m) \in \mathrm{Dom}(D)$ it holds that the difference

$$D(x||m) - D(x||\mu(x)) \tag{4.14}$$

does not depend on the chosen data set x, only on its projection $\mu(x)$. If for a data set model $(\mathbb{X}, \mathbb{M}, \mu, D)$ the function

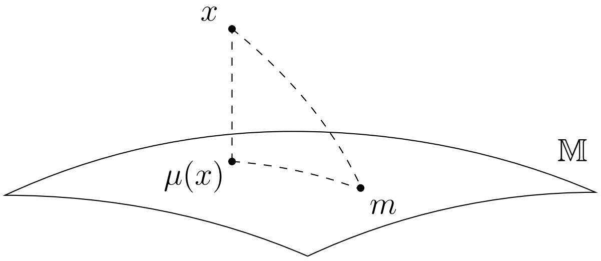
$$m \mapsto D(x||m) \tag{4.15}$$

has its unique global minimum in $\mu(x)$ for all $x \in \mathrm{Dom}(\mu)$, this property determines a proper divergence function on the manifold $\mathbb{M}$. Abusing notation, it is then possible to define

$$D(\mu(x)||m) \overset{\text{def.}}{=} D(x||m) - D(x||\mu(x)). \tag{4.16}$$

It is quickly verified that this construction does indeed give rise to a proper divergence function. Since $\mu(x)$ is assumed to be the unique point where the function (4.15) reaches its global minimum, it holds that $D(x||m) \geqslant D(x||\mu(x))$ and as a consequence $D(\mu(x)||m) \geqslant 0$ with equality only if $\mu(x) = m$. When $\mu(x) = m$, the difference between $D(x||m)$ and $D(x||\mu(x))$ vanishes trivially and so the newly defined divergence equals zero everywhere on the diagonal of $\mathbb{M} \times \mathbb{M}$.

If the left hand side of equation (4.16) is well defined on the domain of D, this equality will be referred to as the generalised Pythagorean theorem. An illustration is given in the drawing, where the divergence between points is represented by the squared length of the dashed line between them.



In some examples, such as in information geometry, it is the case that $\mathbb{M} \subset \mathbb{X}$, making it possible that the divergence function is actually defined on $\mathbb{X} \times \mathbb{X}$ as a whole. Such divergence functions may satisfy a stronger version of the generalised Pythagorean theorem. An example of this is the Kullback-Leibler divergence, for which it holds that

$$D(p||r) = D(p||q) + D(q||r)$$

when q is the projection of p on a particular choice of submanifold containing the distribution r [30]. However, some of these divergences will fail to satisfy the generalised Pythagorean theorem on $\mathbb{X} \times \mathbb{M}$ yet condition (4.14) may still hold for all data sets and model points. The latter condition is thus weaker than the former. In that case it is important to formulate the Pythagorean theorem through expression (4.16) rather than using the original expression for the divergence in all three terms.

For data set models satisfying the property (4.16), it automatically also holds that the metric tensor can have its components written as

$$\begin{aligned} g_{ij}(\theta) &= \partial_i\partial_j D(x||m_\theta) \\ &= \partial_i\partial_j \left[D(x||\mu(x)) + D(\mu(x)||m_\theta)\right] \\ &= \partial_i\partial_j D(\mu(x)||m_\theta). \end{aligned}$$

Indeed, when x is an element of the fibre of m_θ, both arguments of the expression $\partial_i\partial_j D(\mu(x)||m_\theta)$ coincide. This in turn is the definition (3.9) for the metric tensor derived from proper divergence functions presented earlier. Also the definition of the connection coefficients (4.13) can be expressed in terms of the proper divergence. Following an analogous line of reasoning as for the metric tensor, one obtains

$$\begin{aligned} \omega^k{}_{ij}(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_i\partial_j D(X^k(\varepsilon)||m_\theta)\bigg|_{\varepsilon=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_i\partial_j D(\mu(X^k(\varepsilon))||m_\theta)\bigg|_{\varepsilon=0}. \end{aligned}$$

To show the correspondence with the definition (3.10) for the connection coefficients derived from proper divergence functions, fix θ and define $\Theta^{(k)}$ by

$$\mu(X^k(\varepsilon)) = m_{\Theta^{(k)}(\varepsilon)}, \quad X^k(0) = m_\theta.$$

Then it is possible to write

$$\begin{aligned} \delta^l_k &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_k D(m_{\Theta^{(l)}(\varepsilon)}||m_\theta)\bigg|_{\varepsilon=0} \\ &= \frac{\partial}{\partial\xi^s}\partial_k D(m_\xi||m_\theta)\big|_{\xi=\theta}\frac{\mathrm{d}\Theta^{(l)s}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \\ &= -g_{sk}(\theta)\frac{\mathrm{d}\Theta^{(l)s}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0}. \end{aligned}$$

This knowledge can be used to compute the connection coefficients in an analogous way as above:

$$\begin{aligned} \omega^k{}_{ij}(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\partial_i\partial_j D(X^k(\varepsilon)||m_\theta)\bigg|_{\varepsilon=0} \\ &= \left(\frac{\partial}{\partial\xi^s}\partial_i\partial_j D(m_\xi||m_\theta)\big|_{\xi=\theta}\right)\frac{\mathrm{d}\Theta^{(k)s}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \\ &= g^{ks}(\theta)\left[-\frac{\partial}{\partial\xi^s}\partial_i\partial_j D(m_\xi||m_\theta)\bigg|_{\xi=\theta}\right]. \end{aligned}$$

This is indeed the definition found in the literature and a reported upon in the previous chapter, in particular in equation (3.10).

4.4 Identifying exponential families

This section contains two easily obtained results concerning exponential families of probability distributions. In particular, it is shown that a statistical model belongs to the exponential family if and only if it exhibits a Hessian structure. This is useful since this allows for the establishment of both positive and negative results through a straightforward step-by-step process. For exponential families, it is shown that the canonical parameters can be found by solving a system of linear differential equations. I do not claim originality of these results, however. Especially the second one I expect to be contained in at least some books on differential geometry as a method to find the affine coordinates of a connection. The first result on the other hand, genuinely appears to be missing form the important reference works—both on exponential families and on Hessian structures—as well as from the available historically important papers cited in the introduction. Since both properties will be used extensively in the treatment of the examples in the next chapter, it is thus instructive to (re-)derive them here.

It was shown earlier in this chapter that exponential families exhibit a Hessian structure. The converse is also true: the presence of a Hessian structure in a statistical model only occurs for an exponential family—assuming that the divergence is chosen to be that of Kullback and Leibler. To show this, assume a statistical model with a Hessian structure derived from the relative entropy. Since the connection is necessarily flat and torsionless, it allows for affine coordinates θ. In these coordinates the Hessian of the divergence reads

$$\partial_i\partial_j D(p||p_\theta) = -\partial_i\partial_j \int_X p(x)\ln p_\theta(x)\mathrm{d}x$$

and by assumption both sides are of this equality are independent of the distribution p. But this is only possible when

$$x \mapsto \partial_i\partial_j \ln p_\theta(x)$$

is a function of θ only. With a bit of foresight, this function can be identified as $-\Phi$. Then it follows that

$$\ln p_\theta(x) = -\Phi(\theta) - \theta^k \mathcal{H}_k(x)$$

for properly chosen functions $\mathcal{H}_k$. (A term independent of θ could be absorbed in the definition of the measure $\mathrm{d}x$ and can thus be discarded.) This shows the result that was promised to the reader. Due to its practical importance, it is summarised in its own theorem:

Theorem 3. *A statistical model belongs to the exponential family if and only if it exhibits a Hessian structure when it is endowed with the Kullback-Leibler divergence (3.4).*

A recurring property in the work of Efron, Reeds and Amari is that exponential families of probability distributions are flat when endowed with the exponential connection [29,47,49]. Since the connection of the data set model geometry generalises this exponental connection, Theorem 3 can be seen as the data set model generalisation of that information geometric property and its converse.

A useful corollary pertains to the canonical parameters of the exponential families. When employed as coordinates for the model manifold, they serve as the affine coordinates for the flat and torsionless (exponential) connection. This means that these parameters θ^i can be found as a function of arbitrary coordinates ζ^a by studying the connection. More specifically, the connection coefficients are in general defined through (2.8) and this relation can be transformed from general coordinates ζ^a to the affine coordinates $\theta^i = \Theta^i(\zeta)$. This looks like

$$\begin{aligned}
{\omega^c}_{ab}\partial_c &= \nabla_a \partial_b \\
&= \nabla_a\big((\partial_b\Theta^j)\partial_j\big) \\
&= (\partial_a\partial_b\Theta^j)\partial_j + (\partial_b\Theta^j)(\partial_a\Theta^j)(\nabla_i\partial_j) \\
&= (\partial_a\partial_b\Theta^j)\partial_j + (\partial_b\Theta^j)(\partial_a\Theta^j){\omega^k}_{ij}\partial_k,
\end{aligned}$$

where letters from the beginning of the alphabet refer to the arbitrary ζ-coordinates and letters from the middle of the alphabet belong to the affine coordinates θ. The coefficients ${\omega^k}_{ij}$ thus vanish identically by assumption. Computing the θ^j-component of the left hand side and using $\mathrm{d}\theta^j(\partial_c) = \partial_c\Theta^j$ reveals

$$\partial_a\partial_j\Theta^j(\zeta) = {\omega^c}_{ab}(\zeta)\partial_c\Theta^j(\zeta). \tag{4.17}$$

This is a system of linear differential equations which may be solved to find the function $\zeta \mapsto \Theta(\zeta)$ expressing the affine coordinates as a function of arbitrary coordinates.

4.5 Discussion

It is important as well as instructive to compare the data set model formalism with the theory of information geometry upon which it is based.

First amongst these differences is of course the fact that probability theory has been deliberately excluded from the building blocks of the formalism. As is explained in the introduction, this was done in order to identify the crucial link between information theory and the geometry of the formalism, rather than relying on the much-walked path of basing information theory on probability theory. The possibility of using properties ultimately based on probability—on purpose or by accident—is always present when studying information geometry. After all, despite its geometric nature, it still relies in part on properties of the objects it deals with to show theorems and additional properties, as a read-through of the literature referred to in the previous chapter will indicate. Naturally, this cannot be taken as a criticism of information geometry in itself, which has many useful applications in a statistical context which rely on exactly those properties.

A striking contrast with existing methods of information geometry, and perhaps the most promising aspect of the data set model formalism in terms of applications, is the explicit option to model data through qualitatively different mathematical objects. In information geometry, as well as in the study of proper divergence functions, the divergence functions take both arguments from the same set or from a set and a subset thereof, the latter of which then serves as the model for the former. By taking into account the possibility that both arguments of the divergence are qualitatively different a much wider array of possible models can be described and can have their divergence-implied geometry constructed and studied. This general formalism strongly resembles that of pattern recognition in machine learning, see for instance [84] for a good comparison between pattern recognition and information geometry. This generality does, however, come at the price of losing some structure which does exist in other formalisms and contexts. The family of affine connections introduced by Chentsov and Amari is an example of such a structure that could not be replicated.

People familiar with information geometry may perceive the emphasis of the data set model formalism on intrinsic geometry rather than the extrinsic geometry as an important contrast. Nevertheless, this offers a great advantage in terms of generality as not every model manifold has an obvious embedding in some larger set. This embedding is required to enable the study of an extrinsic geometry. Removing the need for an embedding also opens up the possibility of considering the entire simplex of everywhere strictly positive probability distributions over a measurable set as the model manifold. Doing this is rather unnatural in the usual formalism of information geometry since there this simplex serves as the space in which the model is embedded and from which it derives its geometry.

There are also some strong similarities between the data set model formalism and information geometry. The first one is obviously that it is possible to reconstruct the metric tensor and the exponential connection from the latter theory in the former, at least under certain conditions. The reconstruction of a family of connections has not been achieved, however. The metric tensor is used in quantifying information in a way which is very analogous to its use in information geometry: both are a measure for the sensitivity of the modelling process to changes in the data—or more precisely: changes in the fibre containing the data.

It also proved possible to reconstruct a number of useful properties. The most obvious of these is probably the vanishing curvature of statistical manifolds corresponding to exponential families. This result is well-known from information geometry, as it was one of the starting points for the study of connections in this field. The fact that exponential families can be identified by means of their flatness and the correspondence of canonical parameters and affine coordinates was already implied in the work of Amari, who demonstrated this fact through direct computation [29, 30]. However, the above work shows that a similar result can be obtained in a more general context by identifying all data set models which give rise to a Hessian metric as exponential families.

The generalised Pythagorean theorem is also interesting when comparing the data set model formalism the study of proper divergence functions. It makes clear that the data set model geometry is a true extension of the differential geometry induced by proper divergence functions on a manifold.

