Differential Geometry in Statistics

Peadar Coyle

August 26, 2012

Abstract

Some notes on Statistics and Differential Geometry, merely for personal consumption

1 Introduction

Statistics is a science which studies methods of inference, from observed data, concerning the probabilistic structure underlying such data. The class of all the possible probability distributions is usually too wide to consider all its elements as

This a slightly misleading name for applying differential geometry to families of probability distributions, and so to statistical models. Information does however play two roles in it: Kullback-Leibler information, or relative entropy, features as a measure of divergence (not quite a metric, because it's asymmetric), and Fisher information takes the role of curvature. One very nice thing about information geometry is that it gives us very strong tools for proving results about statistical models, simply by considering them as well-behaved geometrical objects. Thus, for instance, it's basically a tautology to say that a manifold is not changing much in the vicinity of points of low curvature, and changing greatly near points of high curvature. Stated more precisely, and then translated back into probabilistic language, this becomes the Cramer-Rao inequality, that the variance of a parameter estimator is at least the reciprocal of the Fisher information. As someone who likes differential geometry, and now is interested in statistics, I find this very pleasing.

As a physicist, I have always been somewhat bothered by the way statisticians seem to accept particular parametrizations of their models as obvious and natural, and build those parameterization into their procedures. In linear regression, for instance, it's reasonably common for them to want to find models with only a few non-zero coefficients. This makes my thumbs prick, because it seems to me obvious that if I regressed on arbitrary linear combinations of my covariates, I have exactly the same information (provided the transformation is invertible), and so I'm really looking at exactly the same model — but in general I'm not going to have a small number of non-zero coefficients any more. In other words, I want to be able to do coordinate-free statistics. Since differential

geometry lets me do coordinate-free physics, information geometry seems like an appealing way to do this. There are various information-geometric model selection criteria, which I want to know more about; I suspect, based purely on this disciplinary prejudice, that they will out-perform coordinate-dependent criteria.

I should also mention that statistical physics, while it does no actual statistics, is also very much concerned with probability distributions. Sun-Ichi Amari, who is the leader of a large and impressive Japanese school of information-geometers, has a nice result (in, e.g., his "Hierarchy of Probability Distributions" paper) showing that maximum entropy distributions are, exactly, the ones with minimal interaction between their variables — the ones which approach most closely to independence. Only local properties of a statistical model are responsible for the asymptotic theory of statistical inference. Local properties are represented by the geometry of the tangent spaces of the manifold. The tangent space has a natural Riemannian metric given by the *Fisher information matrix* in the regular case. It represents on a local property of the model, because the tangent space is nothing but local linearization of the model manifold. In order to obtain larger-scale properties, one needs to define mutual relations of the two different tangent spaces at two neighboring points in the model. This can be done by defining a one-to-one affine correspondence between two tangent spaces, which is called an affine connection in differential geometry. By an affine connection, one can consider local properties around each point beyond the linear approximation. The curvature of a statistical model can be otained by the use of this connection. It is clear that such a differential-geometrical concept provides a tool convenient for studying higher-order asymptotic properties of inference. However, by connecting local tangent spaces further, one can obtain global relations. Hence, the validity of the differential-geometrical method is not limited within the framework of asymptotic theory. It was Rao (1945) who first pointed out the importance of the differential-geometric approach. He introduced the Riemannian metric by using the Fisher information matrix.

2 Geometrical Structure of Statistical Models

2.1 Metric and a-connection

Let S = $(p(x, \theta))$ be a statistical model consisting of probability density functions $p(x, \theta)$ of random variable $x \in X$ with respect to a measure $\mathbb{P}$ on X such that every distribution is uniquely parameterized by an n-dimensional vector parameter $\theta = (\theta^i) = (\theta^1, \cdots, \theta^n)$. Since the set $\{p(x)\}$ of all the density functions on X is a subset of the L_1 space of functions in x, S is considered to be a subset of the L_1 space. A statistical model S is said to be geometrically regular, when it satisfied the following regularity conditions, and S is regarded as an n-dimensional manifold with a coordinate system θ.

1. The domain Θ of the parameter θ is homeomorphic to an n-dimensional Euclidean space $\mathbb{R}^n$.

2. The topology of S induced from $\mathbb{R}^n$ is compatible with the relative topology of S in the L_1 space.

3. The support of p(x,θ) is common for all $\theta \in \Theta$, so that $p(x,\theta)$ are mutually absolutely continuous.

