Chapter 13

Curvature in Riemannian Manifolds

13.1 The Curvature Tensor

If $(M, \langle -, - \rangle)$ is a Riemannian manifold and ∇ is a connection on M (that is, a connection on TM), we saw in Section 11.2 (Proposition 11.8) that the curvature induced by ∇ is given by

$$R(X, Y) = \nabla_X \circ \nabla_Y - \nabla_Y \circ \nabla_X - \nabla_{[X,Y]},$$

for all $X, Y \in \mathfrak{X}(M)$, with $R(X, Y) \in \Gamma(\mathcal{H}om(TM, TM)) \cong \mathrm{Hom}_{C^\infty(M)}(\Gamma(TM), \Gamma(TM))$. Since sections of the tangent bundle are vector fields ($\Gamma(TM) = \mathfrak{X}(M)$), R defines a map

$$R\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \longrightarrow \mathfrak{X}(M),$$

and, as we observed just after stating Proposition 11.8, $R(X, Y)Z$ is $C^\infty(M)$-linear in X, Y, Z and skew-symmetric in X and Y. It follows that R defines a $(1, 3)$-tensor, also denoted R, with

$$R_p\colon T_pM \times T_pM \times T_pM \longrightarrow T_pM.$$

Experience shows that it is useful to consider the $(0, 4)$-tensor, also denoted R, given by

$$R_p(x, y, z, w) = \langle R_p(x, y)z, w \rangle_p$$

as well as the expression $R(x, y, y, x)$, which, for an orthonormal pair, of vectors (x, y), is known as the *sectional curvature*, $K(x, y)$.

This last expression brings up a dilemma regarding the choice for the sign of R. With our present choice, the sectional curvature, $K(x, y)$, is given by $K(x, y) = R(x, y, y, x)$ but many authors define K as $K(x, y) = R(x, y, x, y)$. Since $R(x, y)$ is skew-symmetric in x, y, the latter choice corresponds to using $-R(x, y)$ instead of $R(x, y)$, that is, to define $R(X, Y)$ by

$$R(X, Y) = \nabla_{[X,Y]} + \nabla_Y \circ \nabla_X - \nabla_X \circ \nabla_Y.$$

As pointed out by Milnor [106] (Chapter II, Section 9), the latter choice for the sign of R has the advantage that, in coordinates, the quantity, $\langle R(\partial/\partial x_h, \partial/\partial x_i)\partial/\partial x_j, \partial/\partial x_k \rangle$ coincides

with the classical Ricci notation, R_{hijk}. Gallot, Hulin and Lafontaine [60] (Chapter 3, Section A.1) give other reasons supporting this choice of sign. Clearly, the choice for the sign of R is mostly a matter of taste and we apologize to those readers who prefer the first choice but we will adopt the second choice advocated by Milnor and others. Therefore, we make the following formal definition:

Definition 13.1 Let $(M, \langle -, - \rangle)$ be a Riemannian manifold equipped with the Levi-Civita connection. The *curvature tensor* is the $(1,3)$-tensor, R, defined by

$$R_p(x,y)z = \nabla_{[X,Y]}Z + \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z,$$

for every $p \in M$ and for any vector fields, $X, Y, Z \in \mathfrak{X}(M)$, such that $x = X(p)$, $y = Y(p)$ and $z = Z(p)$. The $(0,4)$-tensor associated with R, also denoted R, is given by

$$R_p(x,y,z,w) = \langle (R_p(x,y)z, w \rangle,$$

for all $p \in M$ and all $x, y, z, w \in T_pM$.

Locally in a chart, we write

$$R\left(\frac{\partial}{\partial x_h}, \frac{\partial}{\partial x_i}\right)\frac{\partial}{\partial x_j} = \sum_l R^l_{jhi}\frac{\partial}{\partial x_l}$$

and

$$R_{hijk} = \left\langle R\left(\frac{\partial}{\partial x_h}, \frac{\partial}{\partial x_i}\right)\frac{\partial}{\partial x_j}, \frac{\partial}{\partial x_k}\right\rangle = \sum_l g_{lk} R^l_{jhi}.$$

The coefficients, R^l_{jhi}, can be expressed in terms of the Christoffel symbols, Γ^k_{ij}, in terms of a rather unfriendly formula (see Gallot, Hulin and Lafontaine [60] (Chapter 3, Section 3.A.3) or O'Neill [119] (Chapter III, Lemma 38). Since we have adopted O'Neill's conventions for the order of the subscripts in R^l_{jhi}, here is the formula from O'Neill:

$$R^l_{jhi} = \partial_i \Gamma^l_{hj} - \partial_h \Gamma^l_{ij} + \sum_m \Gamma^l_{im}\Gamma^m_{hj} - \sum_m \Gamma^l_{hm}\Gamma^m_{ij}.$$