5. EXAMPLES AND APPLICATIONS

This last chapter is devoted to the illustration of the data set formalism. A few familiar examples from statistics and statistical physics will be presented to show how the formalism developed in the previous chapter behaves when applied to these models. There are a total of six examples, each of which demonstrates one or more important properties of the data set model formalism.

5.1 *The normal distributions*

5.1.1 *Using the relative entropy*

In this first example, the Gaussian probability densities (or normal distributions) will be covered as a model for probability densities over $\mathbb{R}$. This family is also used as an example in Chapter 3. The results found there are mostly rederived, now along the lines of the data set model formalism as set out in the previous chapter. It is also illustrated how the canonical coordinates of the exponential can be found in a straightforward manner by constructing the affine connection.

The data sets represent an empirical distribution over the real numbers with finite first and second moments. In information geometry this set would be treated as an infinitely dimensional manifold of which the normal distributions are a submanifold. The data set model formalism, however, does not employ this additional structure. The task at hand in this example is then to find the best fitting member of the two-parameter family of normal distributions, given by the expression

$$p_{\mu,\sigma}(x) = \exp\left\{ -\ln\sqrt{2\pi\sigma^2} - \frac{(x-\mu)^2}{2\sigma^2} \right\}.$$

These distributions form a two-dimensional manifold for which μ and $\sigma > 0$ are used as parameters. (The model map will not be named explicitly in this example and so the symbol μ is free to be used for the mean of the normal distribution, as is conventional.) Some sources use σ^2 as the second parameter, leading to some minor differences between the results obtained

there and the ones derived here.
The best fit will be sought by minimising the Kullback-Leibler divergence

$$\begin{aligned} D(p||p_{\mu,\sigma}) &= \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{p_\theta(x)} \mathrm{d}x \\ &= -S(p) + \ln \sqrt{2\pi\sigma^2} + \frac{1}{2\sigma^2}\mathbb{E}_p[(x-\mu)^2], \end{aligned} \tag{5.1}$$

where S represents the Shannon entropy (3.8) of the probability distribution p. The model map here is entirely implied by the divergence: there is only a single normal distribution minimising this divergence for any of the empirical distributions serving as the first argument and this global minimum is also a local minimum. The derivatives themselves satisfy

$$\begin{aligned} \partial_\mu D(p||p_{\mu,\sigma}) &= -\frac{1}{\sigma^2}\mathbb{E}_p[x-\mu], \\ \partial_\sigma D(p||p_{\mu,\sigma}) &= \frac{1}{\sigma} - \frac{1}{\sigma^3}\mathbb{E}_p[(x-\mu)^2]. \end{aligned}$$

The condition that these expressions vanish simultaneously yields the well known-expressions

$$\mathbb{E}_p[x] = \mu, \tag{5.2}$$

$$\mathbb{E}_p[(x-\mu)^2] = \sigma^2. \tag{5.3}$$

In order to determine the Fisher information metric, it is required to know the matrix of second derivatives of the relative entropy. These satisfy the expressions

$$\begin{aligned} \partial_\mu^2 D(p||p_{\mu,\sigma}) &= \frac{1}{\sigma^2}, \\ \partial_\sigma \partial_\mu D(p||p_{\mu,\sigma}) &= \frac{2}{\sigma^3}\mathbb{E}_p[x-\mu], \\ \partial_\sigma^2 D(p||p_{\mu,\sigma}) &= -\frac{1}{\sigma^2} + \frac{3}{\sigma^4}\mathbb{E}_p[(x-\mu)^2]. \end{aligned}$$

Both expectation values appearing in this matrix can be rewritten as functions of the parameters only. This is achieved by using the conditions (5.2) and (5.3). As such it is implied that the metric is well-defined and has components

$$g_{\mu\mu}(\mu,\sigma) = \frac{1}{\sigma^2}, \quad g_{\mu\sigma}(\mu,\sigma) = 0, \quad g_{\sigma\sigma}(\mu,\sigma) = \frac{2}{\sigma^2}.$$

The next step in constructing the data set model geometry is to determine the connection coefficients. This requires the identification of distributions for which all but one of the derivatives of the divergence vanish. Let us first turn our attention to finding $p^{(\mu)}$, which is used to find the connection coefficients ${\omega^\mu}_{ij}$. Such a distribution must satisfy

$$\partial_\mu D(p^{(\mu)}||p_{\mu,\sigma}) \neq 0 \quad \text{and} \quad \partial_\sigma D(p^{(\mu)}||p_{\mu,\sigma}) = 0.$$

It is thus necessary to find a distribution with a mean different from μ but with its expectation value of $(x-\mu)^2$ equal to σ^2. Take the mean of $p^{(\mu)}$ to be equal to $\mu + \delta\mu$. This can be used to compute the connection coefficients ${\omega^\mu}_{ij}$ through equation (4.12). The result is

$$\begin{aligned} {\omega^\mu}_{\mu\mu}(\mu,\sigma) &= 0, \\ {\omega^\mu}_{\mu\sigma}(\mu,\sigma) &= \frac{2\sigma^{-3}(\delta\mu) - 0}{-\sigma^{-2}(\delta\mu)} = -\frac{2}{\sigma}, \\ {\omega^\mu}_{\sigma\sigma}(\mu,\sigma) &= 0. \end{aligned}$$

The other three independent connection coefficients ${\omega^\sigma}_{ij}$ can be computed in an analogous way. This requires the choice of a data set $p^{(\sigma)}$, which must have μ for its mean but $(\sigma+\delta\sigma)^2$ as its second central moment. Substituting this distribution again into equation (4.12) reveals the expressions for the last three independent connection coefficients to be

$$\begin{aligned} {\omega^\sigma}_{\mu\mu}(\mu,\sigma) &= 0, \\ {\omega^\sigma}_{\mu\sigma}(\mu,\sigma) &= 0, \\ {\omega^\sigma}_{\sigma\sigma}(\mu,\sigma) &= \frac{-\sigma^{-2} + 3\sigma^{-4}(\sigma+\delta\sigma)^2 - 2\sigma^{-2}}{\sigma^{-1} - \sigma^{-3}(\sigma+\delta\sigma)^2} = -\frac{3}{\sigma}. \end{aligned}$$

It can be verified quickly that these coefficients are indeed the correct ones in order to write the metric tensor as the Hessian of the divergence function, regardless of the chosen empirical distribution p. Take, as an example, the component $g_{\sigma\sigma}$. A straightforward computation yields

$$\begin{aligned} &\nabla_\sigma \nabla_\sigma D(p||p_{\mu,\sigma}) \\ &= \partial_\sigma^2 D(p||p_{\mu,\sigma}) - {\omega^\sigma}_{\sigma\sigma}(\mu,\sigma)\partial_\sigma D(p||p_{\mu,\sigma}) \\ &= -\sigma^{-2} + 3\sigma^{-4}\mathbb{E}_p[(x-\mu)^2] - \left(-3\sigma^{-1}\right)\left(\sigma^{-1} - \sigma^{-3}\mathbb{E}_p[(x-\mu)^2]\right) \\ &= \frac{2}{\sigma^2} \\ &= g_{\sigma\sigma}(\mu,\sigma) \end{aligned}$$

as desired. This data set model can thus be concluded to feature a Hessian structure.

It is shown in the text of the previous chapter that this Hessian structure implies the connection to be flat and torsionless, as well to satisfy the Codazzi-Peterson-like equation (4.7). This can be verified explicitly, but it is more instructive to find the affine coordinates of the connection. The function $(\mu, \sigma) \mapsto \Theta^i(\mu, \sigma)$ expressing these affine coordinates must satisfy three partial differential equations, given by equation (4.17). In this example, these equations look like

$$\begin{aligned} \partial_\mu^2 \Theta^i(\mu, \sigma) &= 0, \\ \partial_\mu \partial_\sigma \Theta^i(\mu, \sigma) &= -2\sigma^{-1} \partial_\mu \Theta^i(\mu, \sigma), \\ \partial_\sigma^2 \Theta^i(\mu, \sigma) &= -3\sigma^{-1} \partial_\sigma \Theta^i(\mu, \sigma). \end{aligned}$$

From the first of these equations, it follows that

$$\Theta^i(\mu, \sigma) = A^i(\sigma)\mu + B^i(\sigma),$$

where the functions A^i and B^i still need determining. Substitution into the third partial differential equations yields

$$\mu \partial_\sigma^2 A^i + \partial_\sigma^2 B^i = -3\mu\sigma^{-1} \partial_\sigma A^i - 3\sigma^{-1} \partial_\sigma B^i.$$

Since this must hold for all values of μ, one obtains two independent equations for A^i and B^i. The solutions of these demand that both A^i and B^i are proportional to σ^{-2}. The second of the above partial differential equations is then also satisfied. It is now possible to choose the constants A^i and B^i in such a way as to obtain the well-known canonical parameters

$$\theta^1 = \frac{1}{2\sigma^2} \quad \text{and} \quad \theta^2 = -\frac{\mu}{\sigma^2}.$$

5.1.2 Using a different divergence function

The simplicity of this example offers a good opportunity to study the effects of using a different divergence function—in this case a divergence function different from the relative entropy (5.1). The set $\mathbb{X}$ of data, the model manifold $\mathbb{M}$ and the model map μ will thus remain unchanged. An obvious choice for the new divergence D' is the expression

$$D'(p||p_{\mu,\sigma}) = \frac{1}{2\mu_0^2}\left(\mu - \mathbb{E}_p[x]\right)^2 + \frac{1}{4\sigma_0^4}\left(\mu^2 + \sigma^2 - \mathbb{E}_p[x^2]\right)^2, \qquad (5.4)$$

where μ_0 and σ_0 are strictly positive constants which may be arbitrary otherwise. These constants are introduced for dimensional reasons but they can also be used to change the relative sensitivity of the model map with respect to the values of $\mathbb{E}_p[x]$ and $\mathbb{E}_p[x^2]$. The derivatives of this divergence satisfy

$$\begin{aligned}\partial_\mu D'(p||p_{\mu,\sigma}) &= \frac{1}{\mu_0^2}(\mu - \mathbb{E}_p[x]) + \frac{\mu}{\sigma_0^4}(\mu^2 + \sigma^2 - \mathbb{E}_p[x^2]),\\ \partial_\sigma D'(p||p_{\mu,\sigma}) &= \frac{\sigma}{\sigma_0^4}(\mu^2 + \sigma^2 - \mathbb{E}_p[x^2]).\end{aligned}$$

For these two expressions to vanish simultaneously, it is required that

$$\begin{aligned}\mathbb{E}_p[x] &= \mu,\\ \mathbb{E}_p[x^2] &= \sigma^2 + \mu^2.\end{aligned}$$

By a straightforward computation it can be seen that these two conditions are equivalent to (5.2) and (5.3). This means the model maps implied by the divergence functions (5.1) and (5.4) coincide as intended.

To compute the metric tensor g', the matrix of second derivatives of the divergence is again needed. Another straightforward computation yields

$$\begin{aligned}\partial_\mu^2 D'(p||p_{\mu,\sigma}) &= \frac{1}{\mu_0^2} + \frac{3\mu^2 + \sigma^2 - \mathbb{E}_p[x^2]}{\sigma_0^4},\\ \partial_\mu\partial_\sigma D'(p||p_{\mu,\sigma}) &= \frac{2\mu\sigma}{\sigma_0^4},\\ \partial_\sigma^2 D'(p||p_{\mu,\sigma}) &= \frac{\mu^2 + 3\sigma^2 - \mathbb{E}_p[x^2]}{\sigma_0^4}.\end{aligned}$$

In order to determine the metric tensor, the distribution p needs to be chosen in the fibre of $p_{\beta,\mu}$. This means $\mathbb{E}_p[x^2] = \mu^2 + \sigma^2$ and so the metric tensor has components

$$g'_{\mu\mu}(\mu,\sigma) = \frac{2\mu^2}{\sigma_0^4} + \frac{1}{\mu_0^2}, \quad g'_{\mu\sigma}(\mu,\sigma) = \frac{2\mu\sigma}{\sigma_0^4}, \quad g'_{\sigma\sigma}(\mu,\sigma) = \frac{2\sigma^2}{\sigma_0^4}.$$

Remark that the sum of squares divergence (5.4) yields a more complicated metric tensor than the apparently more complex Kullback-Leibler divergence. It is also easy to verify that the connection constructed from the divergence D' is not the metric connection, since the latter does not vanish—a fact that can be checked equation (2.15). These complications are all in some way a consequence of the presence of the parameter μ in the second term—a presence which is in itself necessitated by the expression for the second (non-central) moment of a normal distribution, which refers to both the mean μ and the variance σ^2.