4. Every density function $p(x,\theta)$ is a smooth function in θ uniformly in x, and the partial derivative $\dfrac{\partial}{\partial\theta^i}$ and integration $\log p(x,\theta)$ with respect to the measure P(x) are always commutative.

5. The moments of the score function $\dfrac{\partial}{\partial\theta^i}\log p(x,\theta)$ exist up to the third order and are smooth in θ.

6. The Fisher information matrix is positive definite.

Let us denote by $\partial_i = \dfrac{\partial}{\partial\theta^i}$ the tangent vector e_i of the i-th coordiate curve θ^i (Fig. 1) at point θ/ Then, n such tangent vectors $e_i = \partial_i$, i = 1, $\cdots$, n, span the tangent space T_θ at point θ of the manifold S. Any tangent vector $A \in T_\theta$ is a linear combination of the basis vectors ∂_i,

$$A = A^i\partial_i,$$

where A^i are the components of vector A and Einstein's summation convention is assumed throughout the paper. The tangent space T_θ is a linearized version of a small neighbourhood at θ of S, and an infinitesimal vector $d\theta = d\theta^i\partial_i$ denotes the vector connecting two neighbouring points θ and $\theta + d\theta$ or two neighbouring distributions $p(x,\theta)$ and $p(x,\theta + d\theta)$.
Let us introduce a metric in the tangent space T_θ. It can be done by defining the inner product $g_{ij}(\theta) = \langle\partial_i, \partial_j\rangle$ of two basis vectors ∂_i and ∂_j at θ. To this end, we represent a vector $\partial_i \in T_\theta$ by a function $\partial_i l(x,\theta)$ in x, where x(l,θ) = $\log p(x,\theta)$ and ∂_i (in $\partial_i l$) is the partial derivative $\dfrac{\partial}{\partial\theta^i}$. Then, it is natural to define the inner product by

$$g_ij(\theta) = \langle\partial_i, \partial_j\rangle = E_\theta[\partial_i l(x,\theta)\partial_j l(x,\theta)], \tag{1}$$

where E_θ denotes the expectation with respect to $p(x,\theta)$. This g_{ij} is the Fisher information matrix. Two vectors A and B are orthogonal when $\langle A, B\rangle = 0$. It is sometimes necessary to companre a vector $A \in T_\theta$ of the tangent space T_θ at one point θ with a vector $B \in T_{\theta'}$, belonging to the tangent space $T_{\theta'}$ at another point θ'. This can be done by comparing the basis vectors ∂_i at T_θ with the basis vectors ∂_i' at $T_{\theta'}$. Since T_θ and $T_{\theta'}$, are two different vector spaces, the two vectors ∂_i and ∂_i' are not directly comparable, and we need some way of identifying T_θ and $T_{\theta'}$ in order to compare the vectors in them. This can be accomplished by introducing an affine connection, which maps a tangent space $T_{\theta+d\theta}$ at $\theta + d\theta$ to the tangent space T_θ at θ. The mappting should reduce to

the identity map as $d\theta \to 0$. Let $m\left(\partial'_j\right)$ be the image of $\partial'_j \in T_{\theta+d\theta}$ mapped to T_θ. It is slightly different from $\partial_j \in T_\theta$. The vector

$$\nabla_{\partial_i}\partial_j = \lim_{d\theta\to 0}\frac{d}{d\theta^i}\left\{m(\partial'_j) - \partial_j\right\}$$

represents the rate at which the j-th basis vector $\partial_j \in T_\theta$ 'intrinsically' changes as the point θ moves from θ to $\theta + d\theta$ in the direction ∂_i. We call $\nabla_{\partial_i}\partial_j$ the covariant derivative of the basis vector ∂_j in the direction ∂_i. Since it is a vector of T_θ, its components are given by

$$\Gamma_{ijk} = \langle \nabla_{\partial_i}\partial_j, \partial_k\rangle \tag{2}$$

