Information geometry for neural networks

Daniel Wagenaar

6th April 1998

Information geometry is the result of applying non-Euclidean geometry to probability theory. The present work introduces some of the basics of information geometry with an eye on applications in neural network research. The Fisher metric and Amari's α-connections are introduced and a proof of the uniqueness of the former is sketched. Dual connections and dual coordinate systems are discussed as is the associated divergence. It is shown how information geometry promises to improve upon the learning times for gradient descent learning.
Due to the inclusion of an appendix about Riemannian geometry, this text should be mostly self-contained.

Information geometry for neural networks by Daniel Wagenaar,
Centre for Neural Networks, King's College London, April 1998

Project option for the MSc course in Information Processing and Neural Networks

Supervisor: Dr A. C. C. Coolen
Second examiner: Prof R. F. Streater

CONTENTS

INTRODUCTION

Information geometry is the result of applying the ideas of non-Euclidean geometry to probability theory. Although interest in this subject can be traced back to the late 1960's, it reached maturity only through the work of Amari in the 1980's. His book [1] is still the canonical reference for anyone wishing to learn about it.

One of the fundamental questions information geometry helps to answer is: 'Given two probability distributions, is it possible to define a notion of "distance" between them?' An important application in neural networks is the gradient descent learning rule, which is used to minimize an error function by repeatedly taking small steps in parameter space. Traditionally, this space was tacitly assumed to have a trivial (flat) geometry. Information geometry shows that this assumption is false, and provides a theoretical recipe to find a gradient descent-style rule for any given network which can be shown to be optimal. In this way it promises a potentially dramatic improvement of learning time in comparison to the traditional rule.

In the present work, I describe some of the basics of information geometry, with the applicability to neural networks as a guide. Since my background in theoretical physics has given me more intuition about geometry than about probability theory, this text takes geometry rather than probability as its starting point. Having said that, I have avoided using the abstract language of modern differential geometry, and opted for the slightly less austere framework of Riemannian geometry. Appendix A introduces all the geometric concepts used in the main text, with the aim of making it accessible for readers with little previous experience with non-Euclidean geometry. That said, I have to admit that the appendix is rather compact, and that textbooks may offer a more gentle introduction into the subject. For example, [2] is a definite recommendation.

The rest of this text is laid out as follows: chapter 1 sets up the basic framework of information geometry, introducing a natural metric and a class of connections for families of probability distributions. Chapter 2 sets out to define some notions of duality in geometry which have a major impact on information geometry. In chapter 3 the results of the previous chapters are used to devise a more natural algorithm for parametric gradient descent that takes the geometry of the parameter space into account. Finally, in chapter 4 the properties of the metric introduced in chapter 1 are investigated in more detail, and a proof is sketched of the uniqueness of this metric.

I'd like to thank Ton Coolen for introducing me to this fascinating subject and for many useful discussions, Dr Streater for his thorough reading of the drafts, Dr Corcuera for sending me a preprint of his article on the classification of divergences and Dr Amari for bringing this article to my attention.

NOTES ON NOTATION

- Throughout this text we shall employ the Einstein summation convention, that is, summation is implied over indices that occur once upstairs and once downstairs in an expression, unless explicitely stated. For example, we write

$$x^i y_i \equiv \sum_i x^i y_i.$$

  In some cases we shall be less than scrupulous about re-using indices that are bound by implicit summation. We might for example write something like

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{tx^i y_i} = x^i y_i \, \mathrm{e}^{tx^i y_i},$$

  where the three pairs of i's are supposed to be completely independent.

- We shall sometimes use the notation

$$f|_{\mathrm{P}}$$

  to denote 'the function f, evaluated at the point P'. This is intended to leave equations more transparent than the notation $f(\mathrm{P})$, and has no other significance.

- We shall use boldface to denote vectors in $\mathbb{R}^n$, e.g. $\boldsymbol{x} \equiv (x^i)_{i=1}^n$. Vectors on manifolds (see appendix A.1) will be denoted by italic uppercase letters, e.g. $X = X^\mu \hat{e}_\mu$ [1)].

- The distinction between downstairs and upstairs labels shall be important throughout this text, as usual when dealing with Riemannian geometry [2)].

- We shall use greek indices μ, ν, ... to enumerate coordinates on manifolds.

[1)] For the benefit of those readers not well acquanted with differential geometry, we shall avoid writing vectors as differential operators, $X = X^\mu \partial_\mu$. This has the unfortunate side effect of making a small number of statements less immediately obvious. Where appropriate the differential operator form of equations shall be given in footnotes.

[2)] In appendix A we review the fundamentals of Riemannian geometry as needed for the main text.

CHAPTER 1

INFORMATION GEOMETRY

As mentioned in the introduction, information geometry is Riemannian geometry applied to probability theory. This chapter, which introduces some of the basic concepts of information geometry, does not presuppose any knowldege of the theory of probability and distributions. Unfortunately however, it does require some knowledge of Riemannian geometry. The reader is referred to appendix A for a brief introduction intended to provide the necessary background for this chapter.

1.1 Probability distributions

We shall begin by defining what we mean by a probability distribution. For our purposes, a probability distribution over some field (or set) X is a distribution $p : X \to \mathbb{R}$, such that

- $\int_X \mathrm{d}\boldsymbol{x}\; p(\boldsymbol{x}) = 1$;
- For any finite subset $S \subset X$, $\int_S \mathrm{d}\boldsymbol{x}\; p(\boldsymbol{x}) > 0$.

In the following we shall consider families of such distributions. In most cases these families will be parametrized by a set of continuous parameters $\boldsymbol{\theta} = (\theta^\mu)_{\mu=1}^N$, that take values in some open interval $M \subseteq \mathbb{R}^N$ and we write $p_{\boldsymbol{\theta}}$ to denote members of the family. For any fixed $\boldsymbol{\theta}$, $p_{\boldsymbol{\theta}} : \boldsymbol{x} \mapsto p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a mapping from X to $\mathbb{R}$.

As an example, consider the Gaussian distributions in one dimension:

$$p_{(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

In this case we may take $\boldsymbol{\theta} = (\theta^1, \theta^2) = (\mu, \sigma)$ as a parametrization of the family. Note that this is not the only possible parametrization: for example $(\theta^1, \theta^2) = (\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2})$ is also commonly used.

1.2 Families of distributions as manifolds

In information geometry, one extends a family of distributions, $F = \{p_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in M\}$, to a manifold $\mathcal{M}$ such that the points $p \in \mathcal{M}$ are in a one to one relation with the distributions $p \in F$. The parameters (θ^μ) of F can thus also be used as coordinates on $\mathcal{M}$. In doing so, one hopes to gain some insight into the structure of such a family. For example, one might hope to discover a reasonable measure of 'nearness' of two distributions in the family.

Having made the link between families of distributions and manifolds, one can try to identify which objects in the language of distributions naturally correspond to objects in the language of manifolds and vice versa. Arguably the most important objects in the language of manifolds are tangent vectors. The tangent space $T_{\boldsymbol{\theta}}$ at the point in $\mathcal{M}$ with coordinates (θ^μ) is seen to be isomorphic to the vector space spanned by the random variables[1)] $\frac{\partial \log p_{\boldsymbol{\theta}}(\cdot)}{\partial \theta^\mu}$, $\mu = 1 \ldots N$. This space is called $T^{(1)}_{\boldsymbol{\theta}}$. A vector field $A(\boldsymbol{\theta}) \in T(\mathcal{M})$:

$$A : \boldsymbol{\theta} \mapsto A(\boldsymbol{\theta}) = A^\mu(\boldsymbol{\theta}) \hat{e}_\mu \tag{1.1}$$

thus is equivalent to a random variable $A_{\boldsymbol{\theta}}(\cdot) \in T^{(1)}(\mathcal{M})$:

$$A_{\boldsymbol{\theta}}(\boldsymbol{x}) = A^\mu(\boldsymbol{\theta}) \frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta^\mu}. \tag{1.2}$$

(Just as $T(\mathcal{M})$ is the space of continuously differentiable mappings that assigns some vector $A(\boldsymbol{\theta}) \in T_{\boldsymbol{\theta}}$ to each point $\boldsymbol{\theta} \in \mathcal{M}$, $T^{(1)}(\mathcal{M})$ assigns a random variable $A_{\boldsymbol{\theta}} \in T^{(1)}_{\boldsymbol{\theta}}$.)

In view of the above equivalence we shall not find it necessary to distinguish between the vector field A and the corresponding random variable $A(\cdot)$.

(1.2) is called the *1-representation* of the vector field A. It is clearly possible to use some other basis of functionals of $p_{\boldsymbol{\theta}}$ instead of $\frac{\partial \log p_{\boldsymbol{\theta}}}{\partial \theta^\mu}$. Our present choice has the advantage that the 1-representation of a vector has zero expectation value:

$$E\big[\frac{\partial \log p_{\boldsymbol{\theta}}}{\partial \theta^\mu}\big] \equiv \int_X \mathrm{d}\boldsymbol{x}\; p_{\boldsymbol{\theta}}(\boldsymbol{x}) \frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta^\mu} = \int_X \mathrm{d}\boldsymbol{x}\; \frac{\partial p_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta^\mu} = \frac{\partial}{\partial \theta^\mu} \int_X \mathrm{d}\boldsymbol{x}\; p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\partial}{\partial \theta^\mu} 1 = 0.$$

Using other functionals can be useful, and in fact the 1-representation turns out to be just one member of the family of α-representations [1].

In order to simplify notation, we shall use

$$\ell(\boldsymbol{\theta}) \equiv \log p_{\boldsymbol{\theta}}$$

in the following. (The argument to ℓ shall be omitted when obvious from the context.) Note that $\ell(\boldsymbol{\theta})$ is a random variable, i.e. a function from X to $\mathbb{R}$: $\ell(\boldsymbol{\theta}) : \boldsymbol{x} \mapsto \log p_{\boldsymbol{\theta}}(x)$.

We shall also use the shorthand

$$\partial_\mu \equiv \frac{\partial}{\partial \theta^\mu}.$$

1.3 Distances between distributions: a metric

In several applications we are interested in distances between distributions. For example, given a distribution $p \in \mathcal{M}$ and a submanifold $\mathcal{S} \subseteq \mathcal{M}$ we may wish to find the distribution $p' \in \mathcal{S}$ that is 'nearest' to p in some sense. To give a specific example: suppose $\mathcal{M}$ is a large family of distributions containing both the gaussian and the binomial distributions of one variable as

[1)] A random variable in this context is just a function from X to $\mathbb{R}$.

subsets. One may wish to approximate a given binomial distribution by the 'nearest' exponential distribution. For this particular case, of course, an approximation formula has been found long ago. However, a general framework for constructing such approximations is seen to be useful. For all such tasks we need a notion of distance on manifolds of distributions. In other words we need a metric.

It turns out that the following is a suitable metric for manifolds of distributions:

Definition: The *Fisher metric* on a manifold of probability distributions is defined as

$$g_{\mu\nu}(\boldsymbol{\theta}) = E\big[\partial_\mu \ell(\boldsymbol{\theta})\, \partial_\nu \ell(\boldsymbol{\theta})\big]\,. \tag{1.3}$$

Obviously, this also gives us an inner product: for vector fields A and B we have:

$$\big\langle A\,,B\big\rangle\big|_{\boldsymbol{\theta}} = g_{\mu\nu}A^\mu_{\boldsymbol{\theta}}B^\nu_{\boldsymbol{\theta}} = E\big[A^\mu_{\boldsymbol{\theta}}\partial_\mu \ell(\boldsymbol{\theta})\, B^\nu_{\boldsymbol{\theta}}\partial_\nu \ell(\boldsymbol{\theta})\big] = E\big[A_{\boldsymbol{\theta}}B_{\boldsymbol{\theta}}\big]\,.$$

This last form is known to statisticians as the Fisher information of the two random variables $A_{\boldsymbol{\theta}}$ and $B_{\boldsymbol{\theta}}$. It is related to the maximum amount of information that can be inferred about $A_{\boldsymbol{\theta}}$ and $B_{\boldsymbol{\theta}}$ by a single measurement through the Cramér-Rao theorem, which we shall not discuss.

At first sight, the definition (1.3) may seem rather ad hoc. However, it has recently been proven by Corcuera and Giummolè [3] to be unique in having the following very appealing properties:

- $g_{\mu\nu}$ is invariant under reparametrizations of the sample space X;
- $g_{\mu\nu}$ is covariant under reparametrizations of the manifold (the parameter space).

This uniqueness will be used later to prove the optimality of the gradient descent rule based on the Fisher metric. The full proof as given in [3] is rather involved. In chapter 4 we shall present an outline of the proof, plus a glossary of the terminology needed to understand [3].

Before going on, let us note that the metric may also be written as

$$g_{\mu\nu} = -E\big[\partial_\mu\partial_\nu \ell(\boldsymbol{\theta})\big]\,,$$

since

$$\begin{aligned} E\big[\partial_\mu\partial_\nu\ell\big] &= \int_X \mathrm{d}\boldsymbol{x}\; p\,\partial_\mu\left(\frac{1}{p}\partial_\nu p\right) = \int_X \mathrm{d}\boldsymbol{x}\;\left\{\partial_\mu\partial_\nu p - \frac{1}{p}\,\partial_\mu p\partial_\nu p\right\} \\ &= 0 - \int_X \mathrm{d}\boldsymbol{x}\; p\left(\frac{1}{p}\partial_\mu p\right)\left(\frac{1}{p}\partial_\nu p\right) = -E\big[\partial_\mu\ell\,\partial_\nu\ell\big]\,. \end{aligned} \tag{1.4}$$

Example: Metric for a single neuron

Consider a single N-input binary neuron with output defined as follows:

$$y(t) = \mathrm{sgn}\left[\tanh\beta h(\boldsymbol{x}) + \eta(t)\right], \tag{1.5}$$

with

$$h(\boldsymbol{x}) = \sum_{i=1}^{N} J^i x_i + J^0.$$

In these formulas, J^i are connection weights, J^0 is the external field or bias, x_i are the (real valued) inputs, and $\eta(t)$ is a source of uniform random noise in $[-1,1]$.

From (1.5) we immediately find that

$$p_{\boldsymbol{J}}(y|\boldsymbol{x}) = \frac{1}{2} + \frac{1}{2} y \tanh \beta h(\boldsymbol{x}) \tag{1.6}$$

Since $p_{\boldsymbol{J}}(y,\boldsymbol{x}) = p_{\boldsymbol{J}}(y|\boldsymbol{x})p(\boldsymbol{x})$, we find

$$\ell(\boldsymbol{J}) \equiv \log p_{\boldsymbol{J}}(y,\boldsymbol{x}) = \log p_{\boldsymbol{J}}(y|\boldsymbol{x}) = \log\left[1 + y \tanh \beta h(\boldsymbol{x})\right].$$

Introducing $x_0 \equiv 1$, we may write $h(\boldsymbol{x}) = \sum_{\mu=0}^{N} J^\mu x_\mu \equiv \boldsymbol{J}\cdot\boldsymbol{x}$, to find

$$\partial_\mu \ell(\boldsymbol{J}) = \frac{y}{1 + y \tanh \beta h(\boldsymbol{x})} [1 - \tanh^2 \beta h(\boldsymbol{x})] \beta x_\mu . \tag{1.7}$$

Therefore

$$\begin{aligned} g_{\mu\nu}(\boldsymbol{J}) &= \sum_{\boldsymbol{x}\in\{-1,1\}^N} \sum_{y\in\{-1,1\}} p(\boldsymbol{x}) p_J(y|\boldsymbol{x}) \partial_\mu \ell \partial_\nu \ell \\ &= \sum_{\boldsymbol{x}} \sum_{y} \frac{1}{2} p(\boldsymbol{x}) \frac{y^2}{1 + y \tanh \beta h(\boldsymbol{x})} [1 - \tanh^2 \beta h(\boldsymbol{x})]^2 \beta^2 x_\mu x_\nu . \end{aligned}$$

Noting that $y = \pm 1$ implies $y^2 = 1$, and using the relation

$$\frac{1}{1+\tanh x} + \frac{1}{1-\tanh x} = \frac{(1-\tanh x) + (1+\tanh x)}{1 - \tanh^2 x} = \frac{2}{1-\tanh^2 x},$$

this can be simplified to

$$g_{\mu\nu}(\boldsymbol{J}) = \sum_{\boldsymbol{x}\in\{-1,1\}^N} p(\boldsymbol{x}) \left[1 - \tanh^2(\beta \boldsymbol{J}\cdot\boldsymbol{x})\right] \beta^2 x_\mu x_\nu . \tag{1.8}$$

Unfortunately, finding the contravariant form of the metric, $g^{\mu\nu}$, is non-trivial since the matrix inverse of (1.8) cannot easily be computed for generic $p(\boldsymbol{x})$. ∎

1.4 Affine connection on a statistical manifold

In this section we shall introduce a family of affine connections based on the 1-representation of vectors on a statistical manifold. These connections have been named α-connections by Amari in [1].

As shown in appendix A.5, an affine connection provides a means of comparing vectors at nearby points, thus providing a non-local notion of parallelism. Through the related notion of affine geodesic, it also provides a notion of straight line between two points in a manifold. We noted that an affine connection is defined by a mapping from the tangent space at P$'$ with coordinates $\boldsymbol{\theta} + \delta\boldsymbol{\theta}$ to the tangent space at P with coordinates $\boldsymbol{\theta}$.

In the 1-representation, the former space, $T_{\mathrm{P}'}^{(1)}$, is spanned by the basis

$$\partial_\mu \ell(\boldsymbol{\theta}+\delta\boldsymbol{\theta}) = \partial_\mu \ell(\boldsymbol{\theta}) + \delta\theta^\nu \partial_\mu \partial_\nu \ell(\boldsymbol{\theta}) + O(\delta\theta^\nu \delta\theta^\mu). \tag{1.9}$$

In trying to construct a connection, we seek a mapping from these functions to some functions in $T_{\mathrm{P}}^{(1)}$, which is spanned by

$$\partial_\mu \ell(\boldsymbol{\theta}). \tag{1.10}$$

It is quite clear that the functions (1.9) cannot be expressed as linear combinations of (1.10), since the expectation values of the latter vanish, while $E\left[\partial_\mu \partial_\nu \ell(\boldsymbol{\theta})\right]$ certainly does not vanish. There are several ways to cure this problem.