Since the matrix of second derivatives does not coincide with the metric tensor, it is necessary to compute the expressions (4.12) to find the connection coefficients ${\varpi^c}_{ab}$ of ∇'. Since only two of the second derivatives of the divergence D' depend on the arbitrary distribution and then only on the second moment of this distribution, only ${\varpi^\sigma}_{\mu\mu}$ and ${\varpi^\sigma}_{\sigma\sigma}$ may be different from zero. The probability distribution $p^{(\sigma)}$ required to compute these coefficients must have a second moment different different from $\sigma^2 + \mu^2$—say $(\sigma + \delta\sigma)^2 + \mu^2$. Using again the definition (4.12) yields for the first coefficient

$$\begin{aligned}{\varpi^\sigma}_{\mu\mu}(\mu, \sigma) &= \frac{\partial_\mu^2 D'(p^{(\sigma)}||p_{\mu,\sigma}) - g_{\mu\mu}(\mu, \sigma)}{\partial_\sigma D'(p^{(\sigma)}||p_{\mu,\sigma})} \\ &= \frac{3\mu^2 + \sigma^2 - [(\sigma + \delta\sigma)^2 + \mu^2] - 2\mu^2}{\sigma(\mu^2 + \sigma^2 - [(\sigma + \delta\sigma)^2 + \mu^2])} \\ &= \frac{1}{\sigma}.\end{aligned}$$

The second one is computed in a very analogous way to obtain

$$\begin{aligned}{\varpi^\sigma}_{\sigma\sigma}(\mu, \sigma) &= \frac{\partial_\sigma^2 D'(p^{(\sigma)}||p_{\mu,\sigma}) - g_{\sigma\sigma}(\mu, \sigma)}{\partial_\sigma D'(p^{(\sigma)}||p_{\mu,\sigma})} \\ &= \frac{\mu^2 + 3\sigma^2 - [(\sigma + \delta\sigma)^2 + \mu^2] - 2\sigma^2}{\sigma(\mu^2 + \sigma^2 - [(\sigma + \delta\sigma)^2 + \mu^2])} \\ &= \frac{1}{\sigma}.\end{aligned}$$

As such, the connection is well-defined but it is a different connection than the one obtained when the relative entropy is used. Computing the curvature tensor explicitly allows one to verify that this connection is flat. However, finding the affine coordinates of this connection will have this as a corollary. To find these special coordinates, the differential equations (4.17) can be used and for this particular example they take the form

$$\begin{aligned}\partial_\mu^2\Theta^i(\mu,\sigma) &= \sigma^{-1}\partial_\sigma\Theta^i(\mu,\sigma),\\ \partial_\mu\partial_\sigma\Theta^i(\mu,\sigma) &= 0,\\ \partial_\sigma^2\Theta^i(\mu,\sigma) &= \sigma^{-1}\partial_\sigma\Theta^i(\mu,\sigma).\end{aligned}$$

The second of these equations demands that

$$\Theta^i(\mu,\sigma) = A^i(\mu) + B^i(\sigma).$$

Inserting this in the third equation yields a differential equation for B^i,

$$\sigma\partial_\sigma^2 B_i(\sigma) = \partial_\sigma B^i(\sigma)$$

and thus $B^i(\sigma) = B_0^i\sigma^2$. The first equation is then satisfied only if

$$\partial_\mu^2 A^i(\mu) = 2B_0^i.$$

A particular solution $A^i = A_0^i\mu$ is obtained when $B^i = 0$, whereas the general solution is $A^i = B_0^i\mu^2$. As such two canonical parameters are revealed to be

$$\eta^1 = \mu \quad \text{and} \quad \eta^2 = \mu^2 + \sigma^2.$$

These are the so-called “expectation parameters” of the family of normal distributions [29].

Note that this treatment never used the explicit expression for the model distributions. A similar “sum of squares” divergence function may thus be defined in a very broad range of cases. The approach does have the disadvantage of lacking interpretation, however. To take a concrete example, the family of Gumbel distributions—discussed in details in the last example—could be adopted as a model when endowed with this same divergence (5.4) as they too can be distinguished uniquely by their first and second moments. In fact this last example shows, *mutatis mutandis*, that this form of the divergence will always yield a flat and torsionless connection—at least when the squares in (5.4) contain sufficiently well-behaved expressions.

5.2 Linear regression

Perhaps the simplest and best known-way of fitting a functional relation to data points is linear regression. The input data in this example takes the form of a set of couples (x_j, y_j) which are believed to exhibit a functional relationship in principle but which have been contaminated by some form of noise. It should be noted that nothing in this example is new—this is merely an illustration that a wide array of problems fits into the data set model formalism.

The set $\mathbb{X}$ contains all collections S consisting of N_S couples $(x_j, y_j) \in \mathbb{R}^2$. Remark that these data sets S must are not required to contain the same number of couples, however they must satisfy

$$N_S \sum_j x_j^2 - \left(\sum_j x_j \right)^2 \neq 0.$$

The model points are the first order polynomials of the form

$$f_{a,b}(x) = ax + b, \quad a, b \in \mathbb{R}.$$

The numbers a and b, indicating the slope and intercept of the polynomial function, are also used as the coordinates for the model manifold $\mathbb{M}$. The obvious choice of divergence, which must indicate the best fitting model, is to adopt the quantity which must be minimised in the least squares method,

$$D(S||f_{a,b}) = \frac{1}{2} \sum_j \left(y_j - a x_j - b\right)^2. \tag{5.5}$$

Minimising this function may happen by setting equal to zero the derivatives of this divergence function, which yields

$$\begin{aligned} \partial_a D(S||f_{a,b}) &= -\sum_j (y_j - a x_j - b) x_j, \\ \partial_b D(S||f_{a,b}) &= -\sum_j (y_j - a x_j - b). \end{aligned}$$

Through a straightforward computation, it is found that these derivatives vanish simultaneously when

$$a = \frac{N_S \sum_j y_j x_j - \sum_j x_j \sum_i y_i}{N_S \sum_j x_j^2 - \left(\sum_j x_j \right)^2} \quad \text{and} \quad b = \frac{1}{N_S} \sum_j (y_j - a x_j).$$

These are of course nothing but the regular expressions for the slope and intersect of the best fit to the data S. Thus, the data set model formalism contains this aspect of the least squares linear regression method.
The next interesting quantity is the metric tensor. To compute this quantity, knowledge of the matrix of second derivatives of the divergence is needed. The independent components are given by

$$\begin{aligned}\partial_a^2 D(S||f_{a,b}) &= \sum_j x_j^2,\\ \partial_a\partial_b D(S||f_{a,b}) &= \sum_j x_j,\\ \partial_b^2 D(S||f_{a,b}) &= N_S.\end{aligned}$$

This matrix depends only on the data and not on the parameters a or b. This is not in itself a problem, as it might in principle be possible to express these quantities only as a function of those parameters by restring these expressions to data sets for which the best fit is a given line. However, it is clear that this will not be possible: the best fit is determined also by the y-values of the couples in the data set and there is no mention of those numbers in the matrix above. As a consequence, the subsets of $\mathbb{X}$ on which the above expressions are constant are not related to the fibres of the model. This shows that the divergence (5.5) does not satisfy the necessary conditions for a divergence in the data set model formalism.
Another divergence which yields the same model map was already mentioned in an earlier publication [74]. It is given by

$$\begin{aligned}D_\lambda(S||f_{a,b}) = &\frac{1}{2\sum_{j,k}(x_i - x_k)^2}\lambda^2 \sum_{j,k}[y_j - y_k - a(x_j - x_k)]^2\\ &+ \frac{1}{2\sum_{j,k}(x_i - x_k)^2}\sum_{j,k}[x_j y_k - x_k y_j - b(x_j - x_k)]^2,\end{aligned}$$

where $\lambda > 0$ is an arbitrary constant introduced for dimensional reasons. The matrix of second derivatives of this divergence is particularly simple, with constant components wholly independent of the data. This means this matrix necessarily coincides with the metric tensor, which is characterised by

$$g_{aa}(a,b) = \lambda^2, \quad g_{ab}(a,b) = 0, \quad g_{bb}(a,b) = 1.$$

Since both objects coincide, all connection coefficients will vanish and the coordinates already in use are affine coordinates for this connection. The model manifold is therefore concluded to be endowed with a flat and torsionless connection.

5.3 The grand canonical ensemble for identical particles

In this first non-trivial example, the goal is to model a system of non-interacting bosonic particles which may occupy states $\{j\}$ with corresponding energy levels $\{\varepsilon_j\}$. This is a well-known and extensively studied model. Hence, no surprising results are to be expected. Nevertheless, such a familiar example may be interesting for the reader as it applies the developed formalism to a system he or she may be already be familiar with.

The set $\mathbb{X}$ of all data sets contains possible outcomes of an experiment to measure the number n_j of bosons occupying the state j. These will be denoted by $(n_i)_i$ or just n where the context makes confusion unlikely to occur. (The letter n is not used for the dimension of $\mathbb{M}$ in this example. A similar remark holds for μ introduced soon.) It will follow implicitly from the discussion that not all possible configurations n are in the domain of the model map. Such configurations are excluded from the onset. The model points making up the manifold $\mathbb{M}$ are the distributions of the grand canonical ensemble. That is, they are the probabilities of the states j having occupations n_j and they are given by

$$\begin{aligned} p_{\beta,\mu}(\{n_i\}) &= \frac{1}{\mathcal{Z}(\beta,\mu)} \prod_j \exp\{-\beta n_j(\varepsilon_j - \mu)\} \\ &= \exp\left\{ -\ln \mathcal{Z}(\beta,\mu) - \beta \sum_j n_j \varepsilon_j + \beta\mu \sum_j n_j \right\}, \end{aligned} \tag{5.6}$$

where $\mathcal{Z}$ is the partition function of the system—also serving as a normalisation factor—, β represents the inverse temperature and μ the chemical potential. It is a well-known result of statistical physics, see for example [4], that

$$\begin{aligned} \mathcal{Z}(\beta,\mu) &= \left(\prod_j \sum_{n_j=0}^{\infty} \right) \exp\left\{ -\beta \sum_i n_i(\varepsilon_i - \mu) \right\} \\ &= \prod_j \frac{1}{1 - \exp\{-\beta(\varepsilon_j - \mu)\}}, \end{aligned}$$

where the expression between large parenthesis is an abuse of notation to indicate a multitude of sums. It is assumed that the remaining product converges. This is not a trivial assumption but it is not native to the data set model formalism and so no excessive attention will be given to this problem. From expression (5.6) it is clear that the model distributions belong to the exponential family with parameters $\theta^1 = \beta$ and $\theta^2 = -\beta\mu$. However, it is more instructive to use β and μ as parameters or coordinates of the model manifold as this will illustrate that the geometric properties of the model are indeed independent of the choice of parameters.

The geometry of a data set model is completely determined by the chosen divergence. The choice made here is given by the expression

$$D(n||p_{\beta,\mu}) = \ln \mathcal{Z}(\beta,\mu) - \sum_i n_i \left(-\beta\varepsilon_i + \beta\mu\right). \tag{5.7}$$

For a realistic experimental outcome there is a highest occupied energy level and a finite number of particles, thereby avoiding additional mathematical difficulties with this definition.

It first needs to be verified whether or not equation (5.7) does indeed define a divergence. The expression is continuous and continuously differentiable as a function of the parameters when $\beta > 0$ and $\mu \notin \{\varepsilon_j\}$. Since from a physical point of view this model only makes sense when $\beta > 0$ and $\mu < \min_j \varepsilon_j$, no difficulties are expected to be encountered in a practical setting.