and $\nabla_{\partial_i}\partial_j = \Gamma^k_{ij}\partial_k$ where $\Gamma_{ijk} = \Gamma^m_{ij}g_{mk}$. We call Γ_{ijk} the components of the affine connection. An affine connection is specified by defining $\nabla_{\partial_i}\partial_j$ or Γ_{ijk}. Let A(θ) be a vector field, which assigns to every point $\theta \in S$ a vector $A(\theta) = A^i(\theta)\partial_i \in T_\theta$. The intrinsic change of the vector $A(\theta)$ as the position θ moves is now given by the covariant derivative in the direction ∂_i of $A(\theta) = A^j(\theta)\partial_j$, defined by

$$\nabla_{\partial_i}A = (\partial_i A^j)\partial_j + A^j(\nabla_{\partial_i}\partial_j) = (\partial_i A^j + \Gamma^j_{ik}A^k)\partial_j \tag{3}$$

in which the change in the basis vectors as well as that in the components $A^i(\theta)$ is taken into account. The covariant derivative in the direction B $= B^i\partial_i$ is given by

$$\nabla_B A = B^i_{\nabla_{\partial_i}}A$$

We have defined the covariant derivative by the use fo the basis vectors ∂_i which are associated with the coordinate system or the parameterization θ. However, the covariant derivative $\nabla_B A$ is invariant under any parameterization, given the same result in any coordinate system. This yields a transformation law for the components of a connection Γ_{ijk}. When another coordinate system $\theta' = \theta'(\theta)$ is used, the basis vectors change from $\{\partial_i\}$ to $\{\partial'_{i'}\}$, where

$$\partial'_{i'} = B^i_{i'}\partial_i,$$

and $B^i_{i'} = \dfrac{\partial\theta^i}{\partial\theta'^{i'}}$ is the inverse matrix of the Jacobian matrix of the coordinate transformation. Since the components $\Gamma'_{i'j'k'}$ of the connection are written as

$$\Gamma'_{i'j'k'} = \langle \nabla\partial_{i'}\partial_{j'}, \partial_{k'}\rangle$$

in this new coordinate system, we easily have the transformation law

$$\Gamma'_{i'j'k'} = B^i_{i'}B^j_{j'}B^k_{k'}\Gamma_{ijk} + B^i_{i'}B^k_{k'}g_{kj}(\partial_i)B^j_{j'})$$

We introduce the α-connection, where α is a real parameter, in the statistical manifold S by the formula.

$$\Gamma^\alpha_{ijk} = E_\theta[\{\partial_i\partial_j l(x,\theta) + \frac{1-\alpha}{2}\partial_i l(x,\theta)\partial_j l(x,\theta)\}\partial_k l(x,\theta)] \tag{4}$$

It is easily checked that the connection defined by 4 satisfies the transformation law. In particular the 1- connection is called the exponential connection, and the -1-connection is called the mixture connection.

2.2 Imbedding and a-curvature

The heading should really be α-curvature. Let us consider an m-dimensional regular statistical model M = $\{q(x,u)\}$, which is inbedded in S = $\{p(x,\theta)\}$ by

$$q(x,u) = p\left\{x, \theta(u)\right\}.$$

Here, u = $(u^a) = \left(u^1, \cdots, u^m\right)$ is a vector parameter specifying distributions of M, and defines a coordinate system of M. We assume that $\theta = \theta(u)$ is smooth and its Jacobian matrix has a full rank. Moreover, it is assumed that M forms an m-dimensional submanifold in S. We identify a point $u \in M$ with the point $\theta = \theta(u)$ imbedded in S. The tangent space $T_u(M)$ at u of M is spanned by m vectors ∂_a, $a = 1, \cdots, m$, where $\partial_a = \dfrac{\partial}{\partial u^a}$ denotes the tangent vector of the coordinate curve u^a in M. The basis ∂_a can be represented by a function $\partial_a l(x,u)$ in x as before, where $l(x,u) = \log q(x,u)$. Since M is imbedded in S, the tangent space $T_u(M)$ of M is regarded as a subspace of the tangent space $T_{\theta(u)}(S)$ of S at $\theta = \theta(u)$. The basis vector $\partial_a \in T_u(M)$ is written as a linear combination of ∂_i,

$$\partial_a = B_a^i(u)\partial_i$$

where $B_a^i = \frac{\partial \theta^i(u)}{\partial u^a}$. This can be understood from the relation