One is to add $g_{\mu\nu}\delta\theta^\nu$ to (1.9), yielding

$$\partial_\mu \ell(\boldsymbol{\theta}) + \left\{\partial_\mu \partial_\nu \ell(\boldsymbol{\theta}) + g_{\mu\nu}\right\} \delta\theta^\nu, \tag{1.11}$$

which has vanishing expectation value, but still does not yet necessarily belong to $T_{\mathrm{P}}^{(1)}$. This we repair by bluntly projecting it down to $T_{\mathrm{P}}^{(1)}$. Since the projection of any random variable $A_{\boldsymbol{\theta}}(\cdot)$ down to $T_{\mathrm{P}}^{(1)}$ is given by

$$A'_{\boldsymbol{\theta}}(\boldsymbol{x}) = E\left[A_{\boldsymbol{\theta}}(\boldsymbol{x})\partial_\nu \ell(\boldsymbol{x};\boldsymbol{\theta})\right]\, g^{\mu\nu}(\boldsymbol{\theta})\partial_\mu \ell(\boldsymbol{x};\boldsymbol{\theta}),^{1)}$$

we find that

$$\phi : \partial_\mu \ell(\boldsymbol{\theta}+\delta\boldsymbol{\theta}) \mapsto \partial_\mu \ell(\boldsymbol{\theta}) + E\left[\left\{\partial_\mu \partial_\nu \ell(\boldsymbol{\theta}) + g_{\mu\nu}(\boldsymbol{\theta})\right\} \delta\theta^\nu \partial_\rho \ell(\boldsymbol{\theta})\right] g^{\rho\lambda}(\boldsymbol{\theta})\partial_\lambda \ell(\boldsymbol{\theta}) \tag{1.12}$$

is a suitable projection from $T_{\mathrm{P}'}^{(1)}$ into $T_{\mathrm{P}}^{(1)}$. As $g_{\mu\nu}$ is an expectation value itself, the expectation value of the second term in braces factorizes, and since the expectation value of $\partial_\lambda \ell$ is zero, this term actually vanishes. The resulting connection therefore is

$$\Gamma_{\mu\nu}{}^{\lambda} = E\left[\partial_\mu \partial_\nu \ell g^{\rho\lambda} \partial_\rho \ell\right] = E\left[\partial_\mu \partial_\nu \ell \partial_\rho \ell\right] g^{\rho\lambda}. \tag{1.13}$$

However, there are other possibilities: (1.12) is not the only projection of $T_{\mathrm{P}'}^{(1)}$ to $T_{\mathrm{P}}^{(1)}$. Since the expectation value of the combination

$$\partial_\mu \partial_\nu \ell + \partial_\mu \ell \partial_\nu \ell$$

vanishes (see (1.4)), we may consider

$$\partial_\mu \ell(\boldsymbol{\theta}) + \left\{\partial_\mu \partial_\nu \ell(\boldsymbol{\theta}) + \partial_\mu \ell \partial_\nu \ell\right\} \delta\theta^\nu \tag{1.14}$$

as a replacement for (1.11)[2)]. Projecting this down to $T_{\mathrm{P}}^{(1)}$ yields

$$\Gamma_{\mu\nu}{}^{\lambda} = E\left[\left\{\partial_\mu \partial_\nu \ell + \partial_\mu \ell \partial_\nu \ell\right\} g^{\rho\lambda} \partial_\rho \ell\right] = E\left[\partial_\mu \partial_\nu \ell \partial_\rho \ell + \partial_\mu \ell \partial_\nu \ell \partial_\rho \ell\right] g^{\rho\lambda}. \tag{1.15}$$

Obviously, any linear combination of (1.11) and (1.14) also has vanishing expectation value, so we have in fact found an entire family of connections, which are called the α-connections:

[1)] cf. $X = X^\mu \hat{e}_\mu \Leftrightarrow X^\mu = \langle X\,,\, \hat{e}_\nu \rangle\, g^{\mu\nu}$.

[2)] Note the subtle difference between adding $g_{\mu\nu}$ and adding $\partial_\mu \ell \partial_\nu \ell$ without taking expectation value straightaway.

Definition: α-connection

The α-connection on a statistical manifold is defined as

$$\Gamma^{(\alpha)\lambda}_{\mu\nu} = E\left[\partial_\mu\partial_\nu\ell\,\partial_\rho\ell + \frac{1-\alpha}{2}\partial_\mu\ell\,\partial_\nu\ell\,\partial_\rho\ell\right] g^{\rho\lambda}. \tag{1.16}$$

As an aside, we show that the metric connection is the same as the 0-connection:

$$\begin{aligned}
\Gamma^{(\text{metric})}_{\mu\nu\rho} &= \frac{1}{2}\left\{\partial_\mu g_{\nu\rho} + \partial_\nu g_{\mu\rho} - \partial_\rho g_{\mu\nu}\right\} \\
&= \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\ \left\{\partial_\mu\left[p\partial_\nu\ell\partial_\rho\ell\right] + \partial_\nu\left[p\partial_\mu\ell\partial_\rho\ell\right] - \partial_\rho\left[p\partial_\mu\ell\partial_\nu\ell\right]\right\} \\
&= \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\ p\left\{\frac{1}{p}\partial_\mu\left[p\partial_\nu\ell\partial_\rho\ell\right] + \frac{1}{p}\partial_\nu\left[p\partial_\mu\ell\partial_\rho\ell\right] - \frac{1}{p}\partial_\rho\left[p\partial_\mu\ell\partial_\nu\ell\right]\right\} \\
&= \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\ p\left\{\partial_\mu\ell\partial_\nu\ell\partial_\rho\ell + 2\partial_\mu\partial_\nu\ell\,\partial_\rho\ell\right\} \\
&= \frac{1}{2}E\left[\partial_\mu\ell\partial_\nu\ell\partial_\rho\ell\right] + E\left[\partial_\mu\partial_\nu\ell\,\partial_\rho\ell\right] = \Gamma^{(0)}_{\mu\nu\rho}.
\end{aligned}$$

We may therefore write

$$\Gamma^{(\alpha)}_{\mu\nu\rho} = \Gamma^{(\text{metric})}_{\mu\nu\rho} + \alpha T_{\mu\nu\rho}, \tag{1.17}$$

where

$$T_{\mu\nu\rho} \equiv -\frac{1}{2}E\left[\partial_\mu\ell\partial_\nu\ell\partial_\rho\ell\right].$$

We round this of chapter with another example.

Example: Continuing our previous example, we may compute the α-connection for a single binary neuron: differentiating (1.7) again, we find:

$$\partial_\mu\partial_\nu\ell = -\left[1-\tanh^2\beta h(\boldsymbol{x})\right]x_\mu x_\nu$$

after some straightforward calculations. Similar calculations yield

$$\partial_\mu\partial_\nu\ell\partial_\rho\ell = -\frac{y}{1+y\tanh\beta h(\boldsymbol{x})}\left[1-\tanh^2\beta h(\boldsymbol{x})\right]^2\beta^3 x_\mu x_\nu x_\rho,$$

and

$$\partial_\mu\ell\partial_\nu\ell\partial_\rho\ell = \frac{y}{\left[1+y\tanh\beta h(\boldsymbol{x})\right]^3}\left[1-\tanh^2\beta h(\boldsymbol{x})\right]^3\beta^3 x_\mu x_\nu x_\rho.$$

Taking expectation values gives us

$$E\left[\partial_\mu\partial_\nu\ell\partial_\rho\ell\right] = 0,$$

and

$$E\left[\partial_\mu\ell\partial_\nu\ell\partial_\rho\ell\right] = \sum_{\boldsymbol{x}} p(\boldsymbol{x})\tanh\beta h(\boldsymbol{x})\left[1-\tanh^2\beta h(\boldsymbol{x})\right]\beta^3 x_\mu x_\nu x_\rho.$$

Therefore,

$$\Gamma^{(\alpha)}_{\mu\nu\rho} = \frac{1-\alpha}{2}\sum_{\boldsymbol{x}} p(\boldsymbol{x})\tanh\beta h(\boldsymbol{x})\left[1-\tanh^2\beta h(\boldsymbol{x})\right]\beta^3 x_\mu x_\nu x_\rho. \tag{1.18}$$

Since we haven't been able to compute $g^{\mu\nu}$, we cannot give $\Gamma^{(\alpha)\lambda}_{\mu\nu}$ either. ∎

CHAPTER 2

DUALITY IN DIFFERENTIAL GEOMETRY

In this chapter we shall investigate the notions of dual connections and dual coordinate systems. The link with chapter 1 will be made at the very end, when we discover the duality properties of the α-connections.

The key result of this chapter will be the introduction of 'divergence', which we shall find to be a measure of the difference between two distributions. For computational purposes the divergence as the important advantage over the Riemannian distance, that it may be calculated without integration along a geodesic.

Most of the ideas presented in this chapter may also be found in [1].

2.1 Dual connections

Consider vector fields X and Y that are defined on a curve $\gamma \subset \mathcal{M}$ as the parallel transports of the vectors $X(0) \equiv X|_{\gamma(0)}$ and $Y(0) \equiv Y|_{\gamma(0)}$ relative to some affine connection $\Gamma_{\mu\nu\rho}$. Parallel transport does not necessarily preserve inner product, that is

$$\left\langle X(t)\,,\,Y(t)\right\rangle \neq \left\langle X(0)\,,\,Y(0)\right\rangle$$

in general. A connection for which the inner product between any pair of vectors is preserved across parallel transport is called a *metric* connection.

For non-metric connections, it may be possible to find another connection, say $\Gamma^*_{\mu\nu\rho}$, such that

$$\left\langle X(t)\,,\,Y^*(t)\right\rangle = \left\langle X(0)\,,\,Y(0)\right\rangle, \tag{2.1}$$

where $Y^*(t)$ is the parallel transport of $Y|_{\gamma(0)}$ relative to $\Gamma^*_{\mu\nu\rho}$. If (2.1) holds for any two vectors X and Y, then the connections $\Gamma_{\mu\nu\rho}$ and $\Gamma^*_{\mu\nu\rho}$ are said to be *dual* to each other[1)]. Metric connections may then be called *self-dual*.

Duality of connections may be locally defined as follows:

Definition: Two covariant derivatives ∇ and ∇^* (and the corresponding connections $\Gamma_{\mu\nu\rho}$ and $\Gamma^*_{\mu\nu\rho}$) are said to be *dual* to each other when for any three vector fields X, Y and Z:

$$X^\mu \partial_\mu \left\langle Y\,,\,Z\right\rangle = \left\langle \nabla_X Y\,,\,Z\right\rangle + \left\langle Y\,,\,\nabla^*_X Z\right\rangle. \tag{2.2}$$

[1)]The notion of a dual connection was introduced in a slightly different manner originally: even in the absence of a metric, one might try to find a way to transport a 1-form ω such that $\omega(X)$ remains constant along any curve if X is transported parallel to that curve. It was found that this is indeed possible, and the connection that transports ω as required was named the dual to the connection used for the parallel transport of X.

Theorem: There exists a dual to *any* affine connection. This dual is unique, and the dual of the dual is the connection itself.

Proof: Substituting $X = \hat{e}_\mu$, $Y = \hat{e}_\nu$ and $Z = \hat{e}_\rho$ in the definition, we find

$$\partial_\mu g_{\nu\rho} = \Gamma_{\mu\nu\rho} + \Gamma^*_{\mu\rho\nu}, \tag{2.3}$$

from which all three claims follow directly. □

Note that the essential step in this proof is: if (2.3) holds for two connections $\Gamma_{\mu\nu\rho}$ and $\Gamma^*_{\mu\rho\nu}$, then they are indeed dual in the sense of (2.2). This equivalence between (2.3) and (2.2) will be important in the proof of the next lemma as well.

Lemma: Any pair of connections ∇ and ∇^* satisfying (2.1) for any vector fields X and Y^* and any curve γ also satisfies the differential definition of duality, and vice versa.

Proof:

$\Rightarrow$ Start out with (2.1) for infinitesimal t, so that we may expand:

$$\begin{aligned} X^\mu(t) &= X^\mu|_{t=0} - t\left[\Gamma_{\lambda\nu}{}^\mu \dot{\theta}^\lambda X^\nu\right]_{t=0} + O(t^2), \\ Y^{*\mu}(t) &= Y^\mu|_{t=0} - t\left[\Gamma^*_{\lambda\nu}{}^\mu \dot{\theta}^\lambda Y^\nu\right]_{t=0} + O(t^2), \\ g_{\mu\nu}(t) &= g_{\mu\nu}\big|_{t=0} + t\left[\partial_\lambda g_{\mu\nu}\dot{\theta}^\lambda\right]_{t=0} + O(t^2), \end{aligned}$$

where $\dot{}$ denotes differentation with respect to t.

Inserting these Taylor expansions into (2.1), we find

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\left[X^\mu(t)Y^{*\nu}(t)g_{\mu\nu}(t)\right] \\ &= \left[-\Gamma_{\lambda\rho}{}^\mu \dot{\theta}^\lambda X^\rho Y^\nu g_{\mu\nu} - X^\nu \Gamma^*_{\lambda\rho}{}^\nu \dot{\theta}^\lambda Y^\rho g_{\mu\nu} + X^\mu Y^\nu \partial_\lambda g_{\mu\nu}\dot{\theta}^\lambda\right]_{t=0} + O(t) \\ &= \left[\left(-\Gamma_{\lambda\mu\nu} - \Gamma^*_{\lambda\nu\mu} + \partial_\lambda g_{\mu\nu}\right) X^\mu Y^\nu \dot{\theta}^\lambda\right]_{t=0}. \end{aligned}$$

Since this is supposed to hold for any vectors X and Y and any curve γ, we may conclude that the first factor on the right hand side must be zero, proving (2.3), and thereby duality of ∇ and ∇^*.

$\Leftarrow$ We shall show that the left hand side of (2.1) cannot differ from the right hand side if $\Gamma_{\mu\nu\rho}$ and $\Gamma^*_{\mu\nu\rho}$ are dual connections. Let the curve γ be parametrized by $\boldsymbol{\theta} = \boldsymbol{\theta}(t)$, and denote the tangent vector field to γ by $\dot{\boldsymbol{\theta}} \in T(\mathcal{M})$. Since X and Y^* are defined as parallel transports along γ, we have

$$\nabla_{\dot{\boldsymbol{\theta}}} X(t) = \nabla^*_{\dot{\boldsymbol{\theta}}} Y^*(t) = 0.$$

Therefore

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left\langle X(t), Y^*(t)\right\rangle &= \frac{\mathrm{d}\theta^\mu}{\mathrm{d}t}\frac{\partial}{\partial\theta^\mu}\left\langle X(\boldsymbol{\theta}(t)), Y^*(\boldsymbol{\theta}(t))\right\rangle \\ &\equiv \dot{\theta}^\mu \partial_\mu \left\langle X(\boldsymbol{\theta}(t)), Y^*(\boldsymbol{\theta}(t))\right\rangle \\ &= \left\langle \nabla_{\dot{\boldsymbol{\theta}}} X(t), Y^*(t)\right\rangle + \left\langle X(t), \nabla^*_{\dot{\boldsymbol{\theta}}} Y^*(t)\right\rangle \\ &= 0 + 0 = 0. \end{aligned}$$

Taking the integral with respect to t yields (2.1). □

2.2 Dual flatness

A manifold is *flat* with respect to an affine connection when there is a set of coordinates such that $\Gamma_{\mu\nu}{}^{\rho} = 0$. The following theorem links flatness with respect to a connection and flatness with respect to its dual:

Theorem: Dual flatness
When a manifold is flat with respect to an affine connection, it is also flat with respect to its dual.

This theorem follows from the following lemma:

Lemma: When $\Gamma_{\mu\nu}{}^{\rho}$ and $\Gamma^{*}_{\mu\nu}{}^{\rho}$ are dual connections, their curvatures obey the following relationship:

$$R_{\mu\nu\rho\lambda} = -R^{*}_{\mu\nu\lambda\rho}.$$

Proof: From its definition (A.11), we see that it is possible to write the curvature as

$$R_{\mu\nu\rho\lambda} = \left\langle \nabla_{\mu}\nabla_{\nu}\hat{e}_{\rho}\,,\,\hat{e}_{\lambda}\right\rangle - \left(\mu \leftrightarrow \nu\right),$$

since

$$\nabla_{\mu}\left(\Gamma_{\nu\rho}{}^{\lambda}\hat{e}_{\lambda}\right) = \left(\partial_{\mu}\Gamma_{\nu\rho}{}^{\lambda}\right)\hat{e}_{\lambda} + \Gamma_{\nu\rho}{}^{\lambda}\nabla_{\mu}\hat{e}_{\lambda}.$$

Using definition (2.2) twice, we find

$$\begin{aligned}
R_{\mu\nu\rho\lambda} &= \left(\partial_{\mu}\left\langle \nabla_{\nu}\hat{e}_{\rho}\,,\,\hat{e}_{\lambda}\right\rangle - \left\langle \nabla_{\nu}\hat{e}_{\rho}\,,\,\nabla^{*}_{\mu}\hat{e}_{\lambda}\right\rangle\right) - \left(\mu \leftrightarrow \nu\right) \\
&= \left(\partial_{\mu}\partial_{\nu}\left\langle \hat{e}_{\rho}\,,\,\hat{e}_{\lambda}\right\rangle - \partial_{\mu}\left\langle \hat{e}_{\rho}\,,\,\nabla^{*}_{\nu}\hat{e}_{\lambda}\right\rangle - \partial_{\nu}\left\langle \hat{e}_{\rho}\,,\,\nabla^{*}_{\mu}\hat{e}_{\lambda}\right\rangle + \left\langle \hat{e}_{\rho}\,,\,\nabla^{*}_{\nu}\nabla^{*}_{\mu}\hat{e}_{\lambda}\right\rangle\right) - \left(\mu \leftrightarrow \nu\right) \\
&= \left(0 - 0 - 0 + \left\langle \hat{e}_{\rho}\,,\,\nabla^{*}_{\nu}\nabla^{*}_{\mu}\hat{e}_{\lambda}\right\rangle\right) - \left(\mu \leftrightarrow \nu\right) \\
&= \left\langle \nabla^{*}_{\nu}\nabla^{*}_{\mu}\hat{e}_{\lambda}\,,\,\hat{e}_{\rho}\right\rangle - \left(\mu \leftrightarrow \nu\right) = R^{*}_{\nu\mu\lambda\rho} \\
&= -R^{*}_{\mu\nu\lambda\rho}.
\end{aligned}$$

□

2.3 Dual coordinate systems

On a dually flat manifold, there exist two special coordinates systems: the affine flat coordinates for each of the connections. These coordinate systems are related to one another by a duality relation of their own: they are *dual coordinate systems*:

Definition: Dual coordinate systems
Two coordinate systems (θ^{μ}) and $(\tilde{\theta}_{\tilde{\nu}})$ are said to be dual to one another when their coordinate basis vectors satisfy:

$$\left\langle \hat{e}_{\mu}\,,\,\tilde{e}^{\tilde{\nu}}\right\rangle = \delta^{\tilde{\nu}}_{\mu},$$

where $\hat{e}_{\mu}$ and $\tilde{e}^{\tilde{\nu}}$ the coordinate basis vectors for the $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ systems respectively.

Note that we use lower indices to denote the components of $\tilde{\boldsymbol{\theta}}$. At the present stage, this has no other significance than notational convenience: it will, for example, allow us to stick to the usual summation convention.

For (θ^μ) and $(\tilde{\theta}_{\tilde{\nu}})$ to be dual to one another, they need not necessarily be affine coordinate systems. However there is no guarantee for general manifolds that a pair of dual coordinate systems exists.

We shall investigate some of the properties of dual coordinate systems.