The derivative of the divergence (5.7) with respect to the inverse temperature β is given by the expression

$$\begin{aligned}\partial_\beta D(n||p_{\beta,\mu}) &= \partial_\beta \ln \mathcal{Z}(\beta,\mu) + \sum_i n_i\varepsilon_i - \mu\sum_i n_i \\ &= -\sum_j \frac{\varepsilon_j - \mu}{\exp\{\beta(\varepsilon_j-\mu)\}-1} + \sum_i n_i\varepsilon_i - \mu\sum_i n_i.\end{aligned} \tag{5.8}$$

Through an analogous computation, it is found that

$$\begin{aligned}\partial_\mu D(n||p_{\beta,\mu}) &= \partial_\mu \ln \mathcal{Z}(\beta,\mu) - \beta\sum_i n_i \\ &= \beta\sum_j \frac{1}{\exp\{\beta(\varepsilon_j-\mu)\}-1} - \beta\sum_i n_i.\end{aligned} \tag{5.9}$$

These two derivatives must vanish simultaneously if n is to be contained in the fibre of the grand canonical distribution $p_{\beta,\mu}$. Rewriting the above expressions shows this is equivalent to demanding that the total energy and total number of bosonic particles can be expressed through the well-known relations

$$\begin{aligned}\sum_i n_i\varepsilon_i &= \sum_j \frac{\varepsilon_j}{\exp\{\beta(\varepsilon_j-\mu)\}-1}, \\ \sum_i n_i &= \sum_j \frac{1}{\exp\{\beta(\varepsilon_j-\mu)\}-1}.\end{aligned}$$

Whether or not a given data set n is acceptable depends on whether or not these equations have solutions for β and μ for this data set. The second derivatives are also important to the data set model formalism. A straightforward computation yields

$$\begin{aligned}
\partial_\beta^2 D(n||p_{\beta,\mu}) &= \sum_j \frac{(\varepsilon_j-\mu)^2}{(\exp\{\beta(\varepsilon_j-\mu)\}-1)^2}\exp\{\beta(\varepsilon_j-\mu)\},\\
\partial_\beta\partial_\mu D(n||p_{\beta,\mu}) &= \sum_j \frac{1}{\exp\{\beta(\varepsilon_j-\mu)\}-1} - \sum_i n_i\\
&\quad - \beta\sum_j \frac{\varepsilon_j-\mu}{(\exp\{\beta(\varepsilon_j-\mu)\}-1)^2}\exp\{\beta(\varepsilon_j-\mu)\},\\
\partial_\mu^2 D(n||p_{\beta,\mu}) &= \beta^2\sum_j \frac{1}{(\exp\{\beta(\varepsilon_j-\mu)\}-1)^2}\exp\{\beta(\varepsilon_j-\mu)\}.
\end{aligned}$$

The mixed derivative still depends on the data set n. However, this dependency only involves the sum of all occupation numbers and can thus be rewritten using the equalities right above when n is in the fibre of $p_{\beta,\mu}$. Furthermore, this will cancel out the first sum in the expression, thereby simplifying the derivative. This means the metric tensor can easily be written down as

$$g(\beta,\mu) = \sum_j \frac{\exp\{\beta(\varepsilon_j-\mu)\}}{(\exp\{\beta(\varepsilon_j-\mu)\}-1)^2}\begin{pmatrix} (\varepsilon_j-\mu)^2 & -\beta(\varepsilon_j-\mu) \\ -\beta(\varepsilon_j-\mu) & \beta^2 \end{pmatrix}. \tag{5.10}$$

This metric can be shown to be positive definite through direct computation. For any vector $\vec{v} \in \mathrm{T}_{\beta,\mu}\mathbb{M}$ it holds that

$$g(\vec{v},\vec{v}) = \sum_j \frac{\exp\{\beta(\varepsilon_j-\mu)\}}{(\exp\{\beta(\varepsilon_j-\mu)\}-1)^2}\left(v^\beta(\varepsilon_j-\mu)-\beta v^\mu\right)^2 \geqslant 0.$$

The next step in the search for the canonical coordinates of the model distributions is to determine the connection. Since only the mixed derivative of the divergence depends on the data set n, it can quickly be seen that

$$\omega^\beta{}_{\beta\beta} = \omega^\mu{}_{\beta\beta} = \omega^\beta{}_{\mu\mu} = \omega^\mu{}_{\mu\mu} = 0.$$

The two remaining independent coefficients, $\omega^\beta{}_{\beta\mu}$ and $\omega^\mu{}_{\beta\mu}$, are only slightly harder to find. $\mathbb{X}$ is a discrete set, but since it can be thought of as embedded in $\mathbb{R}^N$ with N (at least) the number of available energy levels if this is finite, it is possible to use either the expression (4.13) to compute the coefficients or the discretised version (4.12). It is the latter expression that will be used here. This means fixing values for the parameters β and μ and then finding appropriate data sets with which to compute the coefficients of the connection. The computations of connection coefficients each include two data sets: one data set n which is contained in the fibre of $p_{\beta,\mu}$ and another one denoted as $n^{(k)}$ for which

$$\partial_l D(n^{(k)}||p_{\beta,\mu}) = \delta_l^k,$$

or another invertible matrix if this is more practical. Since the derivatives of the divergence only depend on the data sets through the quantities

$$\sum_i n_i \quad \text{and} \quad \sum_i n_i \varepsilon_i,$$

which are constant in the fibres, it is not too difficult to find these "off-fibre" data sets. The condition for $n^{(\beta)}$ is that

$$\partial_\beta D(n^{(\beta)}||p_{\beta,\mu}) \neq 0 \quad \text{and} \quad \partial_\mu D(n^{(\beta)}||p_{\beta,\mu}) = 0.$$

From the second of the conditions, it can be seen that

$$\sum_i n_i^{(\beta)} = \sum_i n_i = \sum_j \frac{1}{\exp\{\beta(\varepsilon_j - \mu)\} - 1}.$$

This is enough to compute the the connection coefficient ${\omega^\beta}_{\beta\mu}$, given by

$$\begin{aligned} {\omega^\beta}_{\beta\mu} &= \frac{\partial_\beta\partial_\mu D(n^{(\beta)}||p_{\beta,\mu}) - \partial_\beta\partial_\mu D(n||p_{\beta,\mu})}{\partial_\beta D(n^{(\beta)}||p_{\beta,\mu})} \\ &= \frac{\sum_i n_i^{(\beta)} - \sum_i n_i}{\partial_\beta D(n^{(\beta)}||p_{\beta,\mu})} \\ &= 0. \end{aligned}$$

This means only the coefficient ${\omega^\mu}_{\beta\mu}$ remains to be determined. The conditions on the data set $n^{(\mu)}$ can be expressed as

$$\partial_\mu D(n^{(\mu)}||p_{\beta,\mu}) \neq 0 \quad \text{and} \quad \partial_\beta D(n^{(\mu)}||p_{\beta,\mu}) = 0. \tag{5.11}$$

The first condition implies

$$\begin{aligned} \partial_\mu D(n^{(\mu)}||p_{\beta,\mu}) &= \beta \sum_j \frac{1}{\exp\{\beta(\varepsilon_j - \mu)\} - 1} - \beta \sum_i n_i^{(\mu)} \\ &= \beta\left(\sum_i n_i - \sum_i n_i^{(\mu)}\right) \end{aligned}$$

must differ from zero. Since $n^{(\mu)}$ is arbitrary within the constraints (5.11), the expression between parentheses can be chosen to be equal to 1. This suffices to determine the final (and only) independent connection coefficient, being

$$\begin{aligned}\omega^{\mu}{}_{\beta\mu} &= \frac{\partial_\beta\partial_\mu D(n^{(\mu)}||p_{\beta,\mu}) - \partial_\beta\partial_\mu D(n||p_{\beta,\mu})}{\partial_\mu D(n^{(\mu)}||p_{\beta,\mu})}\\ &= \frac{-\sum_i n_i^{(\mu)} + \sum_i n_i}{\beta}\\ &= \frac{1}{\beta}.\end{aligned}$$

By simple substitution it can be verified that this connection will indeed make the Hessian of the divergence independent of its first argument and equal to the metric found before, as it is given in expression (5.10).
To find the canonical parameters making up the affine coordinates for this model manifold, attention is turned towards equation (4.17). The partial differential equations to be solved in this example are

$$\begin{aligned}\partial_\beta^2\Theta^i &= 0,\\ \partial_\beta\partial_\mu\Theta^i &= \beta^{-1}\partial_\mu\Theta^i,\\ \partial_\mu^2\Theta^i &= 0.\end{aligned}$$

From the first and the last of these equations, it follows that

$$\Theta^i(\beta,\mu) = A^i\mu\beta + B^i\mu + C^i\beta + D^i,$$

where the Roman capital letters are constants. The second differential equation demands that $A^i = \beta^{-1}(A^i\beta + B^i)$ and thus $B^i = 0$. The other three constants may be chosen freely and with a proper choice one obtains the canonical parameters mentioned earlier,

$$\theta^1 = \beta \quad \text{and} \quad \theta^2 = -\beta\mu.$$

Since the connection is flat and its coefficients can be expressed in a fairly simple way in the coordinates β and μ, it is worthwhile to see what the geodesics and covariant constant vector fields would look like in the coordinate system determined by β and μ.
Given a vector $\vec{v}$ at one particular point of the manifold $\mathbb{M}$, it is possible to construct a covariant constant vector field everywhere through parallel

transport. Along an arbitrary curve parametrised by t, the covariant constant vector field satisfies the equations

$$\frac{\mathrm{d}v^\beta}{\mathrm{d}t} = 0 \quad \text{and} \quad \frac{\mathrm{d}v^\mu}{\mathrm{d}t} = -\frac{1}{\beta}\frac{\mathrm{d}\mu}{\mathrm{d}t}v^\beta - \frac{1}{\beta}\frac{\mathrm{d}\beta}{\mathrm{d}t}v^\mu$$

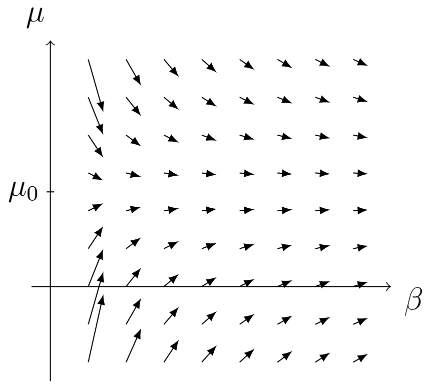
The first equation implies that v^β has the same value everywhere on $\mathbb{M}$. Because this component is constant, it is possible to rewrite the second equation as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\beta v^\mu}{v^\beta}\right) = -\frac{\mathrm{d}\mu}{\mathrm{d}t}.$$

This differential equation has for its solution

$$v^\mu(t) = \frac{\mu_0 - \mu(t)}{\beta(t)}v^\beta,$$

where μ_0 is an integration constant which can be determined by choosing the value of the vector field at a given point. An example of a covariant constant vector field for this connection is depicted in the drawing. This may not look to the reader as a vector field which is parallel with itself everywhere but this is in fact the case. The origin of the possible confusion is that a coordinate system (and thus a coordinate frame) is chosen for the illustration in which the connection coefficients do not vanish.



In order to get a better understanding of the behaviour of this connection, also the geodesics can be considered. The general geodesic equations are given by the two non-linear differential equations

$$\frac{\mathrm{d}^2\beta}{\mathrm{d}t^2} = 0 \quad \text{and} \quad \frac{\mathrm{d}^2\mu}{\mathrm{d}t^2} = -\frac{2}{\beta}\frac{\mathrm{d}\beta}{\mathrm{d}t}\frac{\mathrm{d}\mu}{\mathrm{d}t}.$$

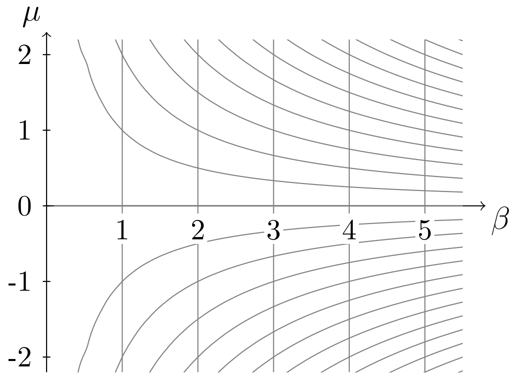
From the first equation it follows that $\beta(t) = At + B$. This allows the second differential equation to be rewritten into the form

$$\frac{\mathrm{d}^2\mu}{\mathrm{d}t^2}(t) = -2\frac{A}{At+B}\frac{\mathrm{d}\mu}{\mathrm{d}t}(t).$$

A particular solution is achieved when $A = 0$, making β constant and $\mu(t) = Ct + D$. Otherwise, this equation can be solved by elementary methods such as separation of variables and order reduction to obtain the result

$$\mu(t) = \mu(t_0) + \dot{\mu}(t_0)\left(\frac{1}{At_0+B} - \frac{1}{At+B}\right).$$

This family of curves excludes the β- or θ^1-curves, which are a particular solution to the geodesic equations corresponding to $A = 0$. The drawing below shows a selection of canonical coordinate curves—all of which are geodesics—as they appear in the original coordinate system of the parameters β and μ.



5.4 The von Mises-Fisher distributions

This example is divided into two separate modelling problems. In both cases, the manifold is a subset of an exponential family. This is very instructive as the geometry of these submanifolds as induced by the divergence is exactly what is expected from a submanifold of a Euclidean space—even though the containing space is not endowed with a Euclidean metric.

5.4.1 Fixing the width of the distribution

The von Mises-Fisher distributions are families of distributions on the sphere [85]. That is, they belong to the set of distributions p for which

$$\sum_i \mathbb{E}_p[x_i]^2 = 1. \tag{5.12}$$

In particular, the von Mises-Fisher distributions take the exponential form

$$p_{\kappa,\mu}(x) = \exp\{-\Phi(\kappa) + \kappa\mu^i x_i\} \quad \text{where} \quad \sum_i (\mu^i)^2 = 1. \tag{5.13}$$

This notation employs the traditional parametrisation but it is easy to see that this is an exponential family by taking as the canonical parameters $\theta^i = -\kappa\mu^i$, analogous to the situation for the grand canonical distribution of bosonic particles treated previously. It is not the intention to repeat the above example with a few extra degrees of freedom. Instead the computations here will be restricted to three-dimensional data and—more importantly—only a subset of the von Mises-Fisher distributions will be considered as the model manifold $\mathbb{M}$. In particular the parameter κ, which is a measure for the width of the distribution, will be held constant. The manifold $\mathbb{M}$ thus obtained is topologically equivalent to the 2-sphere, which can be seen by employing a parametrisation by spherical polar coordinates[1] θ and φ,

$$p_{\theta,\varphi}(x) = \exp\{-\Phi(\kappa) + \kappa(\sin(\theta)\cos(\varphi)x_1 + \sin(\theta)\sin(\varphi)x_2 + \cos(\theta)x_3)\}.$$

The choice for the symbol θ is made in accordance with the usual names of the spherical polar coordinates and it should not be mistaken for a canonical coordinate of an exponential family.