$$\partial_a = B_a^i(u)\partial_i,$$

where $B_a^i = \frac{\partial \theta^i}{\partial u^a}$. This can be understood form the relation

$$\partial_a l(x,u) = B_a^i \partial_i l\left\{x, \theta(u)\right\}.$$

Hence, the tangential directions of M at u is represented by m vectors ∂_a, (a = 1,$\cdots$, m) or $B_a = (B_a^i)$ in the component form with respect to the basis ∂_i of $T_{\theta(u)}(S)$ such that n vectors $\{\partial_a, \partial_\kappa\}$, a = 1,...,m; $\kappa = m+1, \cdots, n$, together form a basis of $T_{\theta(u)}(S)$ and moreover ∂_κ's are orthogonal to ∂_a's

$$g_{a\kappa}(u) = \langle \partial_a, \partial_\kappa \rangle = 0$$

The vectors ∂_κ span the orthogonal complement $\partial_\kappa \in T_u^\perp(M)$ where $\perp$ denotes the orthogonal complement of $T_u(M)$ in $T_{\theta(U)}(S)$.

$$g_{ab}(u) = \langle \partial_a, \partial_b \rangle = B_a^i B_b^j g_{ij} \tag{5}$$

$$g_{a\kappa}(u) = \langle \partial_a, \partial_\kappa \rangle = B_a^i B_\kappa^j g_{ij} \tag{6}$$

$$g_{\kappa\lambda}(u) = \langle \partial_\kappa, \partial_\lambda \rangle = B_\kappa^i B_\lambda^k g_{ik} \tag{7}$$

The basis vector ∂_a may change its direction as point u moves in M. The change is measured by the α- covariant derivative $\nabla^{(\alpha)}_{\partial_b}\partial_a$ is calculated in S as

$$\nabla^{(\alpha)}_{\partial_b}\partial_a = B^i_b \nabla^{(\alpha)}_{\partial_i}(B^j_a \partial_j)$$

$= (\partial_b B^j_a + B^i_b B^k_a \Gamma_i k^{(\alpha)j})\partial_j$. When the directions of the tangent space $T_u(M)$ of M do not change as point u moves in M, the manifold M is said to be α-flat in S, where the tangent directions are compared by the α-connection. Otherwise, M is curved in the sense of the α-connection. The α-covariant derivative $\nabla^{(\alpha)}_{\partial_b}\partial_a$ is decomposed into the tangential component belonging to $T_u(M)$ and the normal component perpendicular to $T_u(M)$. The former component represents the way ∂_a changes within $T_u(M)$, while the latter represents the change of ∂_a in the directions perpendicular to $T_u(M)$, as u moves in M. The normal component is measured by

$$H^{(\alpha)}_{ab\kappa} = \langle \nabla^{(\alpha)}_{\partial_a}\partial_b, \partial_\kappa \rangle = \left(\partial_b B^j_a + B^i_b B^k_a \Gamma_i k^{(\alpha)j}\right) B^m_\kappa g_{mj} \tag{8}$$

which is a tensor called the α-curvature of submanifold M in S. It is usually called the imbedding curvature or Euler-Shouten curvature. This tensor represents how M is curved in S. A tensor is a multi-linear mapping from the number of tangent vectors (or possibly one forms) to the real set. In the present case, for A $= A^a\partial_a \in T_u(M)$, B$= B^b\partial_b \in T_u(M)$ and C$=C^\kappa\partial_\kappa \in T^\perp_u(M)$, we have the multi-linear mapping $H^{(\alpha)}$,

$$H^{(\alpha)}(A, B, C) = H^{(\alpha)}_{ab\kappa} A^a B^b C^\kappa$$

This $H^{(\alpha)}$ is the α-curvature tensor, and $H^{(\alpha)}_{ab\kappa}$ are its components. The submanifold M is α-flat in S when $H^{(\alpha)}_{ab\kappa} = 0$ holds. The m x m matrix

$$[H^{(\alpha)}_M]^2_a b = H^{(\alpha)} = H^{(\alpha)}_{ac\kappa} H^{(\alpha)}_{bd\lambda} g^{\kappa\lambda} g^{cd}$$

represents the square of the α-curvature of M, where $g^{\kappa\lambda}$ and g^{cd} are the inverse matrix of $g_{\kappa\lambda}$ and g_{cd}, respectively. Efron called the scalar