2.3.1 Relation of the metrics

We may express $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ in terms of one another:

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\tilde{\boldsymbol{\theta}}), \qquad \tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\boldsymbol{\theta}).$$

The coordinate basis vectors are therefore related by

$$\hat{e}_\mu = \frac{\partial \tilde{\theta}_{\tilde{\nu}}}{\partial \theta^\mu} \tilde{e}^{\tilde{\nu}}, \qquad \tilde{e}^{\tilde{\mu}} = \frac{\partial \theta^\nu}{\partial \tilde{\theta}_{\tilde{\mu}}} \hat{e}_\nu \,^{1)}.$$

Using these relations, we may express the metrics $g_{\mu\nu} = \langle \hat{e}_\mu \, , \hat{e}_\nu \rangle$ and $g^{\tilde{\mu}\tilde{\nu}} = \langle \tilde{e}^{\tilde{\mu}} \, , \tilde{e}^{\tilde{\nu}} \rangle$ induced by the coordinate systems in terms of the Jacobians:

$$g_{\mu\nu} \equiv \langle \hat{e}_\mu \, , \hat{e}_\nu \rangle = \frac{\partial \tilde{\theta}_{\tilde{\nu}}}{\partial \theta^\mu} \langle \tilde{e}^{\tilde{\nu}} \, , \hat{e}_\nu \rangle = \frac{\partial \tilde{\theta}_{\tilde{\nu}}}{\partial \theta^\mu} \delta^{\tilde{\nu}}_\nu, \tag{2.4a}$$

and

$$\tilde{g}^{\tilde{\mu}\tilde{\nu}} \equiv \langle \tilde{e}^{\tilde{\mu}} \, , \tilde{e}^{\tilde{\nu}} \rangle = \frac{\partial \theta^\mu}{\partial \tilde{\theta}_{\tilde{\mu}}} \langle \hat{e}_\mu \, , \tilde{e}^{\tilde{\nu}} \rangle = \frac{\partial \theta^\mu}{\partial \tilde{\theta}_{\tilde{\nu}}} \delta^{\tilde{\mu}}_\mu. \tag{2.4b}$$

Noting that the Jacobians $[J_{\mu\tilde{\nu}}] \equiv \left[\frac{\partial \tilde{\theta}_{\tilde{\nu}}}{\partial \theta^\mu}\right]$ and $[J^{\tilde{\mu}\nu}] \equiv \left[\frac{\partial \theta^\nu}{\partial \tilde{\theta}_{\tilde{\mu}}}\right]$ are each other's matrix inverse, we find that $[\tilde{g}^{\tilde{\nu}\tilde{\mu}}]$ and $[g_{\mu\nu}]$ are also each other's matrix inverse. Since the matrix inverse of $[g_{\mu\nu}]$ is known to be the contravariant form of the metric, $[g^{\mu\nu}]$, we find that $\tilde{g}^{\tilde{\mu}\tilde{\nu}} = \delta^{\tilde{\mu}}_\mu \delta^{\tilde{\nu}}_\nu g^{\mu\nu}$. In fact, this means that any tensor T expressed in $\boldsymbol{\theta}$-coordinates as $T^{\mu_1 \cdots \mu_m}{}_{\nu_1 \cdots \nu_n}$, may be re-expressed in $\tilde{\boldsymbol{\theta}}$-coordinates as

$$T^{\tilde{\mu}_1 \cdots \tilde{\mu}_m}{}_{\tilde{\nu}_1 \cdots \tilde{\nu}_n} = \delta^{\tilde{\mu}_1}_{\mu_1} \ldots \delta^{\tilde{\mu}_m}_{\mu_m} \, \delta^{\nu_1}_{\tilde{\nu}_1} \ldots \delta^{\nu_n}_{\tilde{\nu}_n} \, T^{\mu_1 \cdots \mu_m}{}_{\nu_1 \cdots \nu_n}.$$

At this stage it is obvious that we may as well clean up our notation by dropping the distinction between labels with and without tildes[2)].

The following theorem allows us to find the functional form of $\boldsymbol{\theta} = \boldsymbol{\theta}(\tilde{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

[1)] Again these relations are obvious in the differential operator formalism: they follow directly from $\hat{e}_\mu \equiv \frac{\partial}{\partial \theta^\mu}$ and $\tilde{e}^\nu \equiv \frac{\partial}{\partial \tilde{\theta}_{\tilde{\nu}}}$

[2)] This explains why we chose to put the contravariant labels in the $\tilde{\boldsymbol{\theta}}$ coordinate system downstairs. The transformation between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ coordinates is equal to the identity only when covariant labels are replaced by contravariant ones and vice versa at the same time. The present convention means that we do not have to change upstairs labels into downstairs labels, thus avoiding a lot of confusion.

Theorem: When (θ^μ) and $(\tilde{\theta}_\mu)$ are dual coordinate systems, there exist *potential functions* $\Theta(\boldsymbol{\theta})$ and $\tilde{\Theta}(\tilde{\boldsymbol{\theta}})$ such that

$$\theta^\mu = \tilde{\partial}^\mu \tilde{\Theta}(\tilde{\boldsymbol{\theta}}) \qquad \text{and} \qquad \tilde{\theta}_\mu = \partial_\mu \Theta(\boldsymbol{\theta}). \tag{2.5}$$

It follows that

$$g_{\mu\nu} = \partial_\mu \partial_\nu \Theta(\boldsymbol{\theta}) \qquad \text{and} \qquad \tilde{g}^{\mu\nu} = \tilde{\partial}^\mu \tilde{\partial}^\nu \tilde{\Theta}(\tilde{\boldsymbol{\theta}}). \tag{2.6}$$

Furthermore,

$$\Theta(\boldsymbol{\theta}) + \tilde{\Theta}(\tilde{\boldsymbol{\theta}}) = \theta^\mu \, \tilde{\theta}_\mu. \tag{2.7}$$

Conversely, when a potential function $\Theta(\boldsymbol{\theta})$ exists such that $g_{\mu\nu} = \partial_\mu \partial_\nu \Theta(\boldsymbol{\theta})$, (2.5) yields a coordinate system $(\tilde{\theta}_\mu)$ which will be dual to (θ^μ), and (2.7) may be used to derive the other potential function $\tilde{\Theta}(\tilde{\boldsymbol{\theta}})$.

Proof: Symmetry of the metric $g_{\mu\nu} = \partial_\mu \tilde{\theta}_\nu(\boldsymbol{\theta})$ shows that $\partial_\mu \tilde{\theta}_\nu - \partial_\nu \tilde{\theta}_\mu = 0$, from which we may conclude that, at least locally, $\tilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the derivative of some function, i.e. there exists a function $\Theta(\boldsymbol{\theta})$ such that $\tilde{\theta}_\mu = \partial_\mu \Theta$. (2.6) follows directly from inserting (2.5) into (2.4a) and (2.4b). Finally, (2.7) is a general fact about Legendre transforms.

The other direction is easy: when $g_{\mu\nu} = \partial_\mu \partial_\nu \Theta(\boldsymbol{\theta})$, we see that $(\tilde{\theta}_\mu) \equiv (\partial_\mu \Theta(\boldsymbol{\theta}))$ is dual to $(\boldsymbol{\theta})$ from the fact that

$$\tilde{e}^\nu = \left(\left[\frac{\partial \tilde{\theta}}{\partial \theta} \right]^{-1} \right)^{\nu\mu} \hat{e}_\mu = ([g_{\mu\nu}]^{-1})^{\mu\nu} \hat{e}_\mu,$$

whence $\left\langle \tilde{e}^\nu, \hat{e}_\mu \right\rangle = \delta^\nu_\mu$, proving duality. □

Example: Binary neuron

Recall that for a binary neuron we obtained

$$g_{\mu\nu}(\boldsymbol{J}) = \sum_{\boldsymbol{x} \in \{-1,1\}^N} p(\boldsymbol{x}) \left[1 - \tanh^2(\beta \boldsymbol{J} \cdot \boldsymbol{x}) \right] \beta^2 x_\mu x_\nu.$$

It is not difficult to integrate this expression and find that

$$\Theta(\boldsymbol{J}) = \sum_{\boldsymbol{x} \in \{1,1\}^N} p(\boldsymbol{x}) \ln \cosh(\beta \boldsymbol{J} \cdot \boldsymbol{x}),$$

whence we compute

$$\tilde{J}_\mu = \sum_{\boldsymbol{x} \in \{1,1\}^N} p(\boldsymbol{x}) \tanh(\beta \boldsymbol{J} \cdot \boldsymbol{x}) \beta x_\mu.$$

In order to find $\tilde{\Theta}(\tilde{\boldsymbol{J}})$, we would have to solve this equation for $\boldsymbol{J}$. This is not easy in general, but we may note that in the case of a single input and no threshold the equations trivialize: we get

$$g_{11} = \sum_{x \in \{-1,1\}} p(x) \left[1 - \tanh^2(\beta J x) \right] \beta^2 x^2.$$

From the antisymmetry of tanh and noting that $x^2 \equiv 1$ this simplifies to

$$g_{11} = \beta^2 \left[1 - \tanh^2(\beta J)\right].$$

We then find $\Theta(J) = \ln\cosh(\beta J)$ and

$$\tilde{J} = \beta \tanh(\beta J).$$

Inverting the metric becomes easy too:

$$g^{11} = \frac{1}{\beta^2 \left[1 - \tanh^2(\beta J)\right]} = \frac{1}{\beta^2 - \tilde{J}^2}.$$

More interestingly, some headway can be made in the limit $N \to \infty$, if we assume the inputs to be uniformly distributed, $p(\boldsymbol{x}) = 2^{-N}$. We may write:

$$\tilde{J}_\mu = \beta \left\langle x_\mu \tanh(\beta J^\nu x_\nu) \right\rangle = \beta \left\langle \tanh(\beta J^\nu x_\nu x_\mu) \right\rangle = \beta \left\langle \tanh\beta(J^\mu + \sum_{\nu\neq\mu} J^\nu x_\nu) \right\rangle,$$

where we performed a gauge transformation $x_\nu \to x_\nu x_\mu$ in the last step.

In the limit $N \to \infty$ the second term, $\sum_{\nu\neq\mu} J^\nu x_\nu$, approaches a Gaussian distribution with zero mean and variance:

$$\sigma^2 = \left\langle (\sum_{\nu\neq\mu} J^\nu x_\nu)^2 \right\rangle = 2^{-N} \sum_{x\in\{-1,1\}} \sum_{\nu,\rho\neq\mu} J^\nu J^\rho x_\nu x_\rho = \sum_{\nu\neq\mu} J^\nu J^\nu \equiv |\boldsymbol{J}|^2 - (J^\mu)^2,$$

where no summation over μ is implied and $|\boldsymbol{J}|$ is the Euclidean (!) length of $\boldsymbol{J}$. We may therefore compute $\tilde{J}_\mu$ by

$$\tilde{J}_\mu = \beta \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}z^2} \tanh\beta \left(J^\mu + z\sqrt{|\boldsymbol{J}|^2 - (J^\mu)^2} \right). \tag{2.8}$$

Assuming that $|\boldsymbol{J}|$ is of order 1, and that for each of the components $|J^\mu|$ is of order $1/\sqrt{N}$, we may expand:

$$\begin{aligned} \tanh\beta \left(J^\mu + z\sqrt{|\boldsymbol{J}|^2 - (J^\mu)^2} \right) &= \tanh\beta \left(J^\mu + z|\boldsymbol{J}| + O(N^{-1}) \right) \\ &= \tanh(\beta z|\boldsymbol{J}|) + \beta J^\mu \left[1 - \tanh^2(\beta z|\boldsymbol{J}|)\right] + O(N^{-1}). \end{aligned}$$

Inserting this into (2.8), the first term vanishes and we are left with:

$$\tilde{J}_\mu = \beta^2 \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}z^2} J^\mu \left[1 - \tanh^2(\beta z|\boldsymbol{J}|)\right].$$

While this integral cannot be computed analytically, numerical methods may be employed, or one could expand in one of the limits $\beta \to 0$ or — with some more care — $\beta \to \infty$. We shall not pursue these possibilities here. ∎

2.3.2 Dual coordinate systems in dually flat manifolds

The following theorem states that a pair of dual coordinate systems exists in a dually flat manifold:

Theorem: When a manifold $\mathcal{M}$ is flat with respect to a dual pair of torsion-free connections ∇ and ∇^*, there is a pair of dual coordinate systems (θ^μ) and $(\tilde{\theta}_\mu)$, such that (θ^μ) is ∇-affine, and $(\tilde{\theta}_\mu)$ is ∇^*-affine.

Proof: ∇-flatness allows us to introduce a coordinate system (θ^μ) in which $\Gamma_{\mu\nu\rho} = 0$. According to (2.3) this means that $\Gamma^*_{\mu\nu\rho} = -\partial_\mu g_{\nu\rho}$ (still in $\boldsymbol{\theta}$ coordinates). Since we assumed that ∇^* is torsion free, we have $\Gamma^*_{\mu\nu\rho} = \Gamma^*_{\nu\mu\rho}$, and therefore $\partial_\mu g_{\nu\rho} = \partial_\nu g_{\mu\rho}$. Combining this with the fact that $g_{\mu\nu} = g_{\nu\mu}$, we may conclude that (again, at least locally) a potential function Θ exists such that $g_{\mu\nu} = \partial_\mu \partial_\nu \Theta$. This allows us to introduce a coordinate system $(\tilde{\theta}_\mu)$ dual to (θ^μ) defined by $\tilde{\theta}_\nu = \partial_\nu \Theta(\boldsymbol{\theta})$.

In order to show that $(\tilde{\theta}_\mu)$ is a ∇^*-affine coordinate system as claimed, we note that for any μ:

$$\partial_\mu \left\langle \hat{e}_\nu \,,\, \tilde{e}^\rho \right\rangle = \partial_\mu \delta^\rho_\nu = 0,$$

since $(\tilde{\theta}_\mu)$ is dual to (θ^μ). On the other hand, (2.2) shows that

$$\begin{aligned}
\partial_\mu \left\langle \hat{e}_\nu \,,\, \tilde{e}^\rho \right\rangle &= \left\langle \nabla_{\hat{e}_\mu} \hat{e}_\nu \,,\, \tilde{e}^\rho \right\rangle + \left\langle \hat{e}_\nu \,,\, \nabla^*_{\hat{e}_\mu} \tilde{e}^\rho \right\rangle \\
&= g^{\rho\lambda} \left\langle \nabla_{\hat{e}_\mu} \hat{e}_\nu \,,\, \hat{e}_\lambda \right\rangle + g_{\nu\lambda} g_{\mu\sigma} \left\langle \tilde{e}^\lambda \,,\, \nabla^*_{\tilde{e}^\sigma} \tilde{e}^\rho \right\rangle \\
&\equiv g^{\rho\lambda} \Gamma_{\mu\nu\lambda} + g_{\nu\lambda} g_{\mu\sigma} (\Gamma^*)^{\sigma\rho\lambda},
\end{aligned}$$

where $(\Gamma^*)^{\sigma\rho\lambda}$ is the connection corresponding to ∇^*[1)].

Since both the left-hand side and the first term on the right are zero, we conclude that $(\Gamma^*)^{\sigma\rho\lambda} = 0$, proving that $(\tilde{\theta}_\mu)$ is a ∇^*-affine coordinate system. □

2.4 Divergence

On a manifold with dual connections we define the divergence between two points as

$$D(\mathrm{P},\mathrm{Q}) = \Theta(\theta_\mathrm{P}) + \tilde{\Theta}(\tilde{\theta}_\mathrm{Q}) - \theta^\mu_\mathrm{P} \tilde{\theta}_{\mathrm{Q},\mu}. \tag{2.9}$$

At first sight this definition may seem meaningless, but in fact the divergence has rather nice properties: it behaves very much like the square of a distance, and it is obviously very easy to compute: one does not have to evaluate integrals as in the calculation of the Riemannian distance on a curved manifold.

More specifically, the properties of the divergence can be stated as follows:

1. $D(\mathrm{P},\mathrm{Q}) \geq 0$, with equality iff $\mathrm{P} = \mathrm{Q}$;[2)]

[1)] Note that for $(\Gamma^*)^{\mu\nu\rho}$ we need not make the distinction between the covariant form in $\tilde{\boldsymbol{\theta}}$ coordinates and the contravariant form in $\boldsymbol{\theta}$ coordinates, just as for tensors.

[2)] Assuming that the metric is positive definite, which the Fisher information certainly is.

2. $\left.\frac{\partial}{\partial\theta^\mu_{\rm P}} D({\rm P},{\rm Q})\right|_{\rm P=Q} = \left.\frac{\partial}{\partial\theta^\mu_{\rm Q}} D({\rm P},{\rm Q})\right|_{\rm P=Q} = 0$;

3. $\left.\frac{\partial}{\partial\theta^\mu_{\rm P}}\frac{\partial}{\partial\theta^\nu_{\rm P}} D({\rm P},{\rm Q})\right|_{\rm P=Q} = g_{\mu\nu}({\rm P})$; (in fact $\frac{\partial}{\partial\theta^\mu_{\rm P}}\frac{\partial}{\partial\theta^\nu_{\rm P}} D({\rm P},{\rm Q}) = g_{\mu\nu}({\rm P})$ for any P, Q);

4. Given three points P, Q and R, then

$$D({\rm P},{\rm R}) \gtrless D({\rm P},{\rm Q}) + D({\rm Q},{\rm R})$$

if the angle between the tangent vectors at Q of the ∇-geodesic joining P and Q, and the ∇^*-geodesic joining Q and R is greater than, equal to, or less than 90°. (This angle is labelled φ in the picture below.)

Properties 2 and 3 follow directly from differentiating the definition. Property 1 then follows from these by noting that $D({\rm P},{\rm Q})$ is strictly convex in $\boldsymbol{\theta}_{\rm Q} - \boldsymbol{\theta}_{\rm P}$, since the metric is strictly positive definite.

Property 4 – which can be viewed as a generalized Pythagoras law – can be proved as follows: Let $\gamma_{\rm PQ}$ be the ∇-geodesic joining P and Q, and $\gamma_{\rm QR}$ the ∇^*-geodesic joining Q and R. Being geodesics, these curves can be written in terms of the affine coordinates θ and $\tilde\theta$ as follows:

$$\gamma_{\rm PQ} : t \mapsto \boldsymbol{\theta}_{\rm P} + (\boldsymbol{\theta}_{\rm Q} - \boldsymbol{\theta}_{\rm P})\, t$$

and

$$\gamma_{\rm QR} : t \mapsto \tilde{\boldsymbol{\theta}}_{\rm Q} + (\tilde{\boldsymbol{\theta}}_{\rm R} - \tilde{\boldsymbol{\theta}}_{\rm Q})\, t.$$

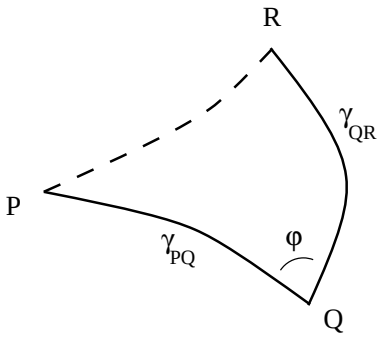

Figure 2.1: *The extension of Pythagoras' law.*

By definition, we have

$$D({\rm P},{\rm R}) = \Theta(\boldsymbol{\theta}_{\rm P}) + \tilde\Theta(\tilde{\boldsymbol{\theta}}_{\rm R}) - \theta^\mu_{\rm P}\tilde\theta_{{\rm R},\mu}.$$

On the other hand,

$$D({\rm P},{\rm Q}) + D({\rm Q},{\rm R}) = \Theta(\boldsymbol{\theta}_{\rm P}) + \tilde\Theta(\tilde{\boldsymbol{\theta}}_{\rm Q}) - \theta^\mu_{\rm P}\tilde\theta_{{\rm Q},\mu} + \Theta(\boldsymbol{\theta}_{\rm Q}) + \tilde\Theta(\tilde{\boldsymbol{\theta}}_{\rm R}) - \theta^\mu_{\rm Q}\tilde\theta_{{\rm R},\mu}.$$

Inserting (2.7) and collecting terms this can be rewritten as:

$$\begin{aligned} D({\rm P},{\rm Q}) + D({\rm Q},{\rm R}) &= D({\rm P},{\rm R}) + \theta^\mu_{\rm Q}\tilde\theta_{{\rm Q},\mu} - \theta^\mu_{\rm P}\tilde\theta_{{\rm Q},\mu} - \theta^\mu_{\rm Q}\tilde\theta_{{\rm R},\mu} + \theta^\mu_{\rm P}\tilde\theta_{{\rm R},\mu} \\ &= D({\rm P},{\rm R}) - (\theta^\mu_{\rm Q} - \theta^\mu_{\rm P})(\tilde\theta_{{\rm R},\mu} - \tilde\theta_{{\rm Q},\mu}) \\ &= D({\rm P},{\rm R}) - \left\langle \dot\gamma_{\rm PQ}\,,\, \dot\gamma_{\rm QR} \right\rangle\Big|_{\rm Q} \\ &= D({\rm P},{\rm R}) - \|\dot\gamma_{\rm PQ}\|\,\|\dot\gamma_{\rm QR}\|\cos(\pi-\varphi), \end{aligned}$$

proving the statement.