The set $\mathbb{X}$ of data sets is the space of all possible probability distributions

[1] The usual remarks for the use of this parametrisation apply, that is when $\theta = 0$ or $\theta = \pi$, the coordinate φ is not well-defined and so θ and φ do not form a proper coordinate system. As this will not hinder the treatment of this example too much, this parametrisation will be used nonetheless.

on the sphere as defined by equation (5.12). The model manifold $\mathbb{M}$ is the two dimensional set of von Mises-Fisher distributions with a fixed value of κ. As the divergence D it proves convenient to opt for the Kullback-Leibler divergence function,

$$\begin{aligned} D(p||p_{\kappa,\mu}) &= -S(p) - \int_X p(x)\left[-\Phi(\kappa) + \kappa\mu^i x_i\right] \mathrm{d}x \\ &= -S(p) + \Phi(\kappa) \\ &\quad - \kappa(\sin(\theta)\cos(\varphi)\mathbb{E}_p[x_1] + \sin(\theta)\sin(\varphi)\mathbb{E}_p[x_2] + \cos(\theta)\mathbb{E}_p[x_3]), \end{aligned}$$

where as usual $S(p)$ is the Shannon entropy (3.8) of the probability distribution p. The derivatives of this divergence are given by

$$\begin{aligned} \partial_\theta D(p||p_{\theta,\varphi}) &= -\kappa(\cos(\theta)\cos(\varphi)\mathbb{E}_p[x_1] + \cos(\theta)\sin(\varphi)\mathbb{E}_p[x_2] - \sin(\theta)\mathbb{E}_p[x_3]), \\ \partial_\varphi D(p||p_{\theta,\varphi}) &= -\kappa(-\sin(\theta)\sin(\varphi)\mathbb{E}_p[x_1] + \sin(\theta)\cos(\varphi)\mathbb{E}_p[x_2]). \end{aligned}$$

Though it is formally possible to solve these equations together with the normalisation condition (5.12) to find which distributions p make up the fibres of the von Mises-Fisher distribution $p_{\theta,\varphi}$, there is an easier way. The divergence $D(p||p_{\theta,\varphi})$ as defined above is minimal when the expression $\mu^i\mathbb{E}_p[x_i]$ attains its maximal value. Since this can be viewed as the inner product of two vectors of unit norm, the maximum is obtained when the two vectors are equal, that is

$$\begin{aligned} \mathbb{E}_p[x_1] &= \sin(\theta)\cos(\varphi), \\ \mathbb{E}_p[x_2] &= \sin(\theta)\sin(\varphi), \\ \mathbb{E}_p[x_3] &= \cos(\theta). \end{aligned}$$

Note that there is also another set of distributions p for which the derivatives of $D(p||p_{\theta,\varphi})$ vanish—those distributions for which the values $\mathbb{E}_p[x_i]$ take the negative of the values above. This is a local maximum of the divergence function, however, and hence it is no candidate for the required solution. The second derivatives of the divergence are found to equal

$$\begin{aligned} \partial_\theta^2 D(p||p_{\theta,\varphi}) &= \kappa(\sin(\theta)\cos(\varphi)\mathbb{E}_p[x_1] + \sin(\theta)\sin(\varphi)\mathbb{E}_p[x_2] + \cos(\theta)\mathbb{E}_p[x_3]), \\ \partial_\theta\partial_\varphi D(p||p_{\theta,\varphi}) &= \kappa(\cos(\theta)\sin(\varphi)\mathbb{E}_p[x_1] - \cos(\theta)\cos(\varphi)\mathbb{E}_p[x_2]), \\ \partial_\varphi^2 D(p||p_{\theta,\varphi}) &= \kappa(\sin(\theta)\cos(\varphi)\mathbb{E}_p[x_1] + \sin(\theta)\sin(\varphi)\mathbb{E}_p[x_2]). \end{aligned}$$

Combined with the expressions for $\mathbb{E}_p[x_i]$ that make up the conditions for p to be an element of the $p_{\theta,\varphi}$-fibre, knowledge of these second derivatives allows for the metric tensor to be expressed. Its components read

$$g_{\theta\theta}(\theta,\varphi) = \kappa, \quad g_{\theta\varphi}(\theta,\varphi) = 0, \quad g_{\varphi\varphi}(\theta,\varphi) = \kappa\sin^2(\theta).$$

This is the metric tensor of a 2-sphere with radius $\kappa^{1/2}$. This is a clear example of a model where the geometry introduced by the divergence coincides with the geometry that is expected from the manifold $\mathbb{M}$ itself, even though the manifold itself is not isomorphic to $\mathbb{R}^n$. However, this does not ensure the connection ∇ will be the metric connection on such a sphere. In fact, at this point there is no guarantee that a connection as introduced in the previous chapter would even exist in this example. A formal investigation is thus required.

The determination of the connection coefficients will proceed by first finding curves $\varepsilon \to X^k(\varepsilon)$ through the space $\mathbb{X}$. Both the first and second derivatives of the divergence function in this example depend only on the expectation values of the random variables x_i, which means it suffices to focus on these expectation values. However, these values are not independent: they too lie on a sphere just as the parameters μ^i do. The use of this knowledge reveals that the result that one family of desirable curves can be obtained (as functions of a parameter ψ) through the expressions

$$\begin{aligned}\mathbb{E}_p[x_1] &= \sin(\theta-\psi)\cos(\varphi),\\ \mathbb{E}_p[x_2] &= \sin(\theta-\psi)\sin(\varphi),\\ \mathbb{E}_p[x_3] &= \cos(\theta-\psi).\end{aligned}$$

Analogously, the second family of curves is obtained (as functions of ξ):

$$\begin{aligned}\mathbb{E}_p[x_1] &= \sin(\theta)\cos(\varphi-\xi),\\ \mathbb{E}_p[x_2] &= \sin(\theta)\sin(\varphi-\xi),\\ \mathbb{E}_p[x_3] &= \cos(\theta).\end{aligned}$$

When either of these parameters are equal to zero, the curves find themselves in a point representing a fibre of the model $p_{\theta,\varphi}$ as intended. Using these curves over the sphere of the expectation values, it is found that

$$\begin{aligned}\left.\frac{\partial^2 D}{\partial\psi\partial\theta}\right|_{\psi=\xi=0} &= \kappa, & \left.\frac{\partial^2 D}{\partial\psi\partial\varphi}\right|_{\psi=\xi=0} &= 0,\\ \left.\frac{\partial^2 D}{\partial\xi\partial\theta}\right|_{\psi=\xi=0} &= 0, & \left.\frac{\partial^2 D}{\partial\xi\partial\varphi}\right|_{\psi=\xi=0} &= \kappa\sin^2(\theta).\end{aligned}$$

This yields exactly the metric tensor, even though this is a coincidental consequence of the choice of the X^k rather than a general property—another choice of curves would have been equally valid but would have yielded different expressions. While the result is not the Kronecker-delta one would desire under ideal circumstances, the invertibility of the metric tensor means these curves can indeed be used to define the connection coefficients. The following step is to use the above expressions for the curves through the expectation value sphere in the expressions for the second derivatives of the divergence. The resulting quantities can be derived with respect to ψ and ξ to obtain the coefficients $\omega_{k,ij} = g_{ks}\omega^{s}{}_{ij}$. The usual coefficients can then be found through a multiplication with the matrix inverse of the metric tensor. There are six derivatives to compute. These are—again suppressing some straightforward function arguments in order to reduce the burden of notation—

$$\begin{aligned}
\frac{\partial}{\partial\psi}\frac{\partial^2 D}{\partial\theta^2}\bigg|_{\psi=\xi=0} &= 0, & \frac{\partial}{\partial\xi}\frac{\partial^2 D}{\partial\theta^2}\bigg|_{\psi=\xi=0} &= 0,\\
\frac{\partial}{\partial\psi}\frac{\partial^2 D}{\partial\theta\partial\varphi}\bigg|_{\psi=\xi=0} &= 0, & \frac{\partial}{\partial\xi}\frac{\partial^2 D}{\partial\theta\partial\varphi}\bigg|_{\psi=\xi=0} &= \kappa\cos(\theta)\sin(\theta),\\
\frac{\partial}{\partial\psi}\frac{\partial^2 D}{\partial\varphi^2}\bigg|_{\psi=\xi=0} &= -\kappa\sin(\theta)\cos(\theta), & \frac{\partial}{\partial\xi}\frac{\partial^2 D}{\partial\varphi^2}\bigg|_{\psi=\xi=0} &= 0.
\end{aligned}$$

Since the matrix multiplication that needs to be performed involves a square matrix, this operation is straightforward and the two independent non-vanishing connection coefficients appear as

$$\omega^{\theta}{}_{\varphi\varphi}(\theta,\varphi) = -\sin(\theta)\cos(\theta) \quad \text{and} \quad \omega^{\varphi}{}_{\theta\varphi}(\theta,\varphi) = \frac{\cos(\theta)}{\sin(\theta)}.$$

These coincide with the coefficients of the metric connection on a spherical surface (see for example [86]). It is not necessary to compute the curvature tensor to see if it vanishes: the reader will no doubt agree that the familiar metric connection of a spherical surface will not be found to exhibit flatness. The example is thus a very instructive one. It shows that for a properly chosen divergence function, the geometry introduced by the formalism of data set models will coincide with the geometry expected from the choice of the model manifold $\mathbb{M}$, something which may come unexpected from readers familiar with information geometry. After all, a prominent role there is played by connections which are not metric. Also, since this connection is not flat, it is an example of a data set model where a connection exists but where the metric is not of the Hessian type.

5.4.2 A cylindrical submanifold

With the knowledge obtained in the previous example regarding von Mises-Fisher distributions, it is easy to consider a different submanifold of the three-dimensional family (5.13). In particular, this example will treat a submanifold homeomorphic to the half-cylinder, that is the two-parameter family of distributions

$$p_{\varphi,\lambda}(x) = \exp\{-\Xi(\kappa,\lambda) + \kappa(\cos(\varphi)x_1 + \sin(\varphi)x_2) - \lambda x_3\},$$

where κ is again a positive constant and $\lambda > 0$ is now an independent parameter. This distribution is essentially a product of the one-dimensional von Mises(-Fisher) distribution and the exponential distribution. The motivation for computing this very similar example is to see if here a flat connection follows. This might be expected from the analogy in the previous example, where the geometry of the model manifold coincided with the sphere with which it is homeomorphic. Cylinder mantles are known to have flat metric connections—a fact which follows from their ability to be unrolled onto $\mathbb{R}^2$ without distorting the intrinsic geometry of the surface.