There is another way of defining the curvature tensor which is useful for comparing second covariant derivatives of one-forms. Recall that for any fixed vector field, Z, the map, $Y \mapsto \nabla_Y Z$, is a $(1,1)$ tensor that we will denote $\nabla_- Z$. Thus, using Proposition 11.5, the covariant derivative $\nabla_X \nabla_- Z$ of $\nabla_- Z$ makes sense and is given by

$$(\nabla_X(\nabla_- Z))(Y) = \nabla_X(\nabla_Y Z) - (\nabla_{\nabla_X Y})Z.$$

Usually, $(\nabla_X(\nabla_- Z))(Y)$ is denoted by $\nabla^2_{X,Y} Z$ and

$$\nabla^2_{X,Y} Z = \nabla_X(\nabla_Y Z) - \nabla_{\nabla_X Y} Z$$

is called the *second covariant derivative* of Z with respect to X and Y. Then, we have

$$\begin{aligned}\nabla^2_{Y,X}Z - \nabla^2_{X,Y}Z &= \nabla_Y(\nabla_X Z) - \nabla_{\nabla_Y X}Z - \nabla_X(\nabla_Y Z) + \nabla_{\nabla_X Y}Z\\ &= \nabla_Y(\nabla_X Z) - \nabla_X(\nabla_Y Z) + \nabla_{\nabla_X Y - \nabla_Y X}Z\\ &= \nabla_Y(\nabla_X Z) - \nabla_X(\nabla_Y Z) + \nabla_{[X,Y]}Z\\ &= R(X,Y)Z,\end{aligned}$$

since $\nabla_X Y - \nabla_Y X = [X,Y]$, as the Levi-Civita connection is torsion-free. Therefore, the curvature tensor can also be defined by

$$R(X,Y)Z = \nabla^2_{Y,X}Z - \nabla^2_{X,Y}Z.$$

We already know that the curvature tensor has some symmetry properties, for example, $R(y,x)z = -R(x,y)z$ but when it is induced by the Levi-Civita connection, it has more remarkable properties stated in the next proposition.

Proposition 13.1 *For a Riemannian manifold, $(M, \langle -,-\rangle)$, equipped with the Levi-Civita connection, the curvature tensor satisfies the following properties:*

(1) $R(x,y)z = -R(y,x)z$

(2) (First Bianchi Identity) $R(x,y)z + R(y,z)x + R(z,x)y = 0$

(3) $R(x,y,z,w) = -R(x,y,w,z)$

(4) $R(x,y,z,w) = R(z,w,x,y)$.

The proof of Proposition 13.1 uses the fact that $R_p(x,y)z = R(X,Y)Z$, for any vector fields X, Y, Z such that $x = X(p)$, $y = Y(p)$ and $Z = Z(p)$. In particular, X, Y, Z can be chosen so that their pairwise Lie brackets are zero (choose a coordinate system and give X, Y, Z constant components). Part (1) is already known. Part (2) follows from the fact that the Levi-Civita connection is torsion-free. Parts (3) and (4) are a little more tricky. Complete proofs can be found in Milnor [106] (Chapter II, Section 9), O'Neill [119] (Chapter III) and Kuhnel [91] (Chapter 6, Lemma 6.3).

If $\omega \in \mathcal{A}^1(M)$ is a one-form, then the covariant derivative of ω defines a $(0,2)$-tensor, T, given by $T(Y,Z) = (\nabla_Y \omega)(Z)$. Thus, we can define the second covariant derivative, $\nabla^2_{X,Y}\omega$, of ω as the covariant derivative of T (see Proposition 11.5), that is,

$$(\nabla_X T)(Y,Z) = X(T(Y,Z)) - T(\nabla_X Y, Z) - T(Y, \nabla_X Z),$$

and so

$$\begin{aligned}(\nabla^2_{X,Y}\omega)(Z) &= X((\nabla_Y\omega)(Z)) - (\nabla_{\nabla_X Y}\omega)(Z) - (\nabla_Y\omega)(\nabla_X Z)\\ &= X((\nabla_Y\omega)(Z)) - (\nabla_Y\omega)(\nabla_X Z) - (\nabla_{\nabla_X Y}\omega)(Z)\\ &= (\nabla_X(\nabla_Y\omega))(Z) - (\nabla_{\nabla_X Y}\omega)(Z).\end{aligned}$$