There is an intimate relation between the divergence just defined and the Kullback-Leibler distance $D(p\|q) = \int_X {\rm d}\boldsymbol{x}\; p(\boldsymbol{x}) \log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$: locally they are equivalent, and for some special families of distributions one may show that they are globally equal.

2.5 Duality of α- and $-\alpha$-connections

We end this chapter by establishing a link between the dualities found above in differential geometry and the information geometry introduced in the previous chapter. This link consists of the following:

Theorem: The α- and $-\alpha$-connections are dual to one another.

Proof: From (1.17) we find that

$$\Gamma^{(\alpha)}_{\mu\nu\rho} + \Gamma^{(-\alpha)}_{\mu\rho\nu} = \Gamma^{(\text{metric})}_{\mu\nu\rho} + \alpha T_{\mu\nu\rho} + \Gamma^{(\text{metric})}_{\mu\rho\nu} - \alpha T_{\mu\rho\nu}.$$

Since $T_{\mu\nu\rho}$ is fully symmetric in its indices, the two terms that involve α cancel. The metric connection being self dual by construction, the remaining terms add up to $\partial_\mu g_{\nu\rho}$ (using (2.3)). We have thus established that

$$\Gamma^{(\alpha)}_{\mu\nu\rho} + \Gamma^{(-\alpha)}_{\mu\rho\nu} = \partial_\mu g_{\nu\rho}.$$

According to (2.3), this means that $\Gamma^{(\alpha)}_{\mu\nu\rho}$ and $\Gamma^{(-\alpha)}_{\mu\nu\rho}$ are dual to one another. □

Corollary: α-flatness implies $-\alpha$-flatness.

CHAPTER 3

OPTIMIZATION

3.1 Minimizing a scalar function: Gradient descent

Minimizing a scalar function is a very common task. For example in neural networks one often wishes to find the point in weight space where the *generalization error* is minimal: In a feed forward network with n inputs $\boldsymbol{x}$ and m outputs $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{x};\boldsymbol{J})$, ($\boldsymbol{J}$ contains all the connection weights defining the network), the generalization error is given by

$$\varepsilon_g(\boldsymbol{J}) = \left\langle \overline{\|\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x};\boldsymbol{J})\|^2} \right\rangle,$$

where $\overline{\cdots}$ indicates averaging over the noise in $\boldsymbol{f}$, $\langle \cdots \rangle$ indicates averaging over the inputs, and $\boldsymbol{T}(\boldsymbol{x})$ is the function that the network is supposed to learn.

In many applications, people use the following *gradient descent learning rule* (in which η is called the *learning rate*):

$$J^\mu \to J^\mu - \eta \frac{\partial}{\partial J^\mu} \varepsilon_g(\boldsymbol{J}).$$

From our study of differential geometry we see immediately that something funny is going on here: in the first term on the right, μ appears as an upstairs index, while in the second term it appears downstairs. Thus the subtraction doesn't yield a proper vector.

The way to fix this problem is obvious: use the inverse metric to raise the index on the second term. This leads us to the following corrected learning rule

$$J^\mu \to J^\mu - \eta g^{\mu\nu} \frac{\partial}{\partial J^\nu} \varepsilon_g(\boldsymbol{J}).$$

Alternatively, this result can be obtained from first principles, by taking a step back, and considering what we are really trying to achieve by gradient descent: we wish to find a $\delta\boldsymbol{J}$ that maximizes

$$-\frac{\varepsilon_g(\boldsymbol{J}+\delta\boldsymbol{J}) - \varepsilon_g(\boldsymbol{J})}{d(\boldsymbol{J}+\delta\boldsymbol{J},\boldsymbol{J})}. \tag{3.1}$$

For infinitesimal $\delta\boldsymbol{J}$ appendix A.4 tells us that

$$d(\boldsymbol{J}+\delta\boldsymbol{J},\boldsymbol{J}) = \sqrt{g_{\mu\nu}\delta J^\mu \delta J^\nu}.$$

Inserting this into (3.1) and extremizing the result with respect to $\delta\boldsymbol{J}$ shows that δJ^μ should be taken to be a scalar multiple of $g^{\mu\nu}\frac{\partial}{\partial J^\nu}\varepsilon_g(\boldsymbol{J})$ as claimed.

A number of studies [4, 5, 6] indicate that this modified learning rule converges much more rapidly than the original, and in particular, that many of the plateaus phases encountered in flat gradient descent are avoided or substantially reduced when one acknowledges that the weight space is not in fact flat.

3.2 Minimizing the divergence

Another important optimalization problem is to find a distribution that approximates another, given distribution as closely as possible under some constraints. An example would be where we wish to find a weight vector for a stochastical neural network, such that it most closely approximates a given distribution, which need not be exactly representable by the network.

Mathematically, this translates into the following: given a manifold $\mathcal{M}$ and a submanifold $\mathcal{S}$, and a point P in $\mathcal{M}$, find the point P$'$ in $\mathcal{S}$ that is nearest to $\mathcal{M}$. Instead of minimizing the Riemann distance — which would involve integrating the metric along the metric geodesic between P and P$'$ at every iteration step — we aim to minimize the divergence introduced in chapter 2.

When $\mathcal{M}$ is flat under the dual connections $\Gamma_{\mu\nu\rho}$ and $\Gamma^*_{\mu\nu\rho}$, and $\mathcal{S}$ is convex[1)] with respect to $\Gamma^*_{\mu\nu\rho}$, this problem may be solved as follows:

- Introduce affine coordinates $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ on $\mathcal{M}$.
- Introduce any convenient coordinate system $\boldsymbol{\vartheta}$ on $\mathcal{S}$.
- The task then reduces to minimizing

$$D(\mathrm{P},\mathrm{P}') \equiv D(\boldsymbol{\theta}_{\mathrm{P}}, \tilde{\boldsymbol{\theta}}(\boldsymbol{\vartheta}_{\mathrm{P}'})) \tag{3.2}$$

  with respect to $\boldsymbol{\vartheta}$. Since (3.2) defines a scalar function on $\mathcal{S}$, this is easy: simply iterate

$$\vartheta^\alpha \to \vartheta^\alpha - \eta g_{\mathcal{S}}^{\alpha\beta} \frac{\partial}{\partial \vartheta^\beta} D(\boldsymbol{\theta}_{\mathrm{P}}, \tilde{\boldsymbol{\theta}}(\boldsymbol{\vartheta})),$$

  where $g_{\mathcal{S}}^{\alpha\beta}$ is the metric on $\mathcal{S}$ induced by the metric on $\mathcal{M}$.

When $\mathcal{S}$ is not convex, the procedure may still be used, but the minimum reached does not necessarily correspond to the global minimum.

3.2.1 Between two constrained points

Sometimes we may not wish to approximate a specific distribution in $\mathcal{M}$ by a distribution in $\mathcal{S}$, but rather to find a distribution in $\mathcal{S}$ that is closest to any distribution in $\mathcal{M}$ that satisfies some property. Mathematically this translates to the possibility that there may be a partitioning of $\mathcal{M}$ such that all distributions in one set are in some sense equivalent. This situation typically occurs in recurrent neural networks with hidden layers: only the states of non-hidden neurons are relevant from an external point of view.

Suppose then that we wish to find a distribution in $\mathcal{S}$ that is closest to any distribution in another submanifold $\mathcal{N}$ of $\mathcal{M}$. In other words, we wish to find the points $\mathrm{P} \in \mathcal{N}$ and $\mathrm{P}' \in \mathcal{S}$ that minimize $D(\mathrm{P},\mathrm{P}') = D(\boldsymbol{\theta}(\boldsymbol{\tau}_{\mathrm{P}}), \tilde{\boldsymbol{\theta}}(\boldsymbol{\vartheta}_{\mathrm{P}'}))$.

Introducing coordinates $\boldsymbol{\tau} = (\tau^a)$ on $\mathcal{N}$, this problem may again be solved by gradient descent[2)]: just iterate

$$(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \to (\boldsymbol{\vartheta}', \boldsymbol{\tau}'),$$

[1)] A submanifold $\mathcal{S}$ is said to be convex with respect to ∇ when the ∇-geodesic between any two points in $\mathcal{S}$ lies entirely within $\mathcal{S}$.

[2)] The solution presented here is based on [7]

where

$$\vartheta'^{\alpha} = \vartheta^{\alpha} - \eta g_{\mathcal{S}}^{\alpha\beta} \frac{\partial}{\partial \vartheta^{\beta}} D(\boldsymbol{\theta}(\boldsymbol{\tau}), \tilde{\boldsymbol{\theta}}(\boldsymbol{\vartheta})),$$

and

$$\tau'^{a} = \tau^{a} - \eta g_{\mathcal{N}}^{ab} \frac{\partial}{\partial \tau^{b}} D(\boldsymbol{\theta}(\boldsymbol{\tau}), \tilde{\boldsymbol{\theta}}(\boldsymbol{\vartheta})),$$

in which $g_{\mathcal{N}}^{ab}$ is the metric on $\mathcal{N}$ induced by the metric on $\mathcal{M}$.

Again this procedure converges to a unique global minimum if $\mathcal{S}$ is convex with respect to $\Gamma^{*}_{\mu\nu\rho}$ and $\mathcal{N}$ is convex with respect to $\Gamma_{\mu\nu\rho}$, but the procedure may be used even in the general case if one accepts that it may not find the global minimum.

3.2.2 Application: Recurrent neural networks

As a specific example, consider a stochastic recurrent neural network with symmetric interaction matrix and without self-interactions[1)]. If the network consists of N cells, we may represent the states of these cells as an N-dimensional vector $\boldsymbol{\sigma} \in \{-1, 1\}^N$. The sequential evolution rule picks one of these cells at random at every timestep and updates its state according to

$$\sigma_i(t+1) = \operatorname{sgn}\left(\tanh(\beta h_i(\boldsymbol{\sigma}(t))) + z_i(t)\right),$$

where $z_i(t)$ is a uniform random variable taking values in $[-1, 1]$ and

$$h_i(\boldsymbol{\sigma}) = \sum_{j=1}^{N} w_{ij}\sigma_j + \theta_i.$$

Here w_{ij} is the interaction matrix and θ_i are the biases.

Iterating this step yields a Markov process, which can be written as

$$P[\boldsymbol{\sigma}(t+1) = \boldsymbol{\sigma}] = \sum_{\boldsymbol{\sigma}'} W[\boldsymbol{\sigma}; \boldsymbol{\sigma}'] P[\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}'],$$

where

$$W[\boldsymbol{\sigma}; \boldsymbol{\sigma}'] = \delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} + \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{W}_i(F_i\boldsymbol{\sigma}) \delta_{\boldsymbol{\sigma}, F_i\boldsymbol{\sigma}'} - \mathcal{W}_i(\boldsymbol{\sigma}) \delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} \right).$$

Here F_i denotes the i-th spin-flip operator: $(F_i\boldsymbol{\sigma})_j = \sigma_j - 2\sigma_j\delta_{ij}$, and

$$\mathcal{W}_i(\boldsymbol{\sigma}) = \frac{1}{2} - \frac{1}{2}\sigma_i \tanh(\beta h_i(\boldsymbol{\sigma})).$$

If the weight matrix is symmetric ($w_{ij} = w_{ji}$) and self-interactions are absent ($w_{ii} = 0$), this process obeys detailed balance and the equilibrium distribution is given by

$$P[\boldsymbol{\sigma}] = \mathrm{e}^{-\beta H(\boldsymbol{\sigma}) + c_0(\boldsymbol{w}, \boldsymbol{\theta})}, \tag{3.3}$$

where

$$H(\boldsymbol{\sigma}) = -\frac{1}{2} \sum_{i \neq j} \sigma_i w_{ij} \sigma_j - \sum_i \theta_i \sigma_i, \tag{3.4}$$

[1)] This example is also treated in [7].

and c_0 is a normalisation constant.

If we assume that the network contains H hidden cells and $V = N - H$ visible cells, the energy function may be rewritten as

$$H(\boldsymbol{\sigma}^V;\boldsymbol{\sigma}^H) = -\sum_{i<j}\sigma_i^H w_{ij}^H \sigma_j^H - \sum_{i<j}\sigma_i^V w_{ij}^V \sigma_j^V - \sum_{i,j}\sigma_i^H w_{ij}^{HV}\sigma_j^V,$$

in which we have implicitly introduced $\sigma_0^H = \sigma_0^V = 1$ and replaced θ_i^H and θ_i^V by w_{0i}^H and w_{0i}^V respectively. Note that the number of degrees of freedom in the interaction matrix is $\frac{1}{2}H(H+1)$ in $\{w_{ij}^H\}$, $\frac{1}{2}V(V+1)$ in $\{w_{ij}^V\}$ and HV in $\{w_{ij}^{HV}\}$. (Recall that $w_{ij} = w_{ji}$ and $w_{ij} = 0$.)

The full space of probability distributions on $X = \{(\boldsymbol{\sigma}^V, \boldsymbol{\sigma}^H)\}$ is a manifold $\mathcal{M}$ of dimension $2^{H+V} - 1$. Only a small part of these distributions can be represented exactly by the neural network: the interaction matrix parametrizes a submanifold $\mathcal{S}$ of $\mathcal{M}$ of dimension $\frac{1}{2}H(H+1) + \frac{1}{2}V(V+1) + VH$. On the other hand, from an external point of view only the distribution on the visible cells is important. This distribution is defined by $2^V - 1$ parameters. This partitions $\mathcal{M}$ into submanifolds $\mathcal{N}_{Q^V}$ of dimension $2^V(2^H - 1)$ each, containing distributions that cannot be distinguished by looking at the visible cells only: Q^V labels the probability distribution $Q^V : \{-1,1\}^V \to [0,1]$ on the visible cells to which all distributions in $\mathcal{N}_{Q^V}$ are equivalent from an external point of view. Our task is to find a neural network that most closely approximates a given Q^V. In other words: we seek points $\mathrm{P} \in \mathcal{N}_{Q^V}$ and $\mathrm{P}' \in \mathcal{S}$ that are as 'close' to each other as possible. We shall show how this may be achieved by minimizing the divergence.

Minimization is quite straightforward if we notice that $\mathcal{S}$ is an exponential family (see appendix B) and that $\mathcal{N}_{Q^V}$ are mixture families. The first fact follows from (3.3) with (3.4), since it is clearly possible to introduce alternative random variables $\boldsymbol{s} = \boldsymbol{s}(\boldsymbol{\sigma})$ such that (3.3) may be rewritten as:

$$P[\boldsymbol{s}] = \mathrm{e}^{\sum w_{ij}s_{ij} + c_0(\boldsymbol{w})}.$$

Having thus established that $\mathcal{S}$ is 1-flat[1)], we can immediately conclude that $\mathcal{S}$ is 1-convex, since the parameter space $\{w_{ij}\}$ stretches out infinitely in every direction.

To see that $\mathcal{N}_{Q^V}$ is a mixture family, consider the following: $\mathcal{M}$ clearly is a mixture family, since it is possible to write any probability distribution $p = p(\boldsymbol{\sigma}) \in \mathcal{M}$ as

$$p(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} \theta^{\boldsymbol{\sigma}'}\delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'},$$

(where each of the $\theta^{\boldsymbol{\sigma}}$'s must be in $[0,1]$ and $\sum_{\boldsymbol{\sigma}}\theta^{\boldsymbol{\sigma}} = 1$ for p to be a properly normalized probability distribution. Since $\mathcal{N}_{Q^V}$ is defined from $\mathcal{M}$ by *linear* constraints:

$$\mathcal{N}_{Q^V} = \left\{ p \in \mathcal{M} \;\middle|\; \sum_{\boldsymbol{\sigma}^H} p(\boldsymbol{\sigma}^V;\boldsymbol{\sigma}^H) = Q^V(\boldsymbol{\sigma}^V) \quad (\forall \boldsymbol{\sigma}^V) \right\},$$

it follows that $\mathcal{N}_{Q^V}$ is also a mixture family, and is therefore -1-flat. Thus $\mathcal{N}_{Q^V}$ is -1-convex.

Our optimization problem can therefore be solved by introducing -1-affine coordinates (θ^μ) and $+1$-affine coordinates $(\tilde{\theta}_\mu)$ on $\mathcal{M}$ and minimizing the divergence

$$D(\mathrm{P},\mathrm{P}') = D(\boldsymbol{\theta}(\boldsymbol{\tau}), \tilde{\boldsymbol{\theta}}(\boldsymbol{\vartheta}))$$

[1)] In appendix B it is shown that the exponential family is 1-flat.

with respect to $\boldsymbol{\vartheta}$ and $\boldsymbol{\tau}$ simultaneously. (As before, (ϑ^α) are coordinates on $\mathcal{S}$ and (τ^a) are coordinates on $\mathcal{N}$.)

The beauty of this technique is that it is not necessary to do the evolution and find the moments of $\boldsymbol{\sigma}^H$ at each timestep, since the divergence is given by $D(\tau,\vartheta) = \Theta(\theta(\tau)+\tilde{\Theta}(\tilde{\theta}(\vartheta)) - \theta^\mu\tilde{\theta}_\mu$, which is specified in terms of the weights only. This makes the optimization process much less time consuming.

If minimizing the divergens seems an arbitrary thing to do, it may help to know that in the case considered in this example, the divergence equals the Kullback-Leibler distance

$$D(P\|Q) = \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \log \frac{P(\boldsymbol{\sigma})}{Q(\boldsymbol{\sigma})}. \tag{3.5}$$

We shall prove that $D(p_{\boldsymbol{w}}, p_{\boldsymbol{w}'}) = D(p_{\boldsymbol{w}'}\|p_{\boldsymbol{w}})$. To reduce the notational clutter, we shall relabel the weights by a single vector $\boldsymbol{w} = (w^\mu) \equiv (w^H_{ij}, w^V_{ij}, w^{HV}_{ij})$.

From the definition of the divergence, we have:

$$D(p_{\boldsymbol{w}}, p_{\boldsymbol{w}}) = \frac{\partial}{\partial w^\mu} D(p_{\boldsymbol{w}}, p_{\boldsymbol{w}'})\bigg|_{\boldsymbol{w}'=\boldsymbol{w}} = 0, \qquad \text{and} \qquad \frac{\partial}{\partial w^\mu}\frac{\partial}{\partial w^\nu} D(p_{\boldsymbol{w}}, p_{\boldsymbol{w}'}) = g_{\mu\nu}(\boldsymbol{w}).$$

In this equation, (w^μ) are supposed to be -1-affine coordinates.

It is not difficult to see that

$$D(p_{\boldsymbol{w}}\|p_{\boldsymbol{w}}) = \frac{\partial}{\partial w^\mu} D(p_{\boldsymbol{w}'}\|p_{\boldsymbol{w}})\bigg|_{\boldsymbol{w}'=\boldsymbol{w}} = 0.$$

All that remains to be shown then is that $\frac{\partial}{\partial w^\mu}\frac{\partial}{\partial w^\nu} D(p_{\boldsymbol{w}'}\|p_{\boldsymbol{w}}) = g_{\mu\nu}(\boldsymbol{w})$ for all $\boldsymbol{w}$ and $\boldsymbol{w}'$.