The data sets which will be modelled by these new cylindrical von Mises-Fisher distributions are those distributions p for which

$$\mathbb{E}_p[x_1]^2 + \mathbb{E}_p[x_2]^2 = 1 \quad \text{and} \quad \mathbb{E}_p[x_3] > 0.$$

Unlike in the previous example, it is actually necessary to know the normalisation function Ξ in order to complete the computations. The distribution $p_{\varphi,\lambda}$ must be normalised and thus

$$\begin{aligned} 1 &= \int_X \exp\{-\Xi(\kappa,\lambda) + \kappa(\cos(\varphi)x_1 + \sin(\varphi)x_2) - \lambda x_3\}\mathrm{d}x \\ &= \exp\{-\Xi(\kappa,\lambda)\} \int_{S^1} \exp\{\kappa(\cos(\varphi)x_1 + \sin(\varphi)x_2)\}\mathrm{d}\ell \\ &\quad \times \int_0^\infty \exp\{-\lambda x_3\}\mathrm{d}x_3 \\ &= \exp\{-\Xi(\kappa,\lambda)\}\frac{2\pi I_0(\kappa)}{\lambda}, \end{aligned}$$

where I_0 is the modified Bessel function of order 0 [86]. This means

$$\Xi(\kappa,\lambda) = \ln\left(\frac{2\pi I_0(\kappa)}{\lambda}\right).$$

The remaining part of the computation is analogous to the treatment above and so less details will be supplied so as not to burden the reader. The choice of divergence is again the Kullback-Leibler divergence, which has derivatives in this example given by

$$\begin{aligned}\partial_\varphi D(p||p_{\varphi,\lambda}) &= -\kappa(-\sin(\varphi)\mathbb{E}_p[x_1] + \cos(\varphi)\mathbb{E}_p[x_2]),\\ \partial_\lambda D(p||p_{\varphi,\lambda}) &= -\frac{1}{\lambda} + \mathbb{E}_p[x_3].\end{aligned}$$

The distributions p in the $p_{\varphi,\lambda}$-fibres are thus those distributions satisfying

$$\begin{aligned}\mathbb{E}_p[x_1] &= \cos(\varphi),\\ \mathbb{E}_p[x_2] &= \sin(\varphi),\\ \mathbb{E}_p[x_3] &= \lambda^{-1}.\end{aligned}$$

The second derivatives of the divergence are equal to

$$\begin{aligned}\partial^2_\varphi D(p||p_{\varphi,\lambda}) &= \kappa(\cos(\varphi)\mathbb{E}_p[x_1] + \sin(\varphi)\mathbb{E}_p[x_2]),\\ \partial_\varphi\partial_\lambda D(p||p_{\varphi,\lambda}) &= 0,\\ \partial^2_\lambda D(p||p_{\varphi,\lambda}) &= \lambda^{-2}.\end{aligned}$$

On the fibres these expressions are constants and so the metric tensor takes the form

$$g(\varphi,\lambda) = \begin{pmatrix} \kappa & 0 \\ 0 & \lambda^{-2} \end{pmatrix}.$$

The $g_{\varphi\varphi}$-component, as well as the off-diagonal components of this metric are what are to be expected for a cylinder with radius $\kappa^{1/2}$. The λ^{-2} serving as the $g_{\lambda\lambda}$-component may seem unexpected for a cylinder, but this should come as no surprise given that the model distributions attach an exponential probability density to the random variable x_3.

In order to compute the connection coefficients, proper curves must be found along which to compute the derivatives serving as the definition of the connection coefficients—just like in the previous example, preference is given to this method for practical purposes. Since only the second derivative with respect to φ of the divergence depends on the distribution p all coefficients except ${\omega^\lambda}_{\varphi\varphi}$ and ${\omega^\varphi}_{\varphi\varphi}$ can already be seen to vanish. The first of these two will also vanish. Indeed, the appropriate derivative must be taken along a curve which is parametrised by $\mathbb{E}_p[x_3]$—as that is the only p-dependency of $\partial_\lambda D(p||p_{\varphi,\lambda})$—but $\partial^2_\varphi D(p||p_{\varphi,\lambda})$ does not depend on this expectation value, making the derivative in (4.13) vanish. In order to determine the last coefficient, ${\omega^\varphi}_{\varphi\varphi}$, take curves parametrised by ξ such that along this curve

$$\mathbb{E}_p[x_1] = \cos(\varphi-\xi), \quad \mathbb{E}_p[x_2] = \sin(\varphi-\xi), \quad \mathbb{E}_p[x_3] = \mathrm{cte},$$

analogous to what happened in the spherical example. Then, along this curve, the second φ-derivative of the divergence takes the form

$$\partial_{\varphi}^{2} D(p||p_{\varphi,\lambda}) = \kappa(\cos(\varphi)\cos(\varphi-\xi)+\sin(\varphi)\sin(\varphi-\xi)).$$

Deriving this expression with respect to ξ and evaluating in $\xi = 0$ shows that also this coefficient vanishes. This means that the chosen parameters φ and λ are indeed the affine coordinates of the connection and thus the canonical parameters of the distribution.

Since all the connection coefficients vanish identically, it is not necessary to compute the curvature tensor. However, it is possible to check whether or not the Codazzi-Peterson-like equation (4.7) is satisfied. As the computation is already taking place in an affine coordinate system, it is sufficient to verify whether or not

$$\partial_{\varphi} g_{\lambda\varphi} = \partial_{\lambda} g_{\varphi\varphi} \quad \text{and} \quad \partial_{\lambda} g_{\varphi\lambda} = \partial_{\varphi} g_{\lambda\lambda}.$$

A quick peek at the coefficients of the metric tensor teaches us that all four of these quantities vanish and so both equalities are satisfied. This would mean that there does indeed exist a Massieu function Φ, the Hessian of which is the metric tensor. This may come as a surprise since the cylinder has a periodic nature in one direction and so it seems impossible that there exists a properly behaved convex function everywhere on the cylinder mantle. The answer to this is of course that there is no such function. After all, the coordinate φ is only a proper coordinate if a curve parametrised by λ is removed from the half-cylinder mantle and the argument leading to the existence of the Massieu function made use of Poincaré's lemma, which holds locally rather than globally. When the points with coordinates (π, λ) are removed, a proper coordinate is obtained and then

$$\Phi(\varphi, \lambda) = \frac{1}{2}\kappa\varphi^{2} - \ln\lambda$$

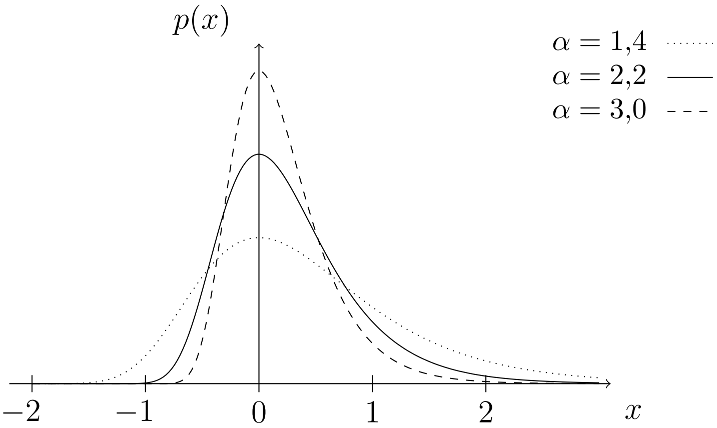
is indeed a convex function of which the matrix of second derivatives equals the metric tensor. This is a consequence of the local nature of differential geometry and one should be careful for this caveat in applications.

5.5 The Gumbel distributions

The last example in this chapter considers the Gumbel distributions [87] as models for empirical probability distributions. As such this example is very similar to the first one. The biggest difference here is that the chosen statistical model is not an exponential family. This means that the construction of the geometrical quantities of the data set model formalism will fail. In this way, it will be confirmed that the model is indeed not an exponential family.
The Gumbel distributions form a two-parameter family of probability densities which are often employed to model the distribution of the minimal or maximal values of a number of statistical samples. The Gumbel distributions, whose domain is the set of real numbers, can be written as

$$p_{\alpha,\mu}(x) = \exp\{\ln\alpha - \alpha(x-\mu) - e^{-\alpha(x-\mu)}\}.$$

The parameters are α and μ. The first one of these, is a strictly positive parameter determining the shape of the distribution. For reasons of notational convenience, the choice for α differs from the one in the literature, where $\beta = \alpha^{-1}$ is more commonly used. The effect of varying the parameter α, for $\mu = 0$, is sketched in the illustration directly below. It can be seen that larger values of α indicate a sharper distribution.



The parameter μ represents the mode of the distribution—and not the mean, as is common for Guassian distributions. The mean of the Gumbel distribution equals $\mu + \alpha^{-1}\gamma$, where γ is the Euler-Mascheroni constant. The effect of changing μ would be to perform a horizontal shift of the distribution over a distance μ and so no distributions with non-zero μ-values are included in the illustration.
It is tempting to choose as the set $\mathbb{X}$ again all possible distributions over the real numbers. Unfortunately this is untenable as this would mean certain

expectation values appearing in the computation below will not exist. This is not a problem specific to the data set model formalism and so it will be ignored by assuming that all distributions $p \in \mathbb{X}$ are such that the expectation values mentioned below do indeed exist.

Due to the appearance of the exponential functions in this expression, an obvious choice for the divergence is again that of Kullback and Leibler. This divergence, with a suitable probability distribution p as its first argument and a Gumbel distribution as its second argument, takes the form

$$\begin{aligned} D(p||p_{\alpha,\mu}) &= -S(p) - \int_X p(x) \ln p_{\alpha,\mu}(x)\ \mathrm{d}x \\ &= -S(p) - \int_X p(x) \left[\ln\alpha - \alpha(x-\mu) - e^{-\alpha(x-\mu)}\right] \mathrm{d}x, \end{aligned} \tag{5.14}$$

where S represents the Shannon entropy (3.8). In order to construct the differential geometry induced by this divergence on the manifold of Gumbel distributions, the derivatives with respect to the parameters are necessary. They are given by the expressions

$$\begin{aligned} \partial_\alpha D(p||p_{\alpha,\mu}) &= -\frac{1}{\alpha} + \mathbb{E}_p[(x-\mu)] - \mathbb{E}_p[e^{-\alpha(x-\mu)}(x-\mu)], \\ \partial_\mu D(p||p_{\alpha,\mu}) &= -\alpha + \alpha\mathbb{E}_p[e^{-\alpha(x-\mu)}]. \end{aligned}$$

This means a distribution p is in the fibre of the Gumbel distribution $p_{\alpha,\mu}$ if and only if it simultaneously satisfies the equations

$$\mathbb{E}_p[e^{-\alpha(x-\mu)}] = 1 \quad \text{and} \quad \mathbb{E}_p[\alpha(x-\mu)\{1 - e^{-\alpha(x-\mu)}\}] = 1. \tag{5.15}$$

Even though this is not a trivial computation, it can be verified that when $p = p_{\alpha,\mu}$ these equations are indeed satisfied. The second derivatives of the divergence (5.14) with respect to the parameters are given by the expressions

$$\begin{aligned} \partial_\alpha^2 D(p||p_{\alpha,\mu}) &= \frac{1}{\alpha^2} + \mathbb{E}_p[(x-\mu)^2 e^{-\alpha(x-\mu)}], \\ \partial_\alpha\partial_\mu D(p||p_{\alpha,\mu}) &= -1 + \mathbb{E}_p[e^{-\alpha(x-\mu)}] - \alpha\mathbb{E}_p[(x-\mu)e^{-\alpha(x-\mu)}], \\ \partial_\mu^2 D(p||p_{\alpha,\mu}) &= \alpha^2\mathbb{E}_p[e^{-\alpha(x-\mu)}]. \end{aligned}$$

It is clear that these expressions in general depend on the chosen probability density p. To define a metric, it is sufficient that these expressions become p-independent when p is in the fibre of $p_{\alpha,\mu}$. However, this is not the case, as is shown by a counterexample. It is only required to show that the first of the derivatives,

$$\partial_\alpha^2 D(p||p_{\alpha,\mu}) = \frac{1}{\alpha^2} + \mathbb{E}_p[(x-\mu)^2 e^{-\alpha(x-\mu)}],$$

cannot be expressed without reference to p, even when using the conditions (5.15) imposed by the vanishing of the first derivatives of the divergence. This shows there exists no metric tensor as defined in the general theoretical outset of the previous chapter. An obvious choice of distribution to provide a counterexample, be it in the largest part for computational convenience, is the family of exponential distributions with density functions

$$p_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0, \\ 0 & x < 0. \end{cases}$$

Fixing a value of λ, it is possible to determine which fibre of the manifold of Gumbel distributions contains p_λ. This computation comprises most of the work needed in demonstrating the counterexample.