$$\gamma^2 = [H^{(l)}_M]^2_a b g^a b$$

the statistical curvature in a one-dimensional model M, whcih is the trace of the square of the exponential- or l-curvature of M in our terminology. Let $\theta = \theta(t)$ be a curve in S parameterized by a scalar t. The curve c: $\theta = \theta(t)$ forms a one dimensional submanifold in S. The tangent vector ∂_t of the curve is represented in component form as

$$\partial_t = \dot{\theta}^i(t)\partial_i$$

or shortly by $\dot{\theta}$ where $\cdot$ denotes the first derivative with respect tot he time. When the direction of the tangent vector $\partial_t = \dot{\theta}$ does not change along the curve in the sense of the α-connection, the curve is called an α-geodesic. By

choosing an appropriate parameter, an α-geodesic $\theta(t)$ satisfies the geodesic equation

$$\nabla^{(\alpha)}_{\dot{\theta}}\dot{\theta} = 0$$

or in component form

$$\ddot{\theta}^i + \Gamma^{(\alpha)i}_{jk}\dot{\theta}^j\dot{\theta}^k = 0 \tag{9}$$

3 Parametric estimation and the Cramer-Rao inequality

Information geometry had its roots in Fishers theory of estimation. Let $\rho_\nu(x)$,$x \in \mathbb{R}$, be a strictly positive differentiable probability density, depending on a parameter $\nu \in \mathbb{R}$. To stress the analogy between the classical case and quantum case a density is also referred to as a state. The Fisher information of ρ_ν u is defined to be (Fisher 1925)

$$G := \int \rho_\nu(x) \left(\frac{\partial \log \rho_\nu(x)}{\partial \nu} \right)^2 dx.$$

We note that this is the variance of the random variable $Y = \frac{\partial \log \rho_\nu}{\partial \nu}$, which has mean zero. Furthermore, G is associated with the family $\mathcal{M} = \{\rho_\nu\}$ of distributions, rather than any one of them. This concept arises in the theory of estimation as follows. Let X be a random variable whose distribution is believed or hoped to be one of those in M. We estimate the value of u by measuring X independently m times, getting the data $x_1 \cdots , x_m$. An *estimator* f is a function of $(x_1, ..., x_m)$ that is used for this estimate. So f is a function of m independent copies of X, and so is a random variable. To be useful, the estimator must be a known function of X, not depending of ν, which we do not (yet) know. We say that an estimator is unbiased if its mean is the desired parameter; it is usual to take f as a function of X and to regard f $(X_i$), $i = 1, \cdots , m$ as samples of f . Then the condition that f is unbiased becomes

$$\rho_\nu(x) \cdot f := \int \rho_\nu(x) f(x) dx = \nu$$

A good estimator should also have only a small chance of being far from the correct value, which is its mean if it is unbiased. This chance is measured by the variance. (Fisher 1925) proved that the variance V of an unbiased estimator f obeys the inequality $V \geq G^{-1}$. This is called the CramèrRao inequality and its proof is based on the CauchySchwarz inequality. We shall show how this is done. If we do N independent measurements for the estimator, and average them, we improve the inequality to $V \geq G^{-1}$ /N . This inequality expresses that, given the family ρ_ν, there is a limit to the reliability with which we can estimate ν. Fisher termed VG^1 the efficiency of the estimator f . Equality in the Schwarz inequality occurs if and only if the two functions are proportional.

In this case, let $\partial\xi/\partial\nu$ denote the factor of proportionality. Then the optimal estimator occurs when

$$\log \rho_\nu(x) = -\int \frac{\partial \xi}{\partial \xi}(f(x) - \nu)d\nu$$

Doing the integral, and adjusting the integration constant by normalisation, leads to

$$\rho_\nu(x) = Z^{-1} \exp\{-\xi f(x)\}$$

which defines the exponential family. This can be generalised to any n-parameter manifold $\mathcal{M} = \rho_\nu$of distributions, $\nu = (\nu_1, \cdots, \nu_n)$ $\nu \in \mathbb{R}^n$. Suppose we have unbiased estimators $(X_1, ..., X_n)$, with covariance matrix V . Fisher introduced the information matrix

$$G^{ij} = \int \rho_\nu(x) \frac{\partial \log \rho_\nu(x)}{\partial \nu_i} \frac{\partial \log \rho_\nu(x)}{\partial \nu_j} dx \tag{10}$$

This looks like a Riemannian metric on $\mathcal{M}$. Cramer and Rao obtained hte analogue of the inequality $V \geq G^{-1}$ when $n > 1$ Put $V_{ij} = \rho_\nu \cdot [(X_i - \nu_i)(X_j - \nu_j)]$ the covariance matrix of the estimators $\{X_i\}$, $i = 1, \cdots, n$, and $Y^i = \dfrac{\partial \rho_\nu}{\partial \nu_i}$. We say that the estimators are locally unbiased if

$$\int \rho_\nu(x) Y^i(x)(X_j(x) - \nu_j)dx = \delta_i j \tag{11}$$

Then we get the CramerRao matrix inequality $V \geq G^{-1}$ as a matrix. For, equation 11 shows that the covariance of $X_j$with Y_i is δ_{ij} , so the covariance matrix of X_j and Y^i is

$$K := \begin{vmatrix} V & I \\ I & G \end{vmatrix} \tag{12}$$

It follows that the matrix 12 is positive semi-definite;

A Definitions

Throughout this article, boldfaced unsubscripted $\mathbf{X}$ and $\mathbf{Y}$ are used to refer to random vectors, and unboldfaced subscripted X_i and Y_i are used to refer to random scalars. If the entries in the column vector

:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

are random variables, each with finite variance, then the covariance matrix Σ is the matrix whose (i,j) entry is the covariance

:

$$\Sigma_{ij} = \operatorname{cov}(X_i, X_j) = \operatorname{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

where
:

$$\mu_i = \mathrm{E}(X_i)$$

is the expected value of the **i**th entry in the vector **X**. In other words, we have

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

The inverse of this matrix,

$$\Sigma^{-1}$$

, if it exists, is the *inverse covariance matrix* The elements of the precision matrix have an interpretation in terms of partial correlations and partial variances
=== Generalization of the variance ===
The definition above is equivalent to the matrix equality

$$\Sigma = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])\,(\mathbf{X} - \mathrm{E}[\mathbf{X}])^{\mathrm{T}}\right]$$

This form can be seen as a generalization of the scalar-valued variance to higher dimensions. Recall that for a scalar-valued random variable **X**

$$\sigma^2 = \mathrm{var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2] = \mathrm{E}[(X - \mathrm{E}(X)) \cdot (X - \mathrm{E}(X))].$$

Indeed, the entries on the diagonal of the covariance matrix Σare the variances of each element of the vector **X**

B What is a Geodesic

Let us recall the definition of a geodesic. A **geodesic** on a smooth manifold M with an **affine connection** ∇ is defined as a curve $\gamma(t)$ such that **parallel transport** along the curve preserves the tangent vector to the curve, so

$$\nabla_{\dot\gamma}\dot\gamma = 0 \tag{13}$$

at each point along the curve, where $\dot\gamma$ is the derivative w.r.t. t. More precisely, in order to define the covariant derivative of $\dot\gamma$ to a continuously differentiable vector field on an open set. However, the resulting value of 13 is independent of the choice of extension. We can however write this in local form,

$$\frac{d^2x^\lambda}{dt^2} + \Gamma^\lambda_{\mu\nu}\frac{dx^\mu}{dt}\frac{dx^\nu}{dt} = 0,$$

where $x^\mu(t)$ are the coordinates of the curve $\gamma(t)$ and $\Gamma^\lambda_{\mu\nu}$ are the Christoffel symbols of the connection ∇. This is just an ODE for the coordinates. It has a unique solution, given an initial position and an initial velocity. Therefore, fromt he point of view of classical mechainics, geodesics can be thought of as trajectories of **free particles** in a manifold. Indeed, the equation $\nabla_{\dot\gamma}\dot\gamma = 0$ means that the acceleration of the curve has no components in the direction of the surface (the direction of the surface is the same as the direction of the affine connection). We can also say that the acceleration of the curve is perpendicular to the tangent plane of the surface at each point of the curve. So the motion is completely determined by the bending of the surface. This also the idea of General Relativity where particles move on geodesics and the bending is caused by gravity.