Differentiating (3.5) twice, we find:

$$\begin{aligned}\frac{\partial}{\partial w^\mu}\frac{\partial}{\partial w^\nu} D(p_{\boldsymbol{w}'}\|p_{\boldsymbol{w}}) &= \sum_{\boldsymbol{\sigma}} p_{\boldsymbol{w}'}(\boldsymbol{\sigma}) \frac{\partial}{\partial w^\mu}\frac{\partial}{\partial w^\nu} p_{\boldsymbol{w}}(\boldsymbol{\sigma}) \\ &= \sum_{\boldsymbol{\sigma}} p_{\boldsymbol{w}}(\boldsymbol{\sigma}) \frac{\partial}{\partial w^\mu}\frac{\partial}{\partial w^\nu} p_{\boldsymbol{w}}(\boldsymbol{\sigma}) - \sum_{\boldsymbol{\sigma}} \big[p_{\boldsymbol{w}'}(\boldsymbol{\sigma}) - p_{\boldsymbol{w}}(\boldsymbol{\sigma})\big] \frac{\partial}{\partial w^\mu}\frac{\partial}{\partial w^\nu} p_{\boldsymbol{w}}(\boldsymbol{\sigma}).\end{aligned} \tag{3.6}$$

The first term on the right will be recognized as $g_{\mu\nu}(\boldsymbol{w})$. The second term vanishes, as is shown by the following argument: we may expand

$$p_{\boldsymbol{w}'}(\boldsymbol{\sigma}) - p_{\boldsymbol{w}}(\boldsymbol{\sigma}) = (w'^\mu - w^\mu)\partial_\mu p_{\boldsymbol{w}} + \frac{1}{2}(w'^\mu - w^\mu)(w'^\nu - w^\nu)\partial_\mu\partial_\nu p_{\boldsymbol{w}} + \cdots.$$

Since (w^μ) as a -1-affine coordinate system on a mixture family, the probability distributions can be written in the form

$$p_{\boldsymbol{w}} = \sum_\mu w^\mu f_\mu(\boldsymbol{\sigma}).$$

Therefore, all higher order terms vanish identically. However, since $\mathcal{M}$ is a mixture family, it is -1-flat and thus also $+1$-flat. Hence the 1-connection vanishes:

$$\Gamma^{(1)}_{\mu\nu\rho} \equiv E\big[\frac{1}{p}\partial_\rho p\,\partial_\mu\partial_\nu p\big] = 0.$$

Therefore the entire second term of (3.6) vanishes, and the equality of the divergence and the Kullback-Leibler distance is proven.

CHAPTER 4

UNIQUENESS OF THE FISHER METRIC

The uniqueness of the Fisher information as a metric on statistical manifolds has recently been proven by Corcuera and Giummolè [3]. In this chapter we present an outline of their proof in a form that is intended to be accessible for those who have less intimate knowledge of probability theory than Corcuera and Giummolè presuppose.

The proof consists of three distinct steps:

1. Classification of metrics on manifolds of probability distributions over a finite number of 'atoms';
2. Generalization to infinite (continuous) sets;
3. Application to parametrized distributions.

The first of these was first performed by Čencov in [8]. The other two steps have been taken by Corcuera and Giummolè in [3].

4.1 The finite case

We are ultimately looking for metrics on manifolds of parametrized probability distributions that are invariant under tranformation of the random variables, and covariant under reparametrization. Starting out with finite sets only helps if we know what invariances we should require in the finite case to end up with the invariances we are looking for in the parametrized case. It turns out that invariance under *Markov embeddings* is the finite case equivalent of reparametrization covariance. We shall therefore start by looking at those.

4.1.1 Markov embeddings

Consider a set X (for example an interval in $\mathbb{R}^N$) and a partition $\boldsymbol{A} = \{A_1, \ldots, A_m\}$ of X. (A collection of subsets is a partition if (1) $\forall i, j : i \neq j \Rightarrow A_i \cap A_j = \emptyset$, and (2) $\bigcup_i A_i = X$.) Furthermore, let $\boldsymbol{B} = \{B_1, \ldots, B_n\}$ be a subpartition of $\boldsymbol{A}$, that is, a partition of X such that each B_i is contained entirely within one of the A_j's. In other words, there exists a partition $\boldsymbol{I} = \{I_1, \ldots, I_m\}$ of $\{1, \ldots, n\}$ such that

$$A_i = \bigcup_{j \in I_i} B_j.$$

One picture says more than a hundred mathematical symbols: figure 4.1 should make the setup clear.

Let $\mathcal{A} \cong \mathbb{R}_+^m$ be the collection of non-negative distributions on $\boldsymbol{A}$[1)] and similarly let $\mathcal{B} \cong \mathbb{R}_+^n$ be the collection of distribution on $\boldsymbol{B}$. A *Markov embedding* is a subpartition $\boldsymbol{B}$ of $\boldsymbol{A}$ together with a mapping f from $\mathcal{A}$ to $\mathcal{B}$, with $f(\boldsymbol{\xi})$ given by

$$f_j(\boldsymbol{\xi}) = \sum_{i=1}^{m} q_{ij}\xi_i,$$

where $q_{ij} = 0$ unless $j \in I_i$. (Thus only one term of the sum is non-zero.) Note that consistency requires that none of the q_{ij}'s be negative, and that $\sum_j q_{ij} = 1$ for all i.

Associated with f is a mapping $\bar{f}$ from $\mathcal{B}$ to $\mathcal{A}$:

$$\bar{f}_i(\boldsymbol{\eta}) = \sum_{j \in I_i} \eta_j.$$

Note that $\bar{f} \circ f = \mathbb{1} : \mathcal{A} \to \mathcal{A}$, but $f \circ \bar{f}$ is not necessarily equal to the identity mapping on $\mathcal{B}$.

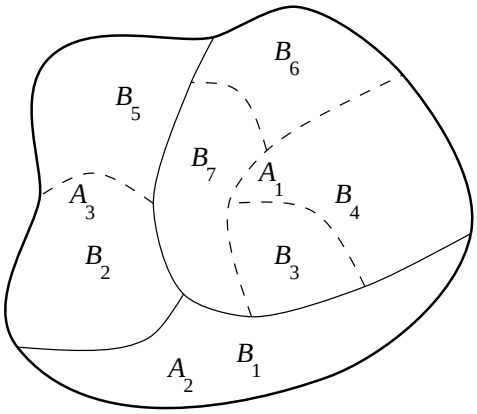

Figure 4.1: *A set X with a partition* $\boldsymbol{A}$ *and a subpartition* $\boldsymbol{B}$*, with dashed lines marking the boundaries of the elements of the subpartition.*

4.1.2 Embeddings and vectors

We may consider $\mathcal{A}$ as an m-dimensional manifold on which we may introduce vectors. In particular we may introduce the coordinate basis $\left\{\hat{\imath} \,\middle|\, i = 1, \ldots, m\right\}$, where $\hat{\imath}$ is the i-th coordinate basis vector in $\mathbb{R}_+^m$. A mapping $f : \mathcal{A} \to \mathcal{B}$ induces a mapping $f_* : T(\mathcal{A}) \to T(\mathcal{B})$ defined by

$$f_* : \hat{\imath} \mapsto f_*(\hat{\imath}) = \sum_{j=1}^{n} q_{ij}\tilde{\jmath},$$

where $\tilde{\jmath}$ is the j-th coordinate basis vector in $\mathbb{R}_+^n$. (These vector relations are the ones that are slightly more involved on the set of probability distributions $\left\{\xi \in \mathbb{R}_+^m \,\middle|\, \sum_i \xi_i = 1\right\}$, since the linear constraint reduces the dimension of the tangent space too.)

4.1.3 Invariant metrics

A sequence of inner products $\left\langle \cdot, \cdot \right\rangle_{(m)}$ on $T(\mathbb{R}_+^m)$ for $m = 2, 3, \ldots$ is said to be *embedding-invariant* if for any Markov embedding $f : \mathbb{R}_+^m \to \mathbb{R}_+^n$ (with $n > m$) the following holds:

$$\left\langle X\,, Y \right\rangle_{(m)}(\boldsymbol{\xi}) = \left\langle f_*(X)\,,\, f_*(Y) \right\rangle_{(n)}(f(\boldsymbol{\xi})).$$

Čencov proved that the only metrics that satisfy this invariance on the submanifold of probability distributions, $\left\{\xi \in \mathbb{R}_+^m \,\middle|\, \sum_i \xi_i = 1\right\}$ are given by

$$g_{ij}^{(m)}(\boldsymbol{\xi}) = c_0 \left\{ \frac{\delta_{ij}}{\xi_i} + \frac{1}{\xi_m} \right\}, \qquad \text{for } i, j = 1, \ldots, m-1,$$

where c_0 is a constant.

[1)] It turns out that the theory is much easier if we do not restrict ourselves to probability distributions, i.e. if we do not require that $\sum_i \xi_i = 1$ for $\boldsymbol{\xi} \in \mathcal{A}$. (This idea is taken from [9], in which a variant of Čencovs result is proven pertaining to such non-negative distributions.)

4.1.4 Rationale

Why are embedding invariant metrics relevant? The solution lies in the fact that a classification of metrics leads to a classification of *divergences*, and for divergences embedding invariance is a very natural property to require.

A divergence, in this context, is any object D that takes two probability distributions over the same partition $\boldsymbol{A}$ of X, and associates a real number to the pair. If we consider a divergence as a measure of the difference between two distributions P and Q, it is natural to require that

$$D(K(P),K(Q)) \leq D(P,Q)$$

for any mapping K from one space of probability distributions to another, since such a mapping cannot add new information, and should therefore not be expected to increase the distance between distributions. Divergences which satisfy this property are called *monotone*.

If there is a mapping $\bar{K}$ such that $\bar{K}(K(P)) = P$ for all P, and D is monotone, then

$$D(K(P),K(Q)) = D(P,Q),$$

because

$$D(P,Q) = D(\bar{K}(K(P)),\bar{K}(K(Q))) \leq D(K(P),K(Q)) \leq D(P,Q),$$

Since Markov embeddings are invertible in this sense, monotone divergences are invariant under Markov embeddings.

One other fact about monotone divergences plays a role in our present quest: since one may consider a mapping K_0 that maps any distribution down to some fixed P_0, we have for any P, Q:

$$D(P_0,P_0) = D(K_0(P),K_0(Q)) \leq D(P,Q).$$

In particular, $D(P,P)$ is a minimum of $D(P,Q)$. We may as well limit ourselves to divergences for which $D(P,P) = 0$ and $D(P,Q) > 0$ unless $P = Q$, consolidating the interpretation of divergences as a measure of the difference between two probability distributions.

For any partition $\boldsymbol{A}$ of X we can then expand

$$D^{(\boldsymbol{A})}(P,Q) = \sum_{i,j=1}^{\#\boldsymbol{A}} D_{ij}^{(\boldsymbol{A})}(P)\left(Q(A_i)-P(A_i)\right)\left(Q(A_j)-P(A_j)\right) + O([Q-P]^3), \tag{4.1}$$

where $D_{ij}^{(\boldsymbol{A})}(P)$ is a strictly positive definite symmetric rank 2 tensor field.

Now we see why embedding invariant metrics are interesting: any metric can be used to play the role of $D_{ij}(P)$ here, but only embedding invariant metrics will yield monotone divergences. Conversely, since any strictly positive definite symmetric rank 2 tensor field can be used as a metric, Čencovs theorem gives a classification of monotone divergences upto second order. It can be shown that the only monotone divergences are those for which $D_{ij}^{(\boldsymbol{A})}(P) = c_0\frac{\delta_{ij}}{P(A_i)}$. (A proof can be found in [3].)

4.2 Generalization to continuous sets

In the previous section we found classifications of metrics and of divergences for finite partitions $\boldsymbol{A}$, and we found that those classifications are tightly related. The next step is to show that a similar relation can be established when X is partitioned into infinitely many sets. We shall not go into the details, but merely state the results and the conditions under which they apply.

One may consider probability distributions on X as probability distributions on an infinitely fine partition of X, which can be viewed as a limit of finer and finer finite partitions. If P is a probability distribution on X, we denote by $P_{\mathcal{A}}$ the restriction of P to $\mathcal{A}$, that is $P_{\mathcal{A}}(A_i) = \int_{A_i} \mathrm{d}x\, P(x)$. Let $(\boldsymbol{A}_{(n)})$ be a sequence of partitions of X, with $\#\boldsymbol{A}_{(n)} = n$. A divergence D is said to be *regular* if

$$\lim_{n\to\infty} D(P_{\mathcal{A}_{(n)}}, Q_{\mathcal{A}_{(n)}}) = D(P,Q)$$

for any probability distributions P and Q on X, and irrespective of the sequence $(\boldsymbol{A}_{(n)})$.

Corcuera and Giummolè show that the only monotone and regular divergences for which $D(P,P) = 0$, are given by

$$D(P,Q) = c_0 \int_X \mathrm{d}x\, \frac{[Q(x)-P(x)]^2}{P(x)} + O([Q-P]^3).$$

Although the mathematics is more complicated, the end result turns out to be a straightforward generalization of the finite case.

We shall not try to write down a metric in this case, since that would force us to deal with the intricacies of infinite dimensional manifolds.

4.3 The parametric case

Instead, we move straight on to our final goal, showing that the Fisher information is unique. To this end, consider a subspace of parametrized probability distributions on X. If P is in this subspace, we may write $P = p_{\boldsymbol{\theta}}(x)$, where $\boldsymbol{\theta}$ is supposed to take values in some interval of $\mathbb{R}^N$ for some $N \in \mathbb{N}$. Any monotone and regular divergence between $P = p_{\boldsymbol{\theta}}(x)$ and $Q = p_{\boldsymbol{\theta}'}(x)$ can then be written as

$$D(P,Q) = D(\boldsymbol{\theta},\boldsymbol{\theta}') = A \int_X \mathrm{d}x\, \frac{[p_{\boldsymbol{\theta}'}(x) - p_{\boldsymbol{\theta}}(x)]^2}{p_{\boldsymbol{\theta}}(x)} + O([p_{\boldsymbol{\theta}'} - p_{\boldsymbol{\theta}}]^3), \tag{4.2}$$

(where we still take $D(\boldsymbol{\theta},\boldsymbol{\theta})$ to be zero).

In writing (4.2), we have implicitly assumed that monotonicity in the parametric case is equivalent to invariance under tranformations of x and $\boldsymbol{\theta}$: those transformations correspond to the invertible mappings of §4.1.4. We shall come back to this crucial point later.

For the moment, we shall proceed assuming that (4.2) is indeed a full classification of invariant divergences upto the stated order: the final term in (4.2), $O([p_{\boldsymbol{\theta}'} - p_{\boldsymbol{\theta}}]^3)$, is supposed to indicate the fact that we are still making expansions similar to the one in (4.1).

Taking this expansion a bit more seriously, we have

$$p_{\boldsymbol{\theta}'}(x) - p_{\boldsymbol{\theta}}(x) = (\theta'^{\mu} - \theta^{\mu})\partial_{\mu} p_{\boldsymbol{\theta}}(x) + O([\boldsymbol{\theta}' - \boldsymbol{\theta}]^2).$$

Inserting this into (4.2) gives:

$$D(\theta,\theta') = c_0 \int_X \mathrm{d}x \, \frac{1}{p_\theta(x)} \partial_\mu p_\theta(x) \, \partial_\nu p_\theta(x) (\theta'^\mu - \theta^\mu)(\theta'^\nu - \theta^\nu) + O([\theta' - \theta]^3).$$

Even now, the fact that any metric could be used as a prefactor for the $(\theta'^\mu - \theta^\mu)(\theta'^\nu - \theta^\nu)$ term remains unchanged. We conclude that the only metrics that give rise to regular and monotone divergences in the parametric case are

$$g_{\mu\nu}(\theta) = c_0 \int_X \mathrm{d}x \, \frac{1}{p_\theta(x)} \partial_\mu p_\theta(x) \, \partial_\nu p_\theta(x), \tag{4.3}$$

since invariance of D is equivalent to covariance of the prefactor for the $(\theta'^\mu - \theta^\mu)(\theta'^\nu - \theta^\nu)$ term. The metrics (4.3) are just the scalar multiples of the Fisher information.

Note that the uniqueness of the Fisher information as a covariant metric also proves the optimality of using the Fisher information in gradient descent learning: since the optimal learning rule should certainly by reparametrization invariant, and the only covariant gradient is $g^{\mu\nu}\partial_\nu$, this gradient cannot but yield the optimal rule, apart from a possible time-dependent scalar pre-factor.

One final word of warning seems to be in order: strictly speaking, we have only proven that the Fisher information is the only metric that can be used to build *regular* divergences. While it is clear that any parametrized divergence of the form (4.2) can be used to construct a monotone and regular divergence for the space of all probability distributions, it is not quite as obvious that *all* parameter invariant divergence must of necessity be extendable in this way. Only when this point is cleared up will the classification of monotone and regular divergences constitute a full mathematical proof of the uniqueness of the Fisher metric.

4.A Appendix: The more formal language of probability theory

Mathematicians have come up with a paradigm to avoid our admittedly vague term 'probablity distributions over finite or infinite partitions of a set X'. In this appendix we shall give a very brief introduction to this language, to serve as an introduction to the literature.

The following definitions have been taken from [10].

Definition: A collection $\mathcal{B}$ of subsets of a set X is called a *σ-algebra* if it satisfies

1. $\emptyset \in \mathcal{B}$;
2. If $A \in \mathcal{B}$, then also $X \backslash A \in \mathcal{B}$;
3. If $A_1, A_2, A_3, \ldots \in \mathcal{B}$, then $\bigcup_i A_i \in \mathcal{B}$.

A pair $(X, \mathcal{B})$ where $\mathcal{B}$ is a σ-algebra over X is called a *borel space*.

Definition: A map $m : \mathcal{B} \to [0,1]$ is called a *probability distribution* if

1. It is *countably additive*, that is

$$m(\bigcup_i A_i) = \sum_i m(A_i)$$

for any (countable) collection $\{A_i\}$ of pairwise disjoint sets in $\mathcal{B}$, and

2. $m(X) = 1$.

Definition: Let $(X_1, \mathcal{B}_1)$ and $(X_2, \mathcal{B}_2)$ be borel spaces, and let $f : X_1 \to X_2$ be a map. Then f is called a *measurable map* if

$$\forall A \in \mathcal{B}_2 : f^{-1}(A) \equiv \left\{ x \in X_1 \,\middle|\, f(x) \in A \right\} \in \mathcal{B}_1.$$

A map $f : X_1 \to Y \subseteq \mathbb{R}$ is similarly called a measurable map if

$$\forall y \in Y : f^{-1}((-\infty, y]) \equiv \left\{ x \in X_1 \,\middle|\, f(x) \leq y \right\} \in \mathcal{B}_1.$$

Definition: Let $(X, \mathcal{B})$ be a borel space. A map $f : X \to Y$ is called a *simple function* if f takes only a finite number of values, and for every $f^{-1}(\{y\}) \in \mathcal{B}$ for all $y \in Y$.

Definition: Integral

Let s be any non-negative simple function on $(X, \mathcal{B})$. Then there exists a partition of X into disjoint sets $A_1, \ldots, A_k$ all of which belong to $\mathcal{B}$, and k numbers $a_1, \ldots, a_k$ in $[0, \infty]$ such that $s = \sum_{i=1}^{k} a_i \chi_{A_i}$, where χ_A is the indicator function: $\chi_A(x) = 1$ if $x \in A$ and zero otherwise.