The equations (5.15) impose relations between the parameters α and μ on one hand and λ on the other, in particular these are

$$\begin{aligned} 1 &= \int_0^\infty \lambda e^{-\lambda x - \alpha(x-\mu)} \mathrm{d}x \\ &= \frac{\lambda e^{\alpha\mu}}{\lambda + \alpha} \end{aligned} \tag{5.16}$$

and

$$\begin{aligned} 1 &= \alpha\lambda \int_0^\infty (x-\mu)\{1 - e^{-\alpha(x-\mu)}\} e^{-\lambda x} \mathrm{d}x \\ &= \alpha\lambda \int_0^\infty (x-\mu) e^{-\lambda x} \mathrm{d}x - \alpha\lambda e^{\alpha\mu} \int_0^\infty (x-\mu) e^{-(\alpha+\lambda)x} \mathrm{d}x \\ &= \frac{\alpha}{\lambda} - \alpha\mu - \frac{\alpha\lambda e^{\alpha\mu}}{(\alpha+\lambda)^2} + \frac{\alpha\lambda\mu e^{\alpha\mu}}{\alpha+\lambda}. \end{aligned} \tag{5.17}$$

This relation can be simplified considerably by using (5.16) to yield

$$\begin{aligned} 1 &= \frac{\alpha}{\lambda} - \alpha\mu - \frac{\alpha}{\alpha+\lambda} + \alpha\mu \\ &= \frac{\alpha^2}{\lambda(\alpha+\lambda)}. \end{aligned}$$

Essentially a quadratic equation in α, this condition can be solved for its positive root

$$\alpha = \frac{\lambda + \sqrt{\lambda^2 + 4\lambda^2}}{2} = \frac{1+\sqrt{5}}{2}\lambda.$$

Negative α-values such as the one obtained by choosing the minus sign are excluded by definition of the Gumbel distributions and so this solution does not need to be given any attention. By substituting this result in relation (5.16), an expression for μ in terms of λ can be obtained as

$$\mu = \frac{1}{\alpha} \ln \left(\frac{\lambda + \alpha}{\lambda} \right) = \frac{1}{\lambda} \cdot \frac{2}{1+\sqrt{5}} \ln \left(\frac{3+\sqrt{5}}{2} \right).$$

Hence the values of the parameters $\alpha(\lambda)$ and $\mu(\lambda)$ of a Gumbel distribution upon which p_λ is mapped are known. It is now possible to show that the second derivative of the divergence (5.14) with respect to α is not constant on the fibres. This derivative contains the term

$$\begin{aligned}
\mathbb{E}_{p_\lambda}[(x-\mu)^2 e^{-\alpha(x-\mu)}] &= \int_{-\infty}^{+\infty} p_\lambda(x)(x-\mu)^2 e^{-\alpha(x-\mu)} \mathrm{d}x \\
&= \lambda e^{\alpha\mu} \int_0^\infty (x-\mu)^2 e^{-(\lambda+\alpha)x} \mathrm{d}x \\
&= \lambda e^{\alpha\mu} \left[\frac{2}{(\lambda+\alpha)^3} - \frac{2\mu}{(\lambda+\alpha)^2} + \frac{\mu^2}{\lambda+\alpha} \right] \\
&\approx \frac{0,5}{\alpha^2} \qquad \text{(using } \alpha(\lambda) \text{ and } \mu(\lambda)\text{)}.
\end{aligned}$$

Since the Gumbel distribution $p_{\alpha,\mu}$ is an element of its own fibre, it must yield this same result when computing this expectation value if the metric tensor is to exist. A direct computation shows that

$$\begin{aligned}
\mathbb{E}_{p_{\alpha,\mu}}[(x-\mu)^2 e^{-\alpha(x-\mu)}] &= \alpha \int_{-\infty}^{\infty} (x-\mu)^2 \exp\{-2\alpha(x-\mu) - e^{-\alpha(x-\mu)}\} \mathrm{d}x \\
&= \frac{1}{\alpha^2} \int_{-\infty}^{\infty} u^2 e^{-2u-e^{-u}} \mathrm{d}u \\
&= \frac{1}{\alpha^2} \int_0^\infty (\ln t)^2 t e^{-t} \mathrm{d}t \\
&= \frac{1}{\alpha^2} \left(\gamma^2 - 2\gamma + \frac{\pi^2}{6} \right) \\
&\approx \frac{0,82}{\alpha^2}.
\end{aligned}$$

It is concluded that—for this data set model—the suggested expression for the metric is not constant along fibres. This does not mean the manifold cannot be endowed with a metric tensor at all. For instance it remains possible to use as a metric tensor the matrix of second derivatives of the divergence restricted to $\mathbb{M} \times \mathbb{M}$, evaluated on the diagonal. Using the equations (5.15) to simplify the expressions, this would yield

$$g(\alpha, \mu) \approx \begin{pmatrix} \frac{1{,}82}{\alpha^2} & \gamma \\ \gamma & \alpha^2 \end{pmatrix}.$$

This metric tensor only contains information about the manifold of Gumbel distributions and, as shown above, is not compatible with the structure of the fibres of the full data set model.
It is instructive to see if a connection could exist for this model. As the metric is not well-defined, there is no Hessian structure. The spherical submanifold of the von Mises-Fisher distribution shows that such structure does not need to be present in order to define an affine connection, however.
Once again, the first step in looking for a connection is to single out curves through the set $\mathbb{X}$ of distributions. The derivatives of the divergence are proportional to the expectation values

$$\mathbb{E}_p[e^{-\alpha(x-\mu)}] \quad \text{and} \quad \mathbb{E}_p[(x-\mu)\{1-e^{-\alpha(x-\mu)}\}]$$

and thus it is convenient to take these values as the parameters for the curves appearing in the strong version of Condition 6. Indeed, differentiating the derivatives of the divergence with respect to these parameters would yield the desired Kronecker-delta as a result. However, when trying to express the second derivatives of the divergence as a function of these parameters, in order to perform the necessary differentiation, a problem arises. Again it is the second derivative with respect to α which causes difficulties. As is stated above, this quantity is given by the expression

$$\partial_\alpha^2 D(p||p_{\alpha,\mu}) = \frac{1}{\alpha^2} + \mathbb{E}_p[(x-\mu)^2 e^{-\alpha(x-\mu)}].$$

Since this quantity is not constant within fibres, different choices of curves X^k satisfying all required conditions are still possible but which yield different results for the connection coefficients. Just as is the case with the metric, a certain choice of connection could be made but there exists no choice compatible with the fibre structure.

6. CONCLUSION AND OUTLOOK

6.1 Conclusion

This doctoral research seeks to contribute to the formulation of an abstract theory of information that is not based on probability. Instead, the mathematical foundation is provided by differential geometry. The resulting framework is called the data set model formalism.

Pursuing information theory without probability may seem counter-intuitive to many readers, not in small part due to the way this discipline is traditionally treated in textbooks. It is nevertheless an idea which has been advocated by a number of authors in the past. Recent interest in this question is motivated by numerous attempts in the literature to base quantum theory on informational principles. This endeavour may be facilitated by the availability of a sufficiently general and abstract perspective on information theory. Also advances in experimental quantum physics could in time vindicate the development of a formalism such as the one presented in this dissertation. These new results enabled by the recently mastered ability to perform weak measurements, may in time argue in favour of—or even demand—the formulation of a novel description of quantum information theory. In that case, having at hand a mathematical foundation such as the one provided by the data set model formalism could be beneficial to the involved research community.

The inspiration for this approach is found in information geometry, a field concerned with the description of probability theory and statistical models through differential geometry. The framework developed here is a proper generalisation of information geometry in the sense that the latter can be re-derived as a special case of the former. The most obvious way in which the abstraction is exercised is by dropping the demand that the data and the models are probability distributions. A related and in all likelihood more important feature is that the models need not be a subset of the data and may even be qualitatively different mathematical objects. Despite these differences, the construction of a geometric structure strongly reminiscent of information geometry is performed in this dissertation.

The geometric structure of the data set model formalism is derived from a generalised divergence function which quantifies how well a data set is described by a given model point. More in particular, a Riemannian metric and an affine connection can be constructed under suitable conditions. The metric is a generalisation of the Fisher information metric and can be employed to express the sensitivity of the inferred parameters to measurable functions of the data. Consequently, when the data set model under consideration is an exponential family of probability distributions, the Cauchy-Schwarz inequality for the metric tensor reduces to the well-known Cramér-Rao bound. The affine connection is flat and torsionless and it is a generalisation of the exponential connection taking up a prominent role in information geometry. Of central importance to the data set model geometry is a Hessian structure, where the metric can be written as the Hessian of a generalised Massieu function. The point of view that the formalism is a proper generalisation of the existing literature is further reinforced by discussing a Pythagorean theorem, which provides the link required to derive the geometry of divergence functions as it has been developed by other researchers from the data set model geometry.

The theoretical discussion is concluded by establishing a straightforward technique to determine whether or not a statistical model belongs to the exponential family and to determine the canonical parameters. While these notions are not expected to be original, they can be quite useful for someone interested in exponential families and the new formalism allows them to be applied in an elegant fashion.

The last chapter of this dissertation is devoted to working out a number of examples. Each of these illustrates one or more prominent aspects of the data set model formalism touched upon in the preceding paragraphs. The examples vary from familiar probability theory and information geometry to a linear regression method. A particular example interesting to physicists is found third in that chapter. It concerns systems of non-interacting bosonic particles and it allows, given experimentally observed occupation numbers of the energy levels of that system, to find the grand canonical distribution function which best describes the state of the system (at the time of measurement). Those readers more attracted to differential or information geometry may have their interest sparked in particular by the two examples in the fourth section. Two statistical models, both submanifolds of the von Mises-Fisher distributions, are studied there and have their data set model geometry constructed. The metric tensor and the affine connection obtained in this way are exactly the ones that would be expected if the submanifolds had been embedded in Euclidean space. This is a remarkable result since the containing space is not endowed with a Euclidean metric.

6.2 Outlook

A number of interesting questions regarding data set models remain open. One question is whether a natural construction for a one-parameter family of affine connections exists for data set models, as is the case in information geometry. A better understanding of the remarkable results of the von Mises-Fisher examples may yield insights in properties of the geometry of data set models. Should these results hold in general, they could serve to detect statistical models which are submanifolds of exponential families. It may also enable a further expansion of the data set model formalism in order to treat also models with a connection which exhibits curvature.

From the perspective of applications, the data set model approach may become a fruitful technique in quantum information theory. Recent—still unpublished—work shows that the work of Petz on positive-operator valued measures (see for instance [16]) is also encompassed by the formalism. It even offers a ground for the belief that a further extension and a mathematical simplification of that research may be possible. Whereas I personally hold the opinion that the preceding suggestion is the application which looks the most promising—at least at the time of writing—the great flexibility offered by the data set model opens up the possibility for plenty of applications. As the examples of linear regression and of the non-interacting bosons show, the data set model formalism is very suitable to function as a mathematical framework for very general fitting procedures. For this reason it could also be useful to researchers in machine learning. Their discipline is concerned with a highly varied collection of different types of what are essentially modelling problems. In particular, a significant part of their field makes use of probability theory whereas an equally important part does not. This community is thus familiar with many powerful techniques which are employed in the study of the former type, which could be extended to the latter type using the data set model formalism as a unifying framework.

CURRICULUM VITAE AND ACADEMIC OVERVIEW

Ben Anthonis

Born on April 2nd 1986, Lier (Belgium)

Contact Information

Departement Fysica
Universiteit Antwerpen (Campus Drie Eiken)
Universiteitsplein 1
B-2610 Wilrijk
Tel: +32 (0)3 265 24 76
e-mail: ben.anthonis@uantwerpen.be

Education

- 2004–2007: Bachelor in Physics at Universiteit Antwerpen (Belgium)
- 2007–2009: Master in Physics at Universiteit Antwerpen (Belgium)
- 2009–2014: PhD in science and teaching assistant at Universiteit Antwerpen (Belgium)

List of publications

- Verhulst T., Anthonis B. and Naudts J.: *Analysis of the N=4 Hubbard ring using a counting operator*, Phys. Lett. A **373** p.2109–2113 (2009)
- Naudts J., Verhulst T. and Anthonis B.: *Counting operator analysis of the discrete spectrum of some model Hamiltonians*, Phys. Lett. A **373**, p.3419–3422 (2009)
- Naudts J. and Anthonis B.: *Data set models and exponential families in statistical physics and beyond*, Mod. Phys. Lett. B **26**, 1250062 (2012)
- Naudts J. and Anthonis B.: *The Exponential Family in Abstract Information Theory*, published in Geometric Science of Information, Lecture Notes in Computer Science **8085** (eds.: Nielsen F. and Barbaresco F.), Springer Berlin Heidelberg, p.265–272 (2013)

Educational assignments

- 2009–2014: Exercise classes for 'Fysica I' to the first-year students in Bio-engineering and Chemistry,
- 2009–2014: Exercise classes for 'Fysica II' to the second-year students in Bio-engineering and Chemistry,
- 2012–2013: Exercise classes for 'Fysica voor biomedisch onderzoek' to the first-year students in Biomedical Sciences,
- Co-supervision of the bachelor theses of Quinten Collier and Amy De Schutter (2009–2010), Inge van Hooydonk (2010–2011) and Romeo Van Snick (2011–2012),
- Co-supervision of the master theses of Geert Van Wauwe (2011–2012) and Inge van Hooydonk (2012–2013),
- 2011–2014: Ombudsperson for the Bachelor in Physics program at Universiteit Antwerpen.