Therefore,

$$\nabla^2_{X,Y}\omega = \nabla_X(\nabla_Y\omega) - \nabla_{\nabla_X Y}\omega,$$

that is, $\nabla^2_{X,Y}\omega$ is formally the same as $\nabla^2_{X,Y}Z$. Then, it is natural to ask what is

$$\nabla^2_{X,Y}\omega - \nabla^2_{Y,X}\omega.$$

The answer is given by the following proposition which plays a crucial role in the proof of a version of Bochner's formula:

Proposition 13.2 *For any vector fields, $X, Y, Z \in \mathfrak{X}(M)$, and any one-form, $\omega \in \mathcal{A}^1(M)$, on a Riemannian manifold, M, we have*

$$((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)(Z) = \omega(R(X,Y)Z).$$

Proof. Recall that we proved in Section 11.5 that

$$(\nabla_X\omega)^\sharp = \nabla\omega^\sharp.$$

We claim that we also have

$$(\nabla^2_{X,Y}\omega)^\sharp = \nabla^2_{X,Y}\omega^\sharp.$$

This is because

$$\begin{aligned}
(\nabla^2_{X,Y}\omega)^\sharp &= (\nabla_X(\nabla_Y\omega))^\sharp - (\nabla_{\nabla_X Y}\omega)^\sharp \\
&= \nabla_X(\nabla_Y\omega)^\sharp - \nabla_{\nabla_X Y}\omega^\sharp \\
&= \nabla_X(\nabla_Y\omega^\sharp) - \nabla_{\nabla_X Y}\omega^\sharp \\
&= \nabla^2_{X,Y}\omega^\sharp.
\end{aligned}$$

Thus, we deduce that

$$((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)^\sharp = (\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega^\sharp = R(Y,X)\omega^\sharp.$$

Consequently,

$$\begin{aligned}
((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)(Z) &= \langle((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)^\sharp, Z\rangle \\
&= \langle R(Y,X)\omega^\sharp, Z\rangle \\
&= R(Y,X,\omega^\sharp,Z) \\
&= R(X,Y,Z,\omega^\sharp) \\
&= \langle R(X,Y)Z, \omega^\sharp\rangle \\
&= \omega(R(X,Y)Z),
\end{aligned}$$

where we used properties (3) and (4) of Proposition 13.1. □

The next proposition will be needed in the proof of the second variation formula. If $\alpha\colon U \to M$ is a parametrized surface, where U is some open subset of $\mathbb{R}^2$, we say that a

vector field, $V \in \mathfrak{X}(M)$, is a *vector field along* α iff $V(x,y) \in T_{\alpha(x,y)}M$, for all $(x,y) \in U$. For any smooth vector field, V, along α, we also define the covariant derivatives, $DV/\partial x$ and $DV/\partial y$ as follows: For each fixed y_0, if we restrict V to the curve

$$x \mapsto \alpha(x, y_0)$$

we obtain a vector field, V_{y_0}, along this curve and we set

$$\frac{DX}{\partial x}(x, y_0) = \frac{DV_{y_0}}{dx}.$$

Then, we let y_0 vary so that $(x, y_0) \in U$ and this yields $DV/\partial x$. We define $DV/\partial y$ is a similar manner, using a fixed x_0.

Proposition 13.3 *For a Riemannian manifold, $(M, \langle -, - \rangle)$, equipped with the Levi-Civita connection, for every parametrized surface, $\alpha\colon \mathbb{R}^2 \to M$, for every vector field, $V \in \mathfrak{X}(M)$ along α, we have*

$$\frac{D}{\partial y}\frac{D}{\partial x}V - \frac{D}{\partial x}\frac{D}{\partial y}V = R\left(\frac{\partial \alpha}{\partial x}, \frac{\partial \alpha}{\partial y}\right)V.$$

Proof. Express both sides in local coordinates in a chart and make use of the identity

$$\nabla_{\frac{\partial}{\partial x_j}} \nabla_{\frac{\partial}{\partial x_i}} \frac{\partial}{\partial x_k} - \nabla_{\frac{\partial}{\partial x_i}} \nabla_{\frac{\partial}{\partial x_j}} \frac{\partial}{\partial x_k} = R\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right)\frac{\partial}{\partial x_k}.$$ □

Remark: Since the Levi-Civita connection is torsion-free, it is easy to check that

$$\frac{D}{\partial x}\frac{\partial \alpha}{\partial y} = \frac{D}{\partial y}\frac{\partial \alpha}{\partial x}.$$

We used this identity in the proof of Theorem 12.18.

The curvature tensor is a rather complicated object. Thus, it is quite natural to seek simpler notions of curvature. The sectional curvature is indeed a simpler object and it turns out that the curvature tensor can be recovered from it.

13.2 Sectional Curvature

Basically, the sectional curvature is the curvature of two-dimensional sections of our manifold. Given any two vectors, $u, v \in T_pM$, recall by Cauchy-Schwarz that

$$\langle u, v\rangle_p^2 \leq \langle u, u\rangle_p \langle v, v\rangle_p,$$

with equality iff u and v are linearly dependent. Consequently, if u and v are linearly independent, we have

$$\langle u, u\rangle_p \langle v, v\rangle_p - \langle u, v\rangle_p^2 \neq 0.$$

In this case, we claim that the ratio

$$K(u,v) = \frac{R_p(u,v,u,v)}{\langle u,u\rangle_p\langle v,v\rangle_p - \langle u,v\rangle_p^2}$$

is independent of the plane, Π, spanned by u and v. If (x,y) is another basis of Π, then

$$\begin{aligned} x &= au + bv \\ y &= cu + dv. \end{aligned}$$

We get

$$\langle x,x\rangle_p\langle y,y\rangle_p - \langle x,y\rangle_p^2 = (ad-bc)^2(\langle u,u\rangle_p\langle v,v\rangle_p - \langle u,v\rangle_p^2)$$

and similarly,

$$R_p(x,y,x,y) = \langle R_p(x,y)x,y\rangle_p = (ad-bc)^2 R_p(u,v,u,v),$$

which proves our assertion.

Definition 13.2 Let $(M,\langle -,-\rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. For every $p \in T_pM$, for every 2-plane, $\Pi \subseteq T_pM$, the *sectional curvature*, $K(\Pi)$, of Π, is given by

$$K(\Pi) = K(x,y) = \frac{R_p(x,y,x,y)}{\langle x,x\rangle_p\langle y,y\rangle_p - \langle x,y\rangle_p^2},$$

for any basis, (x,y), of Π.

Observe that if (x,y) is an orthonormal basis, then the denominator is equal to 1. The expression $R_p(x,y,x,y)$ is often denoted $\kappa_p(x,y)$. Remarkably, κ_p determines R_p. We denote the function $p \mapsto \kappa_p$ by κ. We state the following proposition without proof:

Proposition 13.4 *Let $(M,\langle -,-\rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. The function κ determines the curvature tensor, R. Thus, the knowledge of all the sectional curvatures determines the curvature tensor. Moreover, we have*

$$\begin{aligned} 6\langle R(x,y)z,w\rangle &= \kappa(x+w,y+z) - \kappa(x,y+z) - \kappa(w,y+z) \\ &\quad - \kappa(y+w,x+z) + \kappa(y,x+z) + \kappa(w,x+z) \\ &\quad - \kappa(x+w,y) + \kappa(x,y) + \kappa(w,y) \\ &\quad - \kappa(x+w,z) + \kappa(x,z) + \kappa(w,z) \\ &\quad + \kappa(y+w,x) - \kappa(y,x) - \kappa(w,x) \\ &\quad + \kappa(y+w,z) - \kappa(y,z) - \kappa(w,z). \end{aligned}$$

For a proof of this formidable equation, see Kuhnel [91] (Chapter 6, Theorem 6.5). A different proof of the above proposition (without an explicit formula) is also given in O'Neill [119] (Chapter III, Corollary 42).

Let

$$R_1(x,y)z = \langle x,z\rangle y - \langle y,z\rangle x.$$

Observe that

$$\langle R_1(x,y)x,y\rangle = \langle x,x\rangle\langle y,y\rangle - \langle x,y\rangle^2.$$

As a corollary of Proposition 13.4, we get:

Proposition 13.5 *Let $(M,\langle -,-\rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. If the sectional curvature, $K(\Pi)$ does not depend on the plane, Π, but only on $p \in M$, in the sense that K is a scalar function, $K\colon M \to \mathbb{R}$, then*

$$R = KR_1.$$

Proof. By hypothesis,

$$\kappa_p(x,y) = K(p)(\langle x,x\rangle_p\langle y,y\rangle_p - \langle x,y\rangle_p^2),$$

for all x, y. As the right-hand side of the formula in Proposition 13.4 consists of a sum of terms, we see that the right-hand side is equal to K times a similar sum with κ replaced by

$$\langle R_1(x,y)x,y\rangle = \langle x,x\rangle\langle y,y\rangle - \langle x,y\rangle^2,$$

so it is clear that $R = KR_1$. □

In particular, in dimension $n = 2$, the assumption of Proposition 13.5 holds and K is the well-known *Gaussian curvature* for surfaces.

Definition 13.3 A Riemannian manifold, $(M,\langle -,-\rangle)$ is said to have *constant (resp. negative, resp. positive) curvature* iff its sectional curvature is constant (resp. negative, resp. positive).

In dimension $n \geq 3$, we have the following somewhat surprising theorem due to F. Schur:

Proposition 13.6 *(F. Schur, 1886) Let $(M,\langle -,-\rangle)$ be a connected Riemannian manifold. If $\dim(M) \geq 3$ and if the sectional curvature, $K(\Pi)$, does not depend on the plane, $\Pi \subseteq T_pM$, but only on the point, $p \in M$, then K is constant (i.e., does not depend on p).*

The proof, which is quite beautiful, can be found in Kuhnel [91] (Chapter 6, Theorem 6.7).

If we replace the metric, $g = \langle -, - \rangle$ by the metric $\widetilde{g} = \lambda \langle -, - \rangle$ where $\lambda > 0$ is a constant, some simple calculations show that the Christoffel symbols and the Levi-Civita connection are unchanged, as well as the curvature tensor, but the sectional curvature is changed, with

$$\widetilde{K} = \lambda^{-1} K.$$

As a consequence, if M is a Riemannian manifold of constant curvature, by rescaling the metric, we may assume that either $K = -1$, or $K = 0$, or $K = +1$. Here are standard examples of spaces with constant curvature.

(1) The sphere, $S^n \subseteq \mathbb{R}^{n+1}$, with the metric induced by $\mathbb{R}^{n+1}$, where

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\}.$$

The sphere, S^n, has constant sectional curvature, $K = +1$. This can be shown by using the fact that the stabilizer of the action of $\mathbf{SO}(n+1)$ on S^n is isomorphic to $\mathbf{SO}(n)$. Then, it is easy to see that the action of $\mathbf{SO}(n)$ on $T_p S^n$ is transitive on 2-planes and from this, it follows that $K = 1$ (for details, see Gallot, Hulin and Lafontaine [60] (Chapter 3, Proposition 3.14).

(2) Euclidean space, $\mathbb{R}^{n+1}$, with its natural Euclidean metric. Of course, $K = 0$.

(3) The *hyperbolic space*, $\mathcal{H}_n^+(1)$, from Definition 2.10. Recall that this space is defined in terms of the *Lorentz innner product*, $\langle -, - \rangle_1$, on $\mathbb{R}^{n+1}$, given by

$$\langle (x_1, \ldots, x_{n+1}), (y_1, \ldots, y_{n+1}) \rangle_1 = -x_1 y_1 + \sum_{i=2}^{n+1} x_i y_i.$$

By definition, $\mathcal{H}_n^+(1)$, written simply H^n, is given by

$$H^n = \{x = (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid \langle x, x \rangle_1 = -1,\ x_1 > 0\}.$$

Given any points, $p = (x_1, \ldots, x_{n+1}) \in H^n$, it is easy to see that the set of tangent vectors, $u \in T_p H^n$, are given by the equation

$$\langle p, u \rangle_1 = 0,$$

that is, $T_p H^n$ is orthogonal to p with respect to the Lorentz inner-product. Since $p \in H^n$, we have $\langle p, p \rangle_1 = -1$, that is, u is lightlike, so by Proposition 2.10, all vectors in $T_p H^n$ are spacelike, that is,

$$\langle u, u \rangle_1 > 0, \qquad \text{for all } u \in T_p H^n,\ u \neq 0.$$

Therefore, the restriction of $\langle -, - \rangle_1$ to H^n is positive, definite, which means that it is a metric on $T_p H^n$. The space H^n equipped with this metric, g_H, is called *hyperbolic space* and it has constant curvature, $K = -1$. This can be shown by using the fact that the stabilizer of the action of $\mathbf{SO}_0(n, 1)$ on H^n is isomorphic to $\mathbf{SO}(n)$ (see Proposition 2.11). Then, it is easy to see that the action of $\mathbf{SO}(n)$ on $T_p H^n$ is transitive on 2-planes and from this, it follows that $K = -1$ (for details, see Gallot, Hulin and Lafontaine [60] (Chapter 3, Proposition 3.14).

There are other isometric models of H^n that are perhaps intuitively easier to grasp but for which the metric is more complicated. For example, there is a map, $\mathrm{PD}\colon B^n \to H^n$, where $B^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$ is the open unit ball in $\mathbb{R}^n$, given by

$$\mathrm{PD}(x) = \left(\frac{1 + \|x\|^2}{1 - \|x\|^2}, \frac{2x}{1 - \|x\|^2}\right).$$

It is easy to check that $\langle \mathrm{PD}(x), \mathrm{PD}(x)\rangle_1 = -1$ and that PD is bijective and an isometry. One also checks that the pull-back metric, $g_{\mathrm{PD}} = \mathrm{PD}^* g_H$, on B^n, is given by

$$g_{\mathrm{PD}} = \frac{4}{(1 - \|x\|^2)^2}\,(dx_1^2 + \cdots + dx_n^2).$$

The metric, g_{PD} is called the *conformal disc metric* and the Riemannian manifold, (B^n, g_{PD}) is called the *Poincaré disc model* or *conformal disc model*. The metric g_{PD} is proportional to the Euclidean metric and thus, angles are preserved under the map PD. Another model is the *Poincaré half-plane model*, $\{x \in \mathbb{R}^n \mid x_1 > 0\}$, with the metric

$$g_{\mathrm{PH}} = \frac{1}{x_1^2}\,(dx_1^2 + \cdots + dx_n^2).$$

We already encountered this space for $n = 2$.

The metrics for S^n, $\mathbb{R}^{n+1}$ and H^n have a nice expression in polar coordinates but we prefer to discuss the Ricci curvature next.

13.3 Ricci Curvature

The Ricci tensor is another important notion of curvature. It is mathematically simpler than the sectional curvature (since it is symmetric) but it plays an important role in the theory of gravitation as it occurs in the Einstein field equations. The Ricci tensor is an example of contraction, in this case, the trace of a linear map. Recall that if $f\colon E \to E$ is a linear map from a finite-dimensional Euclidean vector space to itself, given any orthonormal basis, $(e_1, \ldots, e_n)$, we have

$$\mathrm{tr}(f) = \sum_{i=1}^{n} \langle f(e_i), e_i\rangle.$$

Definition 13.4 Let $(M, \langle -, -\rangle)$ be a Riemannian manifold (equipped with the Levi-Civita connection). The *Ricci curvature*, Ric, of M is the $(0,2)$-tensor defined as follows: For every $p \in M$, for all $x, y \in T_pM$, set $\mathrm{Ric}_p(x, y)$ to be the trace of the endomorphism, $v \mapsto R_p(x, v)y$. With respect to any orthonormal basis, $(e_1, \ldots, e_n)$, of T_pM, we have

$$\mathrm{Ric}_p(x, y) = \sum_{j=1}^{n} \langle R_p(x, e_j)y, e_j\rangle_p = \sum_{j=1}^{n} R_p(x, e_j, y, e_j).$$

The *scalar curvature*, S, of M, is the trace of the Ricci curvature, that is, for every $p \in M$,

$$S(p) = \sum_{i \neq j} R(e_i, e_j, e_i, e_j) = \sum_{i \neq j} K(e_i, e_j),$$

where $K(e_i, e_j)$ denotes the sectional curvature of the plane spanned by e_i, e_j.

In view of Proposition 13.1 (4), the Ricci curvature is symmetric. The tensor Ric is a $(0,2)$-tensor but it can be interpreted as a $(1,1)$-tensor as follows: We let $\mathrm{Ric}_p^{\#}$ be the $(1,1)$-tensor given by

$$\langle \mathrm{Ric}_p^{\#} u, v\rangle_p = \mathrm{Ric}(u, v),$$

for all $u, v \in T_pM$. Then, it is easy to see that

$$S(p) = \mathrm{tr}(\mathrm{Ric}_p^{\#}).$$

This is why we said (by abuse of language) that S is the trace of Ric. Observe that if $(e_1, \ldots, e_n)$ is any orthonormal basis of T_pM, as

$$\begin{aligned} \mathrm{Ric}_p(u, v) &= \sum_{j=1}^{n} R_p(u, e_j, v, e_j) \\ &= \sum_{j=1}^{n} R_p(e_j, u, e_j, v) \\ &= \sum_{j=1}^{n} \langle R_p(e_j, u)e_j, v\rangle_p, \end{aligned}$$

we have

$$\mathrm{Ric}_p^{\#}(u) = \sum_{j=1}^{n} R_p(e_j, u)e_j.$$

Observe that in dimension $n = 2$, we get $S(p) = 2K(p)$. Therefore, in dimension 2, the scalar curvature determines the curvature tensor. In dimension $n = 3$, it turns out that the Ricci tensor completely determines the curvature tensor, although this is not obvious. We will come back to this point later.

Since $\mathrm{Ric}(x, y)$ is symmetric, $\mathrm{Ric}(x, x)$ determines $\mathrm{Ric}(x, y)$ completely (Use the polarization identity for a symmetric bilinear form, φ:

$$2\varphi(x, y) = \varphi(x + y) - \varphi(x) - \varphi(y).)$$

Observe that for any orthonormal frame, $(e_1, \ldots, e_n)$, of T_pM, using the definition of the sectional curvature, K, we have

$$\mathrm{Ric}(e_1, e_1) = \sum_{i=1}^{n} \langle (R(e_1, e_i)e_1, e_i\rangle = \sum_{i=2}^{n} K(e_1, e_i).$$

Thus, $\mathrm{Ric}(e_1, e_1)$ is the sum of the sectional curvatures of any $n-1$ orthogonal planes orthogonal to e_1 (a unit vector).

For a Riemannian manifold with constant sectional curvature, we see that

$$\mathrm{Ric}(x, x) = (n-1)Kg(x, x), \qquad S = n(n-1)K,$$

where $g = \langle -, - \rangle$ is the metric on M. Indeed, if K is constant, then we know that $R = KR_1$ and so,

$$\begin{aligned}
\mathrm{Ric}(x, x) &= K \sum_{i=1}^{n} g(R_1(x, e_i)x, e_i) \\
&= K \sum_{i=1}^{n} (g(e_i, e_i)g(x, x) - g(e_i, x)^2) \\
&= K(ng(x, x) - \sum_{i=1}^{n} g(e_i, x)^2) \\
&= (n-1)Kg(x, x).
\end{aligned}$$

Spaces for which the Ricci tensor is proportional to the metric are called Einstein spaces.

Definition 13.5 A Riemannian manifold, (M, g), is called an *Einstein space* iff the Ricci curvature is proportional to the metric, g, that is:

$$\mathrm{Ric}(x, y) = \lambda g(x, y),$$

for some function, $\lambda \colon M \to \mathbb{R}$.

If M is an Einstein space, observe that $S = n\lambda$.

Remark: For any Riemanian manifold, (M, g), the quantity

$$G = \mathrm{Ric} - \frac{S}{2} g$$

is called the *Einstein tensor* (or *Einstein gravitation tensor* for space-times spaces). The Einstein tensor plays an important role in the theory of general relativity. For more on this topic, see Kuhnel [91] (Chapters 6 and 8) O'Neill [119] (Chapter 12).

13.4 Isometries and Local Isometries

Recall that a *local isometry* between two Riemannian manifolds, M and N, is a smooth map, $\varphi\colon M \to N$, so that

$$\langle (d\varphi)_p(u), (d\varphi_p)(v)\rangle_{\varphi(p)} = \langle u, v\rangle_p,$$

for all $p \in M$ and all $u, v \in T_pM$. An *isometry* is a local isometry and a diffeomorphism.

By the inverse function theorem, if $\varphi\colon M \to N$ is a local isometry, then for every $p \in M$, there is some open subset, $U \subseteq M$, with $p \in U$, so that $\varphi \restriction U$ is an isometry between U and $\varphi(U)$.

Also recall that if $\varphi\colon M \to N$ is a diffeomorphism, then for any vector field, X, on M, the vector field, φ_*X, on N (called the *push-forward* of X) is given by

$$(\varphi_*X)_q = d\varphi_{\varphi^{-1}(q)}X(\varphi^{-1}(q)), \qquad \text{for all } q \in N,$$

or equivalently, by

$$(\varphi_*X)_{\varphi(p)} = d\varphi_p X(p), \qquad \text{for all } p \in M.$$

For any smooth function, $h\colon N \to \mathbb{R}$, for any $q \in N$, we have

$$\begin{aligned} X_*(h)_q &= dh_q(X_*(q)) \\ &= dh_q(d\varphi_{\varphi^{-1}(q)}X(\varphi^{-1}(q))) \\ &= d(h\circ\varphi)_{\varphi^{-1}(q)}X(\varphi^{-1}(q)) \\ &= X(h\circ\varphi)_{\varphi^{-1}(q)}, \end{aligned}$$

that is

$$X_*(h)_q = X(h\circ\varphi)_{\varphi^{-1}(q)},$$

or

$$X_*(h)_{\varphi(p)} = X(h\circ\varphi)_p.$$

It is natural to expect that isometries preserve all "natural" Riemannian concepts and this is indeed the case. We begin with the Levi-Civita connection.

Proposition 13.7 *If $\varphi\colon M \to N$ is an isometry, then*

$$\varphi_*(\nabla_X Y) = \nabla_{\varphi_*X}(\varphi_*Y), \qquad \textit{for all } X, Y \in \mathfrak{X}(M),$$

where $\nabla_X Y$ is the Levi-Civita connection induced by the metric on M and similarly on N.

Proof. We use the Koszul formula (Proposition 11.18),

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &= X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) \\ &\quad - \langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle. \end{aligned}$$

We have

$$(\varphi_*(\nabla_X Y))_{\varphi(p)} = d\varphi_p(\nabla_X Y)_p,$$

and as φ is an isometry,

$$\langle d\varphi_p(\nabla_X Y)_p, d\varphi_p Z_p\rangle_{\varphi(p)} = \langle (\nabla_X Y)_p, Z_p\rangle_p,$$

so, Koszul yields

$$\begin{aligned} 2\langle \varphi_*(\nabla_X Y), \varphi_* Z\rangle_{\varphi(p)} &= X(\langle Y, Z\rangle_p) + Y(\langle X, Z\rangle_p) - Z(\langle X, Y\rangle_p) \\ &\quad - \langle Y, [X, Z]\rangle_p - \langle X, [Y, Z]\rangle_p - \langle Z, [Y, X]\rangle_p. \end{aligned}$$

Next, we need to compute

$$\langle \nabla_{\varphi_* X}(\varphi_* Y), \varphi_* Z\rangle_{\varphi(p)}.$$

When we plug $\varphi_* X$, $\varphi_* Y$ and $\varphi_* Z$ into the Koszul formula, as φ is an isometry, for the fourth term on the right-hand side, we get

$$\begin{aligned} \langle \varphi_* Y, [\varphi_* X, \varphi_* Z]\rangle_{\varphi(p)} &= \langle d\varphi_p Y_p, [d\varphi_p X_p, d\varphi_p Z_p]\rangle_{\varphi(p)} \\ &= \langle Y_p, [X_p, Z_p]\rangle_p \end{aligned}$$

and similarly for the fifth and sixth term on the right-hand side. For the first term on the right-hand side, we get

$$\begin{aligned} (\varphi_* X)(\langle \varphi_* Y, \varphi_* Z\rangle)_{\varphi(p)} &= (\varphi_* X)(\langle d\varphi_p Y_p, d\varphi_p Z_p\rangle)_{\varphi(p)} \\ &= (\varphi_* X)(\langle Y_p, Z_p\rangle_{\varphi^{-1}(\varphi(p))})_{\varphi(p)} \\ &= (\varphi_* X)(\langle Y, Z\rangle \circ \varphi^{-1})_{\varphi(p)} \\ &= X(\langle Y, Z\rangle \circ \varphi^{-1} \circ \varphi)_p \\ &= X(\langle Y, Z\rangle)_p \end{aligned}$$

and similarly for the second and third term. Consequently, we get

$$\begin{aligned} 2\langle \nabla_{\varphi_* X}(\varphi_* Y), \varphi_* Z\rangle_{\varphi(p)} &= X(\langle Y, Z\rangle_p) + Y(\langle X, Z\rangle_p) - Z(\langle X, Y\rangle_p) \\ &\quad - \langle Y, [X, Z]\rangle_p - \langle X, [Y, Z]\rangle_p - \langle Z, [Y, X]\rangle_p. \end{aligned}$$

By comparing right-hand sides, we get

$$2\langle \varphi_*(\nabla_X Y), \varphi_* Z\rangle_{\varphi(p)} = 2\langle \nabla_{\varphi_* X}(\varphi_* Y), \varphi_* Z\rangle_{\varphi(p)}$$

for all X, Y, Z, and as φ is a diffeomorphism, this implies

$$\varphi_*(\nabla_X Y) = \nabla_{\varphi_* X}(\varphi_* Y),$$

as claimed. □

As a corollary of Proposition 13.7, the curvature induced by the connection is preserved, that is

$$\varphi_* R(X,Y)Z = R(\varphi_* X, \varphi_* Y)\varphi_* Z,$$

as well as the parallel transport, the covariant derivative of a vector field along a curve, the exponential map, sectional curvature, Ricci curvature and geodesics. Actually, all concepts that are local in nature are preserved by local diffeomorphisms! So, except for the Levi-Civita connection and if we consider the Riemann tensor on vectors, all the above concepts are preserved under local diffeomorphisms. For the record, we state:

Proposition 13.8 *If $\varphi\colon M \to N$ is a local isometry, then the following concepts are preserved:*

(1) The covariant derivative of vector fields along a curve, γ, that is

$$d\varphi_{\gamma(t)} \frac{DX}{dt} = \frac{D\varphi_* X}{dt},$