Let P be a probability distribution on X. We then define the *integral* $\int_X s \, \mathrm{d}P$ of s with respect to P by

$$\int_X s \, \mathrm{d}P = \sum_i a_i P(A_i).$$

If f is a non-negative borel function from X to $\mathbb{R}_+$[1)], we define its integral with respect to P by

$$\int_X f \, \mathrm{d}P = \sup \left\{ \int_X s \mathrm{d}P \,\middle|\, s \text{ is a simple function on } (X, \mathcal{B}), \text{ and } s(x) \leq f(x) \text{ for a } x \in X \right\}.$$

This leads to a more precise definition of mappings between spaces of probability distributions[2)]:

Definition: Given two borel spaces $(X_1, \mathcal{B}_1)$ and $(X_2, \mathcal{B}_2)$, $K : X_1 \times \mathcal{B}_2 \to [0, 1]$ is called a *Markov kernel* if

1. $K(\cdot, A) : X_1 \to [0, 1]$ is a measurable map for any $A \in \mathcal{B}_2$, and
2. $K(x, \cdot) : \mathcal{B}_2 \to [0, 1]$ is a probability distribution on $(X_2, \mathcal{B}_2)$.

[1)] A borel function is a basically a measurable map, with the exception that it need not be defined on subsets $A \in \mathcal{B}$ for which $P(A) = 0$.

[2)] based on [3].

If P is a probability distribution on $(X_1, \mathcal{B}_1)$, then K induces a probability distribution KP on $(X_2, \mathcal{B}_2)$ defined by

$$KP(A) = \int_{X_1} K(x, A) \mathrm{d}P,$$

where the integration is well-defined since $K(\cdot, A)$ is a measurable map.

The only further definition that is needed to understand [3] is the following:

Definition: Let $(X, \mathcal{B})$ be a borel space. A *finite sub-σ-field* is a subset $\mathcal{A} \subset \mathcal{B}$ such that

1. $\mathcal{A}$ is a σ-algebra over X and
2. $\mathcal{A}$ is finite.

4.B A proof of the invariance properties of the Fisher information

For the sake of completeness we shall show that the Fisher metric is indeed invariant under transformations of the random variable and covariant under reparametrizations. Both proofs are straightforward.

4.B.1 Invariance under transformations of the random variable

Suppose that our probability distributions are defined in terms of a random variable $\boldsymbol{x}$ taking values in $X \subseteq \mathbb{R}^n$. Then

$$g_{\mu\nu}(\boldsymbol{\theta}) = \int_X \mathrm{d}\boldsymbol{x} \, \frac{1}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} \partial_\mu p_{\boldsymbol{\theta}}(\boldsymbol{x}) \partial_\nu p_{\boldsymbol{\theta}}(\boldsymbol{x}).$$

We can re-express this in terms of another random variable $\boldsymbol{y}$ taking values in $Y \subseteq \mathbb{R}^n$, if we suppose that $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ is an invertible mapping. We clearly have:

$$\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \int_X \mathrm{d}\boldsymbol{x} \; p_{\boldsymbol{\theta}}(\boldsymbol{x}) \delta(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x})). \tag{4.4}$$

If $\boldsymbol{f}$ is invertible, then we can use the relation

$$\delta(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x})) = \frac{1}{\left|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|} \delta(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}),$$

to find that

$$\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \int_X \mathrm{d}\boldsymbol{x} \; p_{\boldsymbol{\theta}}(\boldsymbol{x}) \frac{1}{\left|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|} \delta(\boldsymbol{f}^{-1}(\boldsymbol{y}) - \boldsymbol{x}) = \left[\frac{1}{\left|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|} p_\theta(\boldsymbol{x}) \right]_{\boldsymbol{x} = \boldsymbol{f}^{-1}(\boldsymbol{y})} {}^{1)}. \tag{4.5}$$

[1)] Historically, this expression is in fact older than (4.4). However, at present the properties of the δ-function seem to be wider known than the properties of distributions in general.

If we further note that $\left|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|$ does not depend on $\boldsymbol{\theta}$, we see that

$$\begin{aligned}\int_Y \mathrm{d}\boldsymbol{y}\, \frac{1}{\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{y})}\partial_\mu \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{y})\partial_\nu \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{y}) &= \int_Y \mathrm{d}\boldsymbol{y}\, \left[\frac{1}{\frac{1}{\left|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|}p(\boldsymbol{x})}\left|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|\partial_\mu p_{\boldsymbol{\theta}}(\boldsymbol{x})\left|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|\partial_\nu p_{\boldsymbol{\theta}}(\boldsymbol{x})\right]_{\boldsymbol{x}=\boldsymbol{f}^{-1}(\boldsymbol{y})} \\ &= \int_X \mathrm{d}\boldsymbol{x}\, \frac{1}{p(\boldsymbol{x})}\partial_\mu p_{\boldsymbol{\theta}}(\boldsymbol{x})\partial_\nu p_{\boldsymbol{\theta}}(\boldsymbol{x}),\end{aligned}$$

since $\int_Y \mathrm{d}\boldsymbol{y} = \int_X \mathrm{d}\boldsymbol{x}\, \left|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|$.

This proves the invariance under transformation of the random variable.

4.B.2 Covariance under reparametrization

Suppose that $(\tilde{\theta}^\mu)$ is a new set of coordinates, specified in terms of the old set through the invertible relationship $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$. Defining $\tilde{p}_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{x}) \equiv p_{\boldsymbol{\theta}(\tilde{\boldsymbol{\theta}})}(\boldsymbol{x})$, we are then able to compute

$$\tilde{g}_{\mu\nu}(\tilde{\boldsymbol{\theta}}) \equiv \int_X \mathrm{d}\boldsymbol{x}\, \frac{1}{\tilde{p}_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{x})}\frac{\partial}{\partial \tilde{\theta}^\mu}\tilde{p}_{\tilde{\theta}}(\boldsymbol{x})\frac{\partial}{\partial \tilde{\theta}^\nu}\tilde{p}_{\tilde{\theta}}(\boldsymbol{x}),$$

in terms of $g_{\mu\nu}(\boldsymbol{\theta})$: since

$$\frac{\partial}{\partial \tilde{\theta}^\mu}\tilde{p}_{\tilde{\boldsymbol{\theta}}} = \frac{\partial \theta^\nu}{\partial \tilde{\theta}^\mu}\frac{\partial}{\partial \theta^\nu}p_{\boldsymbol{\theta}(\tilde{\boldsymbol{\theta}})},$$

we may directly conclude that

$$\tilde{g}_{\mu\nu}(\tilde{\boldsymbol{\theta}}) = \left[\frac{\partial \theta^\rho}{\partial \tilde{\theta}^\mu}\frac{\partial \theta^\lambda}{\partial \tilde{\theta}^\nu}g_{\rho\lambda}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tilde{\boldsymbol{\theta}})}.$$

This is precisely the covariance we claimed.

APPENDIX A

RIEMANNIAN GEOMETRY

This appendix introduces the basics of Riemannian geometry for the purpose of the main text. It is not intended as competition for the excellent textbooks[1)] that are available on Riemannian geometry, either in terms of depth or in terms of educational value.

A.1 Manifolds

For the purpose of this text, we shall not need to be completely formal in the definition of a manifold[2)]. For our purposes, a manifold is a set of points that allows the notion of connecting any two points by a smooth (continuously differentiable) curve. We shall also require that in the neighbourhood of any given point it is possible to define coordinates. These coordinate functions need not be global, and indeed there are many common manifolds that cannot be covered by any single coordinate patch. In such cases we just require the coordinate functions to be compatible: if κ_1, κ_2 and κ_3 are three coordinate functions defined in the vicinity of a point P, we insist that the *transition functions* $\phi_{ij} \equiv \kappa_i \circ \kappa_j^{-1}$ are differentiable and satisfy $\phi_{ij} \circ \phi_{jk} = \phi_{ik}$ for any i, j and k.

Example: The unit sphere $S^2 = \left\{ \boldsymbol{x} \in \mathbb{R}^3 \,\middle|\, \|\boldsymbol{x}\| = 1 \right\}$ is a manifold. It cannot be covered by a single coordinate function, but we may take $\kappa_\pm : \boldsymbol{x} \mapsto (\frac{x}{1 \pm z}, \frac{y}{1 \pm z})$, which are valid on the entire S^2 except on the south and north pole respectively. Together they clearly cover the sphere. It is to show that $\phi_{+-} \circ \phi_{-+}$ is the identity on $S^2 \backslash \{(0,0,1),\ (0,0,-1)\}$. ∎

In the present work we shall not worry too much about such global issues, and we shall assume that a single coordinate patch covers the entire manifold.

We will therefore use the following – very sloppy – definition of a manifold:

Definition: Manifold

A *manifold* $\mathcal{M}$ is a set of points that allows the notion of smooth curves connecting any two points, and for which we can define a continuously differentiable injection $\kappa : \mathcal{M} \to \mathbb{R}^n$ that is invertible on its co-domain.

---

1)[2] is a good introduction to general relativity. [11] is much more elaborate and mathematically precise. I found [12] and in particular [13] very useful sources on differential topology. However, they are both in Dutch and may not be easy to acquire.

2)In fact, giving a mathematically precise definition of 'manifold' seems to be quite difficult: one very useful introduction to differential geometry [13] gives a definition that starts with 'A manifold is a (para-)compact Hausdorff space satisfying the following properties:'. Even so the author warns that his definition is an 'intuitive' one!

A.2 Vectors

In Euclidean geometry one usually thinks of a vector as the straight line connecting two points. This definition doesn't make sense on manifolds, since we haven't yet defined the notion of straightness: a curve that appears to be straight in $\mathbb{R}^n$ in one coordinate mapping will generally not be straight in $\mathbb{R}^n$ in another.

A.2.1 Tangent vectors

To avoid these complications, we wish to define vectors locally, and independently of coordinates. Instead of thinking of a vector as the line segment connecting two points, we shall use the following:

Definition: The *tangent vector* X_{P} in a point P to a curve $\gamma(t)$ for which $\gamma(0) = \mathrm{P}$ is defined as

$$X_{\mathrm{P}} = \left.\frac{\mathrm{d}\gamma}{\mathrm{d}t}\right|_{t=0}.$$

Given a set of coordinate functions, that is a mapping

$$\kappa : \mathrm{P} \mapsto \left(\theta^1(\mathrm{P}), \theta^2(\mathrm{P}), \ldots, \theta^n(\mathrm{P})\right),$$

we may introduce a basis of coordinate vectors as follows: let $\gamma_{(\mu)}$ be the curve through P that satisfies $\kappa(\gamma_{(\mu)}(t)) = \kappa(\mathrm{P}) + t\hat{\mu}$, where $\hat{\mu}$ is the μ-th unit vector in $\mathbb{R}^n$. We then define $\hat{e}_\mu$, for $\mu = 1 \ldots n$, by

$$\hat{e}_\mu = \left.\frac{\mathrm{d}\gamma_{(\mu)}}{\mathrm{d}t}\right|_{t=0}. \tag{A.1}$$

Note that in the expression $\hat{e}_\mu$, μ does *not* label the components of a vector $\hat{e}$, rather, for each μ, $\hat{e}_\mu$ is a different vector.

The set of all tangent vectors at a point P is denoted by T_{P}. Any vector $X \in T_{\mathrm{P}}$ can be decomposed with respect to the basis (A.1):

$$X = X^\mu \hat{e}_\mu,$$

where summation over μ is left implicit, in line with the Einstein summation convention which we shall employ throughout this appendix.

If X is the tangent vector to a curve γ, then $X^\mu = \frac{\mathrm{d}}{\mathrm{d}t}\theta^\mu(\gamma(t))\big|_{t=0}$.

Note that at this stage we have not yet introduced an inner product, so we have no way of talking about the 'length' of a vector. In particular, we should note that it makes no sense to define $\|X\| \equiv \sqrt{\sum_\mu (X^\mu)^2}$, since this would depend critically on the choice of coordinates.

Remark: Mathematicians take it one step further. Given a manifold $\mathcal{M}$, and a curve $\gamma(t) \subset \mathcal{M}$ with $\gamma(0) = \mathrm{P}$, they consider the vector X_{P} tangent to γ at $t = 0$ as the derivative operator that takes the derivative of functions $f : \mathcal{M} \to \mathbb{R}$ in the direction of γ:

$$X f \equiv \nabla_X f = \left.\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\right|_{t=0}.$$

Using the shorthand $\partial_\mu f$ to denote the functions $\partial_\mu f : \mathrm{P} \mapsto \frac{\partial}{\partial\theta^\mu}\left[f \circ \kappa^{-1}(\boldsymbol{\theta})\right]\big|_{\boldsymbol{\theta}=\kappa(\mathrm{P})}$, they write $X_\mathrm{P} f = X^\mu \partial_\mu f$, or

$$X_\mathrm{P} \equiv X^\mu_\mathrm{P} \partial_\mu \text{ [1)]}.$$

In particular, the basis vectors $\hat{e}_\mu$ defined above may be identified with the derivative operators ∂_μ.

These are the first building blocks of the beautiful formalism of differential topology. In this thesis I shall avoid it as much as possible, in the hope of expanding my potential audience. This approach clearly has its disadvantages, and in places where the derivative operator formalism allows shortcuts or useful insights, these will be mentioned in footnotes. ▮

A.2.2 Vector fields

A vector-valued function defined on a manifold is called a vector field:

Definition: A *(contravariant) vector field X* over a manifold $\mathcal{M}$ is a mapping that assigns a vector X_P to each point $\mathrm{P} \in \mathcal{M}$. The set of all smooth vector fields over $\mathcal{M}$ is denoted by $T(\mathcal{M})$.

Definition: Coordinate basis
Given a manifold $\mathcal{M}$ and a coordinate function $\kappa : \mathcal{M} \to \mathbb{R}^n$, the μ-th coordinate basis vector field $\hat{e}_\mu$ is the vector field that assigns the μ-th basis vector $\hat{e}_\mu\big|_\mathrm{P}$ to each point $\mathrm{P} \in \mathcal{M}$.

A.2.3 Transformation behaviour

Given two sets of coordinates, (θ^μ) and $(\tilde{\theta}^{\tilde{\mu}})$ two bases for $T_\mathcal{M}$ are induced: $\{\hat{e}_\mu\}$ and $\{\hat{e}_{\tilde{\mu}}\}$. Vector fields may then be decomposed in either of two ways:

$$X = X^\mu \hat{e}_\mu \qquad \text{or} \qquad X = X^{\tilde{\mu}} \hat{e}_{\tilde{\mu}}.$$

The components are related by

$$X^\mu = J^\mu{}_{\tilde{\mu}} X^{\tilde{\mu}}, \tag{A.2}$$

where $J^\mu{}_{\tilde{\mu}} \equiv \frac{\partial\theta^\mu}{\partial\tilde{\theta}^{\tilde{\mu}}}$ [2)] is the Jacobian of the transformation. Some texts actually take this transformation behaviour as the defining property of a vector field.

[1)] This definition of the concept *vector* may seem to be far removed from the original, yet it is clear that both definitions contain the same information, and are therefore equally valid.

[2)] This is most easily seen when identifying $\hat{e}_\mu \equiv \partial_\mu$, since it is quite obvious that

$$\partial_\mu = \frac{\partial\tilde{\theta}^{\tilde{\mu}}}{\partial\theta^\mu}\partial_{\tilde{\mu}}.$$

Requiring that $X^\mu \hat{e}_\mu = X^{\tilde{\mu}} \hat{e}_{\tilde{\mu}}$ then implies (A.2).

A.3 Tensor fields

Vector fields are not the only objects that have 'sensible' transformation behaviour in terms of the Jacobian. More generally, one may consider rank (n,m)-tensor fields. For the purpose of this text, these are simply objects $T^{\mu_1\cdots\mu_m}{}_{\nu_1\cdots\nu_n}$ which transform according to

$$T^{\mu_1\cdots\mu_m}{}_{\nu_1\cdots\nu_n} = J^{\mu_1}{}_{\tilde{\mu}_1}\cdots J^{\mu_m}{}_{\tilde{\mu}_m} J_{\nu_1}{}^{\tilde{\nu}_1}\cdots J_{\nu_n}{}^{\tilde{\nu}_n} T^{\tilde{\mu}_1\cdots\tilde{\mu}_m}{}_{\tilde{\nu}_1\cdots\tilde{\nu}_n},$$

where $J_\mu{}^{\tilde{\mu}}$ is the matrix inverse of $J^\mu{}_{\tilde{\mu}}$. Because of this transformation behaviour, tensors may be multiplied in arbitrary ways to yield new tensors as long as one sticks to the summation convention. For example, if A is a $(2,0)$-tensor field $A_{\mu\nu}$ and X and Y are vector fields, then $A(X,Y) \equiv A_{\mu\nu}X^\mu Y^\nu$ is a scalar field in its transformation behaviour: the Jacobians cancel one another.

Remark: More formally, a rank (n,m)-tensor field is a linear mapping that maps m covariant[1)] and n contravariant vector fields into real valued functions on $\mathcal{M}$:

$$T(A^{(1)},\ldots,A^{(m)},X_{(1)},\ldots X_{(n)}) = T^{\mu_1\cdots\mu_m}{}_{\nu_1\cdots\nu_n} A^{(1)}_{\mu_1}\cdots A^{(m)}_{\mu_m} X^{\nu_1}_{(1)}\cdots X^{\nu_n}_{(n)}.$$

The space of (n,m)-tensor fields over $\mathcal{M}$ is written as $T_{(n,m)}(\mathcal{M})$.

An (n,m)-tensor field can also be viewed as mapping $(0,n)$-tensor fields into $(m,0)$-tensor fields, or as mapping $m-1$ 1-forms and n vector fields into vector fields. To give a specific example, consider a $(2,1)$-tensor field T. If X and Y are vector fields, and A is a 1-form, then T may be used as any of the following mappings:

$$\begin{aligned} T:(A,X,Y) &\mapsto T^\mu{}_{\nu\rho}A_\mu X^\nu X^\rho \in \mathbb{R} \\ T:A &\mapsto [T^\mu{}_{\nu\rho}A_\mu] \in T_{(2,0)}(\mathcal{M}) \\ T:(X,Y) &\mapsto T^\mu{}_{\nu\rho}X^\nu Y^\rho \hat{e}_\mu \in T(\mathcal{M}). \end{aligned}$$

If T is symmetric in its covariant indices, one may also define $T(A,X)$ and $T(X)$ without ambiguity. ∎

A.4 Metrics

A metric is an object $g_{\mu\nu}$ that assigns length to vectors and to curves in the following way: given a vector A_{P} at some point P in a manifold $\mathcal{M}$, the length of this vector is defined as

$$\|A_{\mathrm{P}}\| \equiv \sqrt{g(A,A)}\Big|_{\mathrm{P}} \equiv \sqrt{g_{\mu\nu}A^\mu A^\nu}\Big|_{\mathrm{P}}$$

in any coordinate system.

[1)] A covariant vector field or 1-form A is a linear map from contravariant vectors fields into functions: $A(X) = A_\mu X^\mu$. 1-forms transform according to $A_\mu = J_\mu{}^{\tilde{\mu}}A_{\tilde{\mu}}$. 1-forms may be expanded in terms of basis 1-forms ω^μ which are defined by $\omega^\mu(\hat{e}_\nu) = \delta^\mu_\nu$ in terms of a basis for contravariant vector fields.

The length s of a curve $\gamma : [0,1] \to \mathcal{M}$ is given by

$$s = \int \mathrm{d}t \, \sqrt{g_{\mu\nu}(\gamma(t)) \frac{\mathrm{d}\theta^\mu(\gamma(t))}{\mathrm{d}t} \frac{\mathrm{d}\theta^\nu(\gamma(t))}{\mathrm{d}t}}. \tag{A.3}$$

A metric also induces an inner product on $T(\mathcal{M})$: for any two vectors A and B defined at the same point $\mathrm{P} \in \mathcal{M}$, we may define the inner product $\langle A, B \rangle$ as

$$\langle A, B \rangle = g(A,B) = g_{\mu\nu} A^\mu B^\nu.$$

A metric must be symmetric and transform as a proper tensor field: given two sets of coordinates (θ^μ) and $(\tilde{\theta}^{\tilde{\mu}})$ we must have

$$g_{\tilde{\mu}\tilde{\nu}} = \frac{\partial \theta^\mu}{\partial \tilde{\theta}^{\tilde{\mu}}} \frac{\partial \theta^\nu}{\partial \tilde{\theta}^{\tilde{\nu}}} g_{\mu\nu},$$

since otherwise the induced inner product would not be coordinate independent or symmetric.

We shall sometimes have use for a contravariant form of the metric, $g^{\mu\nu}$, which is defined as the matrix inverse of $g_{\mu\nu}$, i.e. $g^{\mu\nu} g_{\nu\rho} = \delta^\mu_\rho$.

Metrics may be used to raise and lower tensor indices, ie given a (n,m)-tensor $T^{\mu_1 \cdots \mu_m}{}_{\nu_1 \cdots \nu_n}$ we may define the $(n+1, m-1)$-tensor $T_\mu{}^{\mu_2 \cdots \mu_m}{}_{\nu_1 \cdots \nu_n} \equiv g_{\mu\mu_1} T^{\mu_1 \cdots \mu_m}{}_{\nu_1 \cdots \nu_n}$, and the $(n-1, m+1)$-tensor $T^{\mu_1 \cdots \mu_m \nu}{}_{\nu_2 \cdots \nu_n} \equiv g^{\nu\nu_1} T^{\mu_1 \cdots \mu_m}{}_{\nu_1 \cdots \nu_n}$, etcetera.

A.5 Affine and metric connection

One is often interested in the rate of change of various fields as one moves around a manifold. In the case of a scalar field (a function), this is no problem: one just takes the values at two nearby points to compute the gradient. However, for vector fields (or tensor fields in general), this is not so easy, since the tangent spaces at two different points are unequal: whereas for a real valued function f the difference $f(\mathrm{P}') - f(\mathrm{P})$ is again a real number, for a vector field X, we cannot meaningfully compute $X(\mathrm{P}') - X(\mathrm{P})$, since the basis vectors at P' are not in T_P and vice versa. Is there a way around this? Yes and no.

No, because there is no unique way to solve this problem: given a manifold $\mathcal{M}$ there is no unique way to define the rate of change of vector fields. However, if one is willing to accept that, the problem may be solved by introducing a linear mapping Φ that takes vectors at P' and maps them into T_P, allowing comparison between $\Phi(X(\mathrm{P}'))$ and $X(\mathrm{P})$.

A.5.1 Affine connection and parallel transport

We shall be slightly more precise: consider again the set of curves $\{\gamma_{(\mu)}\}$ passing through P and satisfying $\dot{\gamma}_{(\mu)}(\mathrm{P}) = \hat{e}_\mu(\mathrm{P})$, as in §A.1. Let $\mathrm{P}'_{(\mu,\delta t)}$ be points on $\gamma_{(\mu)}$ near P, in the sense that $\mathrm{P}'_{(\mu,\delta t)} = \gamma_{(\mu)}(\delta t)$ for infinitesimal δt, and let $\Phi_{(\mu,\delta t)}$ be linear mappings from $T_{\mathrm{P}'_{(\mu,\delta t)}}$ to T_P, that reduce to the identity as $\delta t \to 0$. Linearity means that $\Phi_{(\mu,\delta t)}$ are completely defined by their actions on the coordinate vectors as follows:

$$\Phi_{(\mu,\delta t)} : \hat{e}^{(\mu,\delta t)}_\rho \mapsto \Phi^\nu_{(\mu,\delta t)}(\hat{e}^{(\mu,\delta t)}_\rho)\, \hat{e}_\nu,$$

where $\{\hat{e}_\rho^{(\mu,\delta t)}\}$ and $\{\hat{e}_\nu\}$ are the coordinate bases at $\mathrm{P}'_{(\mu,\delta t)}$ and P respectively.

If $\Phi_{(\mu,\delta t)}$ are to reduce to the identity as $\delta t \to 0$, we can write

$$\Phi_{(\mu,\delta t)}(\hat{e}_\nu^{(\mu,\delta t)}) - \hat{e}_\nu = \delta t\, \Gamma_{\mu\nu}{}^\rho\, \hat{e}_\rho.$$

for small δt. The constants $\Gamma_{\mu\nu}{}^\rho$ are called the coefficients of the *affine connection*.

Just as one defines

$$\partial_\mu f(\mathrm{P}) = \lim_{\delta t\to 0} \frac{f(\mathrm{P}'_{(\mu,\delta t)}) - f(\mathrm{P})}{\delta t}$$

for scalar functions f, we may now define the *covariant derivatives* of $\hat{e}_\nu$ as

$$\nabla_\mu \hat{e}_\nu = \lim_{\delta t\to 0} \frac{\Phi_{(\mu,\delta t)}(\hat{e}_\nu^{(\mu,\delta t)}) - \hat{e}_\nu}{\delta t} = \Gamma_{\mu\nu}{}^\rho\, \hat{e}_\rho. \tag{A.4}$$

Note that for any pair (μ,ν), $\nabla_\mu \hat{e}_\nu$ is a vector.

A.5.1.1 Formal derivation

In this section we shall derive the action of the covariant derivative on vector fields and general tensor fields. Those who are not interested in the mathematical details, may wish to skip it.

We define the covariant derivative of a function to be the ordinary derivative:

$$\nabla_\mu f \equiv \partial_\mu f, \tag{A.5}$$

and demand that ∇ behaves like a proper derivative operator, that is:

1. $\nabla_\mu(\alpha T) = \alpha \nabla_\mu(T)$ (for any tensor T and constant α),
2. $\nabla_\mu(T+S) = \nabla_\mu(T) + \nabla_\mu(S)$ (for any two tensors T and S of the same rank),
3. $\nabla_\mu(T\cdot S) = T\cdot\nabla_\mu(S) + \nabla_\mu(T)\cdot S$ (for any two tensors T and S),

where $\cdot$ may represent either tensor multiplication or some contraction.

These properties and (A.4) allow us to conclude that for general vector fields X:

$$\nabla_\mu X = \nabla_\mu(X^\nu \hat{e}_\nu) = X^\nu \nabla_\mu(\hat{e}_\nu) + \nabla_\mu(X^\nu)\hat{e}_\nu = X^\nu \Gamma_{\mu\nu}{}^\rho \hat{e}_\rho + \partial_\mu X^\nu \hat{e}_\nu. \tag{A.6}$$

They also allow us to conclude that

$$\nabla_\mu \omega^\nu = -\Gamma_{\mu\rho}{}^\nu \omega^\rho, \tag{A.7}$$

since $\omega^\nu(\hat{e}_\rho) = \delta^\nu_\rho$ implies that

$$0 = \partial_\mu(\delta^\nu_\rho) = \omega^\nu(\nabla_\mu \hat{e}_\rho) + (\nabla_\mu \omega^\nu)(\hat{e}_\rho) = \omega^\nu(\Gamma_{\mu\rho}{}^\lambda \hat{e}_\lambda) + (\nabla_\mu \omega^\nu)(\hat{e}_\rho) = \Gamma_{\mu\rho}{}^\lambda \delta^\nu_\lambda + (\nabla_\mu \omega^\nu)(\hat{e}_\rho),$$

whence

$$(\nabla_\mu \omega^\nu)_\rho = -\Gamma_{\mu\rho}{}^\nu,$$

as claimed.

Finally, we may combine (A.4), (A.5) and (A.7) to find that for a general rank (n,m)-tensor T:

$$(\nabla_\rho T)^{\nu_1\cdots\nu_n}_{\mu_1\cdots\mu_m} = \partial_\rho T^{\nu_1\cdots\nu_n}_{\mu_1\cdots\mu_m} + \Gamma_{\rho\lambda}{}^{\nu_1} T^{\lambda\nu_2\cdots\nu_n}_{\mu_1\cdots\mu_m} + \cdots + \Gamma_{\rho\lambda}{}^{\nu_n} T^{\nu_1\cdots\nu_{n-1}\lambda}_{\mu_1\cdots\mu_m} - \Gamma_{\rho\mu_1}{}^{\lambda} T^{\nu_1\cdots\nu_n}_{\lambda\mu_2\cdots\mu_m} - \cdots - \Gamma_{\rho\mu_m}{}^{\lambda} T^{\nu_1\cdots\nu_n}_{\mu_1\cdots\mu_{m-1}\lambda}. \tag{A.8}$$

Luckily we shall scarcely ever need to take the covariant derivative of anything but functions and vector fields.

A.5.1.2 Working definitions

For the purposes of the rest of this text the following 'working definitions' of the covariant derivative are sufficient:

Definition: The *covariant derivative of a scalar function* is defined to be the ordinary derivative:

$$\nabla_\mu f \equiv \partial_\mu f.$$

Definition: The *covariant derivative of a vector field* is given by

$$\nabla_\mu X \equiv \left[\partial_\mu X^\rho + \Gamma_{\mu\nu}{}^\rho X^\nu\right] \hat{e}_\rho.$$

Definition: Covariant derivative in the direction of a vector

The covariant derivative of a function f in the direction of the vector X is defined by:

$$\nabla_X f \equiv X^\mu \nabla_\mu f = X^\mu \partial_\mu f.$$

The covariant derivative of a vector field Y in the direction of a vector X is defined by

$$\nabla_X Y \equiv X^\mu \nabla_\mu Y.$$

As mentioned before, $\Gamma_{\mu\nu}{}^\rho$ are called the components of the affine connection:

Definition: An *affine connection* is an object $\Gamma_{\mu\nu}{}^\rho$ that induces a covariant derivative as in the definition above. Requiring that $\nabla_Y X$ be a vector implies that the affine connection is not a $(2,1)$-tensor field. In fact, if (θ^μ) and $(\tilde{\theta}^{\tilde{\mu}})$ are two sets of coordinates, one may show that

$$\Gamma_{\tilde{\mu}\tilde{\nu}}{}^{\tilde{\rho}} = \frac{\partial\theta^\mu}{\partial\tilde{\theta}^{\tilde{\mu}}}\frac{\partial\theta^\nu}{\partial\tilde{\theta}^{\tilde{\nu}}}\frac{\partial\tilde{\theta}^{\tilde{\rho}}}{\partial\theta^\rho}\Gamma_{\mu\nu}{}^\rho - \frac{\partial\theta^\mu}{\partial\tilde{\theta}^{\tilde{\mu}}}\frac{\partial\theta^\nu}{\partial\tilde{\theta}^{\tilde{\nu}}}\frac{\partial^2\tilde{\theta}^{\tilde{\rho}}}{\partial\theta^\mu\partial\theta^\nu}. \tag{A.9}$$

This transformation behaviour can be regarded as the defining property for affine connections[1)].

The antisymmetric part of an affine connection is called the torsion tensor:

[1)] One should perhaps write 'an affine connection is an object Γ with components $\Gamma_{\mu\nu}{}^\rho$', but Γ not being a tensor we shall always write it in component form.

Definition: For any affine connection, the *torsion tensor* is defined as

$$T^{\rho}_{\mu\nu} \equiv \Gamma_{\mu\nu}{}^{\rho} - \Gamma_{\nu\mu}{}^{\rho}.$$

From (A.9) we see that the torsion tensor *does* transform as a proper $(2,1)$-tensor.

An affine connection may be used to define what we mean by parallel transport:

Definition: Parallel transport

A vector field X is said to be *parallelly transported*, or *parallelly propagated*, along a curve γ with tangent vector field Y if

$$\nabla_Y X = 0.$$

A curve for which the tangent vector is propagated parallel to itself is called an affine geodesic:

Definition: A curve γ with tangent vector field X is called an *affine geodesic* if there exists a parametrisation of the curve such that

$$\nabla_X X = 0$$

at all points along the curve.

In terms of coordinates, we may write $X^{\mu}(\gamma(t)) = \dot{x}^{\mu}(t)$, where $x^{\mu}(t)$ are the coordinates of $\gamma(t)$, and the geodesic equation becomes

$$\ddot{x}^{\rho} + \Gamma_{\mu\nu}{}^{\rho}\dot{x}^{\mu}\dot{x}^{\nu} = 0.$$

A.5.2 Metric connection

A metric induces a special affine connection: the *metric connection*, or *Riemannian connection*. It is derived from the notion of *metric geodesic*:

Definition: A *metric geodesic* connecting two points is a[1)] shortest curve between those points in the sense of (A.3).

It is possible to show (see eg. [2]) that metric geodesics are also affine geodesics for the following connection:

$$\Gamma_{\mu\nu}{}^{\rho} = \frac{1}{2}g^{\rho\lambda}\left[\partial_{\mu}g_{\nu\lambda} + \partial_{\nu}g_{\mu\lambda} - \partial_{\lambda}g_{\mu\nu}\right] \tag{A.10}$$

and vice versa. The connection (A.10) is called the metric connection. It satisfies $\Gamma_{\mu\nu}{}^{\rho} = \Gamma_{\nu\mu}{}^{\rho}$, and so the torsion tensor vanishes: the metric connection is *torsion free*. (Another important property of the metric connection is that it satisfies $\nabla_{\rho}g_{\mu\nu} = 0$.)

On a space with metric connection, a global notion of distance is provided by the *Riemannian distance*: the length of the metric geodesic between two points.

[1)] For infinitesimally separated points this geodesic is unique, but more generally there may be more than one local minimum. As an example, consider walking through a hilly area. When the hills are not very steep, there will be one shortest path between two points that are nearby. However, in the vicinity of a fairly high hill, there may be two locally-shortest paths: one around each side of the hill.

A.6 Curvature

To top of the *exposé* of Riemannian geometry presented in this appendix, we introduce just one more important tensor that will be used in the main text: the curvature tensor. We shall start with an intuitive introduction, followed by a more formal definition.

A.6.1 Intuitive introduction

In general, when one transports a vector parallelly from one point to another point in a curved space, the result is not path independent. The canonical example is carrying a spear around the earth: start of from the north pole pointing a spear in the direction of the Greenwich meridian and walk all the way down to the equator, still pointing the spear horizontally along the meridian. Not changing the direction of the spear, turn around and walk along the equator until you reach the 90° meridian. Still not rotating the spear, walk back to the north pole. The spear now points in the direction of the 90° meridian, in contrast to the original situation.

How does this happen? The answer must lie in the observation that parallel transport according to the metric connection on the earth is not the same as parallel transport in the three dimensional space in which the earth is embedded. On a sphere one finds that the angle of deviation is equal to the covered spherical angle. We shall wish to find a quantitative result for more general manifolds though.

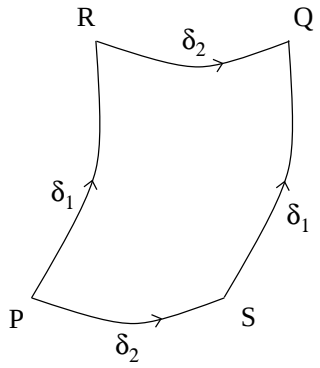

Figure A.1: *Parallel transport from a point* P *to a point* Q *along two paths. On a curved manifold, the result of parallel transport is path-dependent.*

To this end, consider carrying a vector X from a point P to a nearby point Q along two paths (see figure). Let's assign coordinates $\boldsymbol{\theta} = (\theta^\mu)$ to P, $\boldsymbol{\theta} + \boldsymbol{\delta}_1$ to R, $\boldsymbol{\theta} + \boldsymbol{\delta}_2$ to S and $\boldsymbol{\theta} + \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2$ to Q. We shall assume that all components of both $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are infinitesimal.

First carry X from P to R:

$$X_{\mathrm{R}} = X_{\mathrm{P}} - \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}}\, \delta_1^\mu X_{\mathrm{P}}^\nu \hat{e}_\rho .$$

(This follows from requiring that the covariant derivative of X in the direction of $\boldsymbol{\delta}_1$ should vanish. Integrating $0 = \nabla_\mu X \equiv \partial_\mu X + \Gamma_{\mu\nu}{}^{\rho} X^\nu \hat{e}_\mu$ in the direction of $\boldsymbol{\delta}_1$, we find that the first term integrates to $X_{\mathrm{R}} - X_{\mathrm{P}}$, while the second term yields $\Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}}\, \delta_1^\mu X_{\mathrm{P}}^\nu \hat{e}_\rho$ to first order in δ_1.)

Expanding

$$\Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{R}} = \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}} + \delta_1^\lambda \partial_\lambda\, \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}} + O\left([\boldsymbol{\delta}_1]^2\right),$$

we next carry X from R to Q:

$$\begin{aligned}
X_{\mathrm{Q\ (via\ R)}} &= X_{\mathrm{R}} - \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{R}}\, \delta_2^\mu X_{\mathrm{R}}^\nu \hat{e}_\rho \\
&= \left(X_{\mathrm{P}} - \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}}\, \delta_1^\mu X_{\mathrm{P}}^\nu \hat{e}_\rho\right) - \left(\Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}} + \delta_1^\lambda \partial_\lambda\, \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}}\right) \delta_2^\mu \left(X_{\mathrm{P}}^\nu - \Gamma_{\sigma\tau}{}^{\mu}|_{\mathrm{P}}\, \delta_1^\sigma X_{\mathrm{P}}^\tau\right) \hat{e}_\rho \\
&= X_{\mathrm{P}} - \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}} \left(\delta_1^\mu + \delta_2^\mu\right) X_{\mathrm{P}}^\nu \hat{e}_\rho + \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}}\, \Gamma_{\sigma\tau}{}^{\nu}|_{\mathrm{P}}\, \delta_2^\mu \delta_1^\sigma X_{\mathrm{P}}^\tau \hat{e}_\rho + \\
&\quad - \left(\partial_\lambda\, \Gamma_{\mu\nu}{}^{\rho}\big|_{\mathrm{P}}\right) \delta_1^\lambda \delta_2^\mu X_{\mathrm{P}}^\nu \hat{e}_\rho + O\left([\boldsymbol{\delta}_a]^3\right).
\end{aligned}$$

$X_{\mathrm{Q\,(via\,S)}}$ may be found from scratch, or more easily by interchanging $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ in the expression for $X_{\mathrm{Q\,(via\,R)}}$. What is interesting, of course, is the difference between these two:

$$X_{\mathrm{Q\,(via\,S)}} - X_{\mathrm{Q\,(via\,R)}} = \left(\Gamma_{\mu\sigma}{}^{\lambda}\big|_{\mathrm{P}}\,\Gamma_{\nu\rho}{}^{\sigma}\big|_{\mathrm{P}} - \Gamma_{\nu\sigma}{}^{\lambda}\big|_{\mathrm{P}}\,\Gamma_{\mu\rho}{}^{\sigma}\big|_{\mathrm{P}} - \partial_\nu\,\Gamma_{\mu\rho}{}^{\lambda}\big|_{\mathrm{P}} + \partial_\mu\,\Gamma_{\nu\rho}{}^{\lambda}\big|_{\mathrm{P}}\right)\delta_1^\mu\,\delta_2^\nu X_{\mathrm{P}}^\rho\hat{e}_\lambda.$$

The expression between brackets is called the Riemann tensor $R_{\mu\nu\rho}{}^{\lambda}$. It expresses the λ-th component of the difference between the result of parallelly transporting $\hat{e}_\rho$ first in the ν-th direction and then in the μ-th direction, and the result of doing it in the reverse order.

Note for the attentive reader: in calculating X_{Q} to second order in $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$, we should of course really have taken second order terms into account when calculating X_{R}. However, careful inspection shows that any $[\boldsymbol{\delta}_1]^2$ and $[\boldsymbol{\delta}_2]^2$ terms cancel in the final result.

A.6.2 Formal definition

For any connection $\Gamma_{\mu\nu}{}^{\rho}$ the curvature tensor, or *Riemann tensor* is defined by

$$R_{\mu\nu\rho}{}^{\lambda} = \partial_\mu\Gamma_{\nu\rho}{}^{\lambda} - \partial_\nu\Gamma_{\mu\rho}{}^{\lambda} + \Gamma_{\mu\sigma}{}^{\lambda}\Gamma_{\nu\rho}{}^{\sigma} - \Gamma_{\nu\sigma}{}^{\lambda}\Gamma_{\mu\rho}{}^{\sigma}. \tag{A.11}$$

There are several further observations one can make about the Riemann tensor:

- From the definition we see that it is antisymmetric in its first two indices:

$$R_{\mu\nu\rho}{}^{\lambda} = -R_{\nu\mu\rho}{}^{\lambda}. \tag{A.12}$$

- For a symmetric (or torsion free) connection the last two terms cancel. The Riemann tensor for a torsion free connection then satisfies

$$R_{\mu\nu\rho}{}^{\lambda} + R_{\nu\rho\mu}{}^{\lambda} + R_{\rho\mu\nu}{}^{\lambda} = 0 \qquad \text{(for torsion free connections)}.$$

- For the metric connection we can also establish the following identity:

$$R_{\mu\nu\rho\lambda} = R_{\rho\lambda\mu\nu} \qquad \text{(for the metric connection)}.$$

  Combining this with (A.12) we find $R_{\mu\nu\rho\lambda} = -R_{\mu\nu\lambda\rho}$. Finally the curvature for the metric connection also satisfies the Bianchi identities:

$$\nabla_\mu R_{\sigma\nu\rho}{}^{\lambda} + \nabla_\nu R_{\sigma\rho\mu}{}^{\lambda} + \nabla_\rho R_{\sigma\mu\nu}{}^{\lambda} = 0 \qquad \text{(for the metric connection)}.$$

  (Both of these relations are easy to prove in geodesic coordinates, see § 6.6 of [2].)

A.6.3 Affine flatness

A manifold on which the Riemann tensor vanishes everywhere is called (affine) flat. On an affine flat manifold, it is possible to introduce a coordinate system in which the connection vanishes everywhere[1)]. Such a coordinate system is called an affine coordinate system.

[1)] A proof of this fact can be found in § 6.7 of [2].

A.7 Submanifolds

Consider an n-dimensional subspace $\mathcal{S}$ of an m-dimensional manifold $\mathcal{M}$ with coordinates θ^μ, $\mu = 1 \ldots m$. On $\mathcal{S}$ we may introduce coordinates ϑ^α, $\alpha = 1 \ldots n$. We consider $\mathcal{S}$ to be embedded in $\mathcal{M}$: any point in $\mathcal{S}$ is also a point in $\mathcal{M}$. The $\mathcal{M}$-coordinates of such a point are a function of the $\mathcal{S}$-coordinates: $\theta = \theta(\vartheta)$. When this function is smooth and the coordinate vectors

$$\hat{e}_\alpha = \frac{\partial \theta^\mu}{\partial \vartheta^\alpha} \hat{e}_\mu \text{ [1)]} \tag{A.13}$$

are linearly independent, that is, the Jacobian

$$B^\mu_\alpha \equiv \frac{\partial \theta^\mu}{\partial \vartheta^\alpha}$$

is non-zero definite[2)], $\mathcal{S}$ is called a submanifold of $\mathcal{M}$.

A metric on $\mathcal{M}$ naturally induces a metric on $\mathcal{S}$:

$$g_{\alpha\beta} \equiv \langle \hat{e}_\alpha \, , \, \hat{e}_\beta \rangle = \langle B^\mu_\alpha \hat{e}_\mu \, , \, B^\nu_\beta \hat{e}_\nu \rangle = B^\mu_\alpha B^\nu_\beta \, g_{\mu\nu}.$$

We may also study the covariant derivative of vector fields in $T(\mathcal{S})$:

$$\begin{aligned} \nabla_\alpha \hat{e}_\beta &= \nabla_\alpha (B^\nu_\beta \hat{e}_\nu) = (\partial_\alpha B^\nu_\beta) \hat{e}_\nu + B^\nu_\beta \nabla_{\hat{e}_\alpha} \hat{e}_\nu \\ &= (\partial_\alpha B^\nu_\beta) \hat{e}_\nu + B^\nu_\beta B^\mu_\alpha \nabla_\mu \hat{e}_\nu = \left\{ \partial_\alpha B^\rho_\beta + B^\mu_\alpha B^\nu_\beta \Gamma_{\mu\nu}{}^\rho \right\} \hat{e}_\rho \end{aligned} \tag{A.14}$$

When $\nabla_\alpha \hat{e}_\beta$ lies entirely in $T(\mathcal{S})$ for any α and β, $\mathcal{S}$ is called a *flat* submanifold of $\mathcal{M}$. In general, however, $\nabla_\alpha \hat{e}_\beta$ has components orthogonal to $T(\mathcal{S})$, and thus cannot be viewed as a covariant derivative on $\mathcal{S}$. This may be cured by projection: the projection on $T(\mathcal{S})$ of a vector field in $T(\mathcal{M})$ is given by

$$\hat{X} = \langle X \, , \, \hat{e}_\alpha \rangle \, g^{\alpha\beta} \hat{e}_\beta.$$

By applying this projection to (A.14) we obtain a proper covariant derivative on $\mathcal{S}$:

$$\hat{\nabla}_\alpha \hat{e}_\beta = \left\{ \partial_\alpha B^\rho_\beta + B^\mu_\alpha B^\nu_\beta \Gamma_{\mu\nu}{}^\rho \right\} \langle \hat{e}_\rho \, , \, \hat{e}_\gamma \rangle \, g^{\gamma\delta} \hat{e}_\delta. \tag{A.15}$$

Noting that, given a metric, the components of an affine connection may be computed from the covariant derivative of the coordinate vectors:

$$\Gamma_{\mu\nu}{}^\rho = g^{\rho\lambda} \langle \nabla_\mu \hat{e}_\nu \, , \, \hat{e}_\lambda \rangle \, ,$$

(A.15) yields the following connection on $\mathcal{S}$:

$$\Gamma_{\alpha\beta}{}^\gamma \equiv g^{\gamma\delta} \left\langle \hat{\nabla}_\alpha \hat{e}_\beta \, , \, \hat{e}_\delta \right\rangle = g^{\gamma\delta} \langle \nabla_\alpha \hat{e}_\beta \, , \, \hat{e}_\delta \rangle = g^{\gamma\delta} \left\{ \partial_\alpha B^\rho_\beta + B^\mu_\alpha B^\nu_\beta \Gamma_{\mu\nu}{}^\rho \right\} B^\lambda_\delta g_{\rho\lambda} \, .$$

Note that a flat manifold may very well contain non-flat submanifolds, as is exemplified by the possibility of having a sphere as a submanifold of $\mathbb{R}^3$. Also note that affine flatness of $\mathcal{S}$ does *not* imply that $\mathcal{S}$ is a flat submanifold of $\mathcal{M}$, even if $\mathcal{M}$ is itself affine flat[3)]. On the other hand, a flat submanifold of an affine flat manifold *is* affine flat.

[1)] We shall write $\hat{e}_\alpha$ for the basis vectors of $T(\mathcal{S})$, relying on the choice of indices $\alpha, \beta, \ldots$ to distinguish them from the basis vectors $\hat{e}_\mu$ of $T(\mathcal{M})$. Note again that (A.13) is obvious if we identify $\hat{e}_\mu \equiv \frac{\partial}{\partial \theta^\mu}$, and $\hat{e}_\alpha \equiv \frac{\partial}{\partial \vartheta^\alpha}$.

[2)] or – equivalently – has full rank.

[3)] Consider for example a torus (which is flat with respect to the metric connection) as a submanifold of $\mathbb{R}^3$.

APPENDIX B

SOME SPECIAL FAMILIES OF PROBABILITY DISTRIBUTIONS

Two families of probability distributions that appear time and time again are the exponential family and the mixture family. We shall introduce them briefly and discuss the link with α-connections.

B.1 Exponential family

An exponential family of probability distributions is a family of distributions that can be written as

$$\mathcal{M} = \left\{ p_{\boldsymbol{\theta}} \,\middle|\, p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathrm{e}^{\sum_\mu \theta^\mu x_\mu + c_0(\boldsymbol{\theta})} P_0(\boldsymbol{x}) \right\},$$

where $P_0(\boldsymbol{x})$ is some fixed 'carrier' measure (which may be discrete or continuous) and c_0 ensures that the distribution is normalized.

An important subset of the exponential family is the set of Gaussian distributions: while in the normal way of writing there is an x^2 term in the exponent which seems to disqualify them, we may write

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

showing that in fact the normal distributions *do* form an exponential family.

We shall now compute the metric and α-connections for exponential families. We have:

$$\ell = \log p = \theta^k x_k + c_0(\boldsymbol{\theta}) + \log P(\boldsymbol{x}).$$

Therefore:

$$\partial_\mu \ell = x_\mu + \partial_\mu c_0(\boldsymbol{\theta}),$$

(where, as usual, $\partial_\mu \equiv \frac{\partial}{\partial \theta^\mu}$).

We thus have:

$$g_{\mu\nu} \equiv \int \mathrm{d}\boldsymbol{x}\; p\, \partial_\mu \ell\, \partial_\nu \ell = \int \mathrm{d}\boldsymbol{x}\; p \left[x_\mu x_\nu + x_\mu \partial_\nu c_0 + \partial_\mu c_0 x_\nu + \partial_\mu c_0 \partial_\nu c_0 \right].^{1)}$$

We see that

$$x_\mu p = \left(\partial_\mu \mathrm{e}^{\theta^k x_k} \right) \mathrm{e}^{c_0} P_0$$

[1)] Here, and in the following, the integration should be replaced by summation if P_0 is a discrete distribution.

and

$$x_\mu x_\nu p = \left(\partial_\mu \partial_\nu \mathrm{e}^{\theta^k x_k}\right) \mathrm{e}^{c_0} P_0.$$

Noting that $\int \mathrm{d}\boldsymbol{x}\; p = 1$ implies

$$\int \mathrm{d}\boldsymbol{x}\; \mathrm{e}^{\theta^k x_k} P_0(\boldsymbol{x}) = \mathrm{e}^{-c_0(\boldsymbol{\theta})},$$

we find:

$$\begin{aligned} g_{\mu\nu}(\boldsymbol{\theta}) &= \int \mathrm{d}\boldsymbol{x}\; \left[\partial_\mu \partial_\nu \mathrm{e}^{\theta^k x_k} + \partial_\mu \mathrm{e}^{\theta^k x_k}\, \partial_\nu c_0 + \partial_\mu c_0\, \partial_\nu \mathrm{e}^{\theta^k x_k} + \partial_\mu \partial_\nu c_0\right] e^{c_0(\boldsymbol{\theta})} P_0(\boldsymbol{x}) \\ &= \left[\partial_\mu \partial_\nu \mathrm{e}^{-c_0} + \partial_\mu \mathrm{e}^{-c_0}\, \partial_\nu c_0 + \partial_\mu c_0\, \partial_\nu \mathrm{e}^{-c_0} + \partial_\mu c_0\, \partial_\nu c_0\right] e^{c_0} \\ &= -\partial_\mu \partial_\nu c_0(\boldsymbol{\theta}). \end{aligned}$$

Similarly boring calculations show that

$$E(\partial_\mu \partial_\nu \ell\, \partial_\rho \ell) = 0,$$

while

$$E(\partial_\mu \ell\, \partial_\nu \ell\, \partial_\rho \ell) = -\partial_\mu \partial_\nu \partial_\rho c_0(\boldsymbol{\theta}).$$

Inserting these two into the definition (1.16), we find

$$\Gamma^{(\alpha)}_{\mu\nu\rho}(\boldsymbol{\theta}) = \frac{\alpha - 1}{2} \partial_\mu \partial_\nu \partial_\rho c_0(\boldsymbol{\theta}).$$

In particular, we see that the exponential family is $+1$-flat.

B.2 Mixture family

Any family of probability distributions that can be written as

$$\mathcal{M} = \left\{ p_{\boldsymbol{\theta}}(\boldsymbol{x}) \,\middle|\, p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta^\mu P_\mu(\boldsymbol{x}) + c_0(\boldsymbol{\theta}) P_0(\boldsymbol{x}) \right\}$$

is called a mixture family. In this definition, P_0 and each P_μ should be properly normalized probability distributions, each of the θ^μ should be in $[0,1]$, and c_0 is a normalization constant: $c_0 = 1 - \sum_\mu \theta^\mu$, which should also be in $[0,1]$.

We shall again try to compute the metric and α-connections. To this end, note that

$$\partial_\mu \ell(\boldsymbol{\theta}) = \frac{1}{p_{\boldsymbol{\theta}}} \partial_\mu p_{\boldsymbol{\theta}} = \frac{1}{p_{\boldsymbol{\theta}}} \left[P_\mu - P_0\right].$$

Therefore:

$$g_{\mu\nu}(\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{x}\; \frac{1}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} \left[P_\mu(\boldsymbol{x}) - P_0(\boldsymbol{x})\right] \left[P_\nu(\boldsymbol{x}) - P_0(\boldsymbol{x})\right].^{1)}$$

[1)] Again, integration should be replaced by summation if the P_μ are discrete distributions.

It does not seem possible to calculate this integral for general $\{P_\mu\}$. However, in one special case we may: namely when we limit ourselves to distributions over a countable set of atoms:

$$p_{\boldsymbol{\theta}}(x) = \sum_{\mu=1}^{N} \theta^\mu \delta_{x,a_\mu} + c_0(\boldsymbol{\theta})\delta_{x,a_0}, \tag{B.1}$$

where the same conditions on the θ^μ apply as before. As long as these conditions are met, N may be infinity.

In this case we find

$$\partial_\mu \ell(\boldsymbol{\theta}) = \frac{1}{p_{\boldsymbol{\theta}}}\partial_\mu p_{\boldsymbol{\theta}} = \frac{1}{p_{\boldsymbol{\theta}}}\left[\delta_{x,a_\mu} - \delta_{x,a_0}\right].$$

Therefore:

$$g_{\mu\nu}(\boldsymbol{\theta}) = \sum_x \frac{1}{p_{\boldsymbol{\theta}}}\partial_\mu p_{\boldsymbol{\theta}}\partial_\nu p_{\boldsymbol{\theta}} = \sum_{\rho=0}^{N} \frac{(\delta_{\rho\mu}-\delta_{\rho 0})(\delta_{\rho\nu}-\delta_{\rho 0})}{\sum_{\lambda=1}^{N}\theta^\lambda\delta_{\rho\lambda} + c_0(\boldsymbol{\theta})\delta_{\rho 0}}.$$

The numerator is always zero except when $\rho = 0$, or when $\rho = \mu = \nu$. We thus find:

$$g_{\mu\nu}(\boldsymbol{\theta}) = \frac{\delta_{\mu\nu}}{\theta^\mu} + \frac{1}{1-\sum_\lambda \theta^\lambda}, \tag{B.2}$$

a result first obtained by Čencov in 1972 [8].

Calculating the α-connections one stumbles across similar difficulties as for the metric, however for one special value of α these difficulties collapse:

$$\begin{aligned}\Gamma^{(\alpha)}_{\mu\nu\rho} &= \int \mathrm{d}\boldsymbol{x}\; p\left[\partial_\mu\partial_\nu\ell\,\partial_\rho\ell + \frac{1-\alpha}{2}\partial_\mu\ell\,\partial_\nu\ell\,\partial_\rho\ell\right] \\ &= \int \mathrm{d}\boldsymbol{x}\; p\left[\left(\frac{1}{p}\partial_\mu\partial_\nu p - \frac{1}{p^2}\partial_\mu p\,\partial_\nu p\right)\frac{1}{p}\partial_\rho p + \frac{1-\alpha}{2}\frac{1}{p^3}\partial_\mu p\,\partial_\nu p\,\partial_\rho p\right].\end{aligned}$$

Noting that $\partial_\mu\partial_\nu p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\partial}{\partial\theta^\mu}(P_\nu(\boldsymbol{x}) - P_0(\boldsymbol{x})) = 0$, this reduces to

$$\Gamma^{(\alpha)}_{\mu\nu\rho}(\boldsymbol{\theta}) = -\frac{1+\alpha}{2}\int \mathrm{d}\boldsymbol{x}\;\frac{1}{p_{\boldsymbol{\theta}}^2}\partial_\mu p_{\boldsymbol{\theta}}\,\partial_\nu p_{\boldsymbol{\theta}}\,\partial_\rho p_{\boldsymbol{\theta}},$$

and in particular $\Gamma^{(-1)}_{\mu\nu\rho} = 0$. The mixture family is thus found to be -1-flat.

As an aside we note that the calculation may be performed for general α if we limit ourselves to distributions over a countable set of atoms again: in the special case of (B.1) we find:

$$\begin{aligned}\Gamma^{(\alpha)}_{\mu\nu\rho}(\boldsymbol{\theta}) &= \frac{1+\alpha}{2}\sum_{\lambda=0}^{N}\frac{1}{p_{\boldsymbol{\theta}}^2}(\delta_{\lambda\mu}-\delta_{\lambda 0})(\delta_{\lambda\nu}-\delta_{\lambda 0})(\delta_{\lambda\rho}-\delta_{\lambda 0}) \\ &= \frac{1+\alpha}{2}\left[\delta_{\mu\nu}\delta_{\mu\rho}\left(\frac{1}{\theta^\mu}\right)^2 + \left(\frac{1}{1-\sum_\lambda\theta^\lambda}\right)^2\right].\end{aligned}$$

BIBLIOGRAPHY

[1] S. Amari, *Differential-geometrical methods in statistics*, Lecture notes in statistics, Springer-Verlag, Berlin, 1985.

[2] R. d'Inverno, *Introducing Einstein's relativity*, Clarendon Press, Oxford, 1992.

[3] J. M. Corcuera and F. Giummolè, *A characterization of monotone and regular divergences*, Accepted by Annals of the Institute of Statistical Mathematics, 1998.

[4] S. Amari, *Natural gradient works efficiently in learning*, Neural Computation **10** (1998) 251.

[5] Ma, Ji, and Farmer, *An efficient EM-based training algorithm for feedforward neural networks*, Neural Networks **10** (1997) 243.

[6] M. Rattray and D. Saad, *Transients and asymptotics of natural gradient learning*, Submitted to ICANN 98, 1998.

[7] A. Fujiwara and S. Amari, *Gradient systems in view of information geometry*, Physica **D 80** (1995) 317.

[8] N. Čencov, *Statistical decision rules and optimal inference*, Transl. Math. Monographs, vol. 53, Amer. Math. Soc., Providence USA, 1981, (translated from the Russian original, published by Nauka, Moscow in 1972).

[9] L. L. Campbell, *An extended Čencov characterization of the information metric*, Proc. AMS **98** (1986) 135.

[10] K. R. Parthasarathy, *Introduction to probability and measure*, MacMillan, Delhi, 1977.

[11] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, W. H. Freeman and company, San Fransisco, 1973.

[12] W. A. van Leeuwen, *Structuur van Ruimte-tijd, syllabus naar het college van G. G. A. Bäuerle en W. A. van Leeuwen, uitgewerkt door F. Hoogeveen*, in Dutch; obtainable from the University of Amsterdam, Institute for Theoretical Physics, Valckenierstraat 65, Amsterdam.

[13] R. H. Dijkgraaf, *Meetkunde, topologie en fysica, syllabus voor het college Wiskunde voor Fysici II*, in Dutch; obtainable from the University of Amsterdam, Dept. of Mathematics, Plantage Muidergracht 24, Amsterdam.