BIBLIOGRAPHY

[1] Jaynes E. T.: *Information Theory and Statistical Mechanics*, Phys. Rev. **106** p.620–630 (1957)

[2] Jaynes E. T.: *On the Rationale of Maximum Entropy Methods*, Proc. IEEE **70** p.939–952 (1982)

[3] Barndorf-Nielsen, O.: *Information and Exponential Families in Statistical Theory*, J. Wiley & Sons, New York (1978)

[4] Naudts J.: *Generalised Thermostatistics*, Springer Verlag (2011)

[5] Kullback S. and Leibler R. A.: *On Information and Sufficiency*, Annals of Math. Stat. **22**, p.79–86 (1951)

[6] Ingarden R. S. and Urbanik K.: *Information without probability*, Colloq. Math. **9** p.131150 (1962)

[7] Kolmogorov A. N.: *Combinatorial foundations of information theory and the calculus of probabilities*, Russian Math. Surveys **38** p.29–40 (1983)

[8] Wheeler A.: *Information, physics, quantum: the search for links*, published in Complexity, Entropy, and the Physics of Information (ed.: Zurek W.), Addison-Wesley (1990)

[9] Frieden, R. B.: *Physics from Fisher Information*, Cambridge University Press (1998)

[10] Fuchs C. A.:*Quantum Mechanics as Quantum Information (and only a little more)* arXiv:quant-ph/0205039 (2002)

[11] Brukner C. and Zeilinger A.: *Information and Fundamental Elements of the Structure of Quantum Theory*, published in *Time, Quantum and Information* p.323–354 (2003)

[12] Clifton R., Bub J. and Halvorson H.: *Characterizing Quantum Theory in Terms of Information-Theoretic Constraints*, Foundations of Physics **33** p.1561–1591 (2003)

[13] Chiribella G., D'Ariano G. M. and Perinotti P.: *Informational derivation of quantum theory*, Phys. Rev. A **84**, 012311 (2011)

[14] Masanes L. and Müller M.: *A derivation of quantum theory from physical requirements*, New J. Phys. **13** 063001 (2011)

[15] Nielsen M. A. and Chang I. L.: *Quantum Computation and Quantum Information*, Cambridge University Press (2000)

[16] Petz D.: *Quantum Information Theory and Quantum Statistics*, Springer Verlag (2008)

[17] Aharanov Y., Albert D. Z. and Vaidman L.: *How the Result of a Measurement of a Component of the Spin of a Spin-$\frac{1}{2}$ Particle Can Turn Out to be 100*, Phys. Rev. Lett. **60** p.1351–1354 (1988)

[18] Guerlin C., Bernu J., Deléglise S., Sayrin C., Gleyzes S., Kuhr S., Brune M., Raimond J.-M. and Haroche S.: *Progressive field-state collapse and quantum non-demolition photon counting*, Nature **448** 06057 (2007)

[19] Aharonov Y., Popescu S. and Tollaksen J.: *A time-symmetric formulation of quantum mechanics*. Phys. Today **63** p.2732 (2010)

[20] Aharanov Y., Vaidman L.: *The Two-State Vector Formalism of Quantum Mechanics*, published in *Time in Quantum Mechanics*, Lecture Notes in Physics **72** p369–412 (2002)

[21] Tsang M.: *Continuous Quantum Hypothesis Testing*, Phys. Rev. Lett. **108** 170502 (2012)

[22] Lundeen J. S., Sutherland B., Patel A., Stewart C. and Bamber C.: *Direct measurement of the quantum wavefunction*, Nature **474** p.188–191 (2011)

[23] Lundeen J. S. and Bamber C.: *Procedure for Direct Measurement of General Quantum States Using Weak Measurement*, Phys. Rev. Lett. **108** 070402 (2012)

[24] Hall B. C.: *Quantum Theory for Mathematicians*, Graduate Text in Mathematics **267**, Springer (2013)

[25] Alpaydin E.: *Introduction to Machine Learning*, 2nd edn. The MIT Press (2009)

[26] Smola A. and Vishwanatan S. V. N.: *Introduction to Machine Learning*, Cambridge University Press (2008)

[27] Pressley, A.: *Elementary Differential Geometry*, 2nd edn. Springer Undergraduate Mathematics Series (2012)

[28] Frankel, T.: *The Geometry of Physics: An Introduction*, 3rd edn. Cambridge University Press (2012)

[29] Amari S.-I.: *Differential geometrical methods in statistics*, Lecture notes in statistics, Springer-Verlag, (1985)

[30] Amari S.-I. and Nagaoka H.: *Methods of information geometry*, Translations of mathematical monographs **191**, American Mathematical Society (2000)

[31] Adler, I.: *A New Look at Geometry.* Originally published by John Day Co. (1966), republished by Dover Publications (2012)

[32] Euclid of Alexandria, Heath T.L. and Densmore D.: *Euclid's Elements* (Heath translation) originally published by Cambridge University Press (1908), republished by Green Lion Press (2002)

[33] Riemann, G. B.: *Über die Hypothesen, welche der Geometrie zu Grunde liegen*, inaugural lecture at University of Göttingen (1854), published later by Dedekind R., in Abhandlungen der Kniglichen Gesellschaft der Wissenschaften zu Göttingen **13** (1867)

[34] Willard, S.: *General Topology*, originally published by Addison-Wesley (1970), republished by Dover Publications (2004)

[35] Pistone, G.: *Examples of the Application of Nonparametric Information Geometry to Statistical Physics*, Entropy **15**, p.4042–4065 (2013)

[36] Pistone, G. and Sempi, C.: *An Infinite-Dimensional Geometric Structure on the Space of all the Probability Measures Equivalent to a Given One*, Ann. Stat. **23** p.1543–1561 (1995)

[37] Taubes C. H.: *Differential Geometry: Bundles, Connections, Metrics and Curvature*, Oxford Graduate Texts in Mathematics **23**, Oxford University Press (2011)

[38] Fisher R. A.: *Theory of Statistical Estimation*, Math. Proc. of the Cambridge Phil. Soc. **22** p.700–725 (1925)

[39] Rao C. R.: *Information and the Accuracy Attainable in the Estimation of Statistical Parameters*, Bull. Calcutta Math. Soc. **37** , p.81–91 (1945)

[40] Ohara A.: *Geometric Study for the Legendre Duality of generalized entropies and its applications to the porous medium equation*, Eur. Phys. J. B, **70** p.15–28 (2009)

[41] Skovgaard L. T.: *A Riemannian Geometry of the Multivariate Normal Model*, Scandinavian Journal of Statistics **11** p.211–223 (1984)

[42] Amari S.-I., Barndorff-Nielsen O. E., Kass R. E., Lauritzen S. L. and Rao C. R.: *Differential geometry in statistical inference*, published by the Institute of Mathematical Statistics (1987)

[43] Tsallis, C.: *What are the numbers experiments provide?*, Química Nova **17** p.468–471 (1994)

[44] Naudts, J.: *Estimators, escort probabilities, and ϕ-exponential families in statistical physics*, J. Ineq. Pure Appl. Math. **5** Art. 102 (2004)

[45] Jeffreys H.: *An Invariant Form for the Prior Probability in Estimation Problems*, Proc. Roy. Soc. London Ser. A **186** p.453–461 (1946)

[46] Chentsov, N. N.: *Statistical Decision Rules and Optimal Inference*, Translations of Mathematical Monographs **53**, American Mathematical Society (1982)

[47] Efron B.: *Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency)*, Annals of Statistics **3** p.1189-1242 (1975)

[48] Massieu F.: *Mémoire sur les Fonctions caractéristiques des divers fluides et sur la théorie des vapeurs*, Mémoires des Savants étrangers **22** p.1–92 (1876)

[49] Reeds J.: *Discussion to Efron's paper* (reference [47]), Annals of Statistics **3** 1234–1238 (1975)

[50] Dawid A. P.: *Discussion to Efron's paper* (reference [47]), Annals of Statistics **3**, 1231–1234 (1975)

[51] Madsen L. T.: *The Geometry of Statistical Models—A Generalization of Curvature* Research Report **79** Stat. Res. Unit, Danish Medical Research Council (1979)

[52] Amari S.-I.: *Differential Geometry of Curved Exponential Families—Curvatures and Information Loss*, Annals of Statistics **10** p.357–385 (1982)

[53] Amari S.-I.: *Information Geometry on Hierarchy of Probability Distributions*, IEEE Transactions on Information Theory **47** p.1701-1711 (2001)

[54] Shannon. C. E.: *A Mathematical Theory of Communication*, Bell System Technical Journal **27**, p.379–423,623–656 (1948)

[55] Csiszár I.: *Information-type measures of difference of probability distributions and indirect observation* Studia Scientiarum Mathematicarum Hungarica **2** p.229-318 (1967)

[56] Bregman L. M.: *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comp. Math. and Math. Phys. **7** p200–217 (1966)

[57] Murate N., Takenouchi T., Kanamori T. and Eguchi S.: *Information Geometry of U-Boost and Bregman Divergence*, Neural Computation **16** p.1437–1481 (2004)

[58] Eguchi S.: *Geometry of minimum contrast*, Hiroshima Math. J. **22** p.631-647 (1992)

[59] Amari S.-I. and Cichocki A.: *Information Geometry of Divergence functions*, Bull. of the Polish Ac. of Sciences **58** p.183–195 (2010)

[60] Eguchi S.: *Second order efficiency of minimum contrast estimators in a curved exponential family*, Annals of Statistics **11** p.793–803 (1983)

[61] Eguchi S.: *A differential geometric approach to statistical inference on the basis of contract functionals*, Hiroshima Math. J. **15** p.341–391 (1985)

[62] Carathéodory C.: *Untersuchen über die Grundlagen der Thermodynamik*, Math. Ann. **67** p.355–386 (1909)

[63] Carathéodory C.: *Über die Bestimmung der Energie und der absoluten Temperatur mit Hilfe von reversiblen Prozessen*, Sitzber. Preuss. Akad. Wiss. Phys. K1 p.39–47 (1925)

[64] Pogliano L. and Berberan-Santos M. N.: *Constantin Carathéodory and the axiomatic thermodynamics*, J. Math. Chem. **28** p.313–324 (2000)

[65] Tisza L.: *Generalized Thermodynamics*, The MIT Press (1966)

[66] Griffiths R. B. and Wheeler J. C.: *Critical Points in Multicomponent Systems*, Phys. Rev. A **2** p.1047–1064 (1970)

[67] Weinhold F.: *Metric geometry of equilibrium thermodynamics I–V*, J. Chem. Phys. **63** p.2479–2501 (1975)

[68] Andresen B., Berry S. R., Gilmore R., Ihrig E. and Salamon P.: *Thermodynamic geometry and the metrics of Weinhold and Gilmore*, Phys. Rev. A **37** p.845–848 (1988)

[69] Ruppeiner G.: *Thermodynamics: A Riemannian geometric model*, Phys. Rev. A **20** p.1608–1613 (1979)

[70] Crooks G. E.: *Measuring thermodynamic length*, Phys. Rev. Lett. 100602 (2007)

[71] Gilmore R.: *Length and curvature in the geometry of thermodynamics*, Phys. Rev. A **30** p.1994–1997 (1984)

[72] Gilmore R.: *Uncertainty relations of statistical mechanics*, Phys. Rev. A. **31** p.3237–3239 (1985)

[73] Gibbs W.: *Elementary Principles in Statistical Mechanics*, Yale University Press (1902)

[74] Naudts J. and Anthonis B.: *Data set models and exponential families in statistical physics and beyond*, Mod. Phys. Lett. B **26**, 1250062 (2012)

[75] Topsøe F.: *Game Theoretical Optimization inspired by Information Theory*, J. Global Optim. **43** p.553-564 (2009)

[76] Topsøe F.: *Towards operational interpretations of generalized entropies*, J. Phys.: Conf. Ser. **201** (2010)

[77] Topsøe F.: *Elements of the cognitive universe*, unpublished, available online at http://www.math.ku.dk/~topsoe/isit2011.pdf (2011)

[78] Naudts J. and Anthonis B.: *The Exponential Family in Abstract Information Theory*, published in Geometric Science of Information, Lecture Notes in Computer Science **8085** (eds.: Nielsen F. and Barbaresco F.), Springer Berlin Heidelberg, p.265–272 (2013)

[79] Ambrosio L., Gigli N. and Savaré G.: *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics, ETH Zürich, Birkhäuser Verlag (2005)

[80] Amari, S.-I.: *Information Geometry and Its Applications: Convex Function and Dually Flat Manifold*, published as a chapter in Emerging Trends in Visual Computing, Lecture Notes in Computer Science **5416** p.75–102 (2009)

[81] Ohara A., Matsuzoe H. and Amari S.-I.: *Conformal Geometry of Escort Probability and its Applications*, Mod. Phys. Lett. B **26** 1250063 (2012)

[82] Codazzi, D.: *Sulle coordinate curvilinee d'una superficie dello spazio*, Ann. Math. Pura applicata **2** p.101-119 (1869)

[83] Shima H.: *The Geometry of Hessian Structures*, World Scientific (2007)

[84] Eguchi S.: *Information Geometry and Statistical Pattern Recognition*, Sugaku Expositions **19** p.197–216 (2006)

[85] Fisher R. A.: *Dispersion on a sphere*, Proc. Roy. Soc. London Ser. A **217** p.295–305 (1953)

[86] Arfken G. B., Weber H. J. and Harris F. E.: *Mathematical Methods for Physicists: A comprehensive guide*, 6th edn. Academic Press (2005)

[87] Gumbel E.J.: *Les valeurs extrêmes des distributions statistiques*, Ann. Inst. Henri Poincaré **5** p.115-158 (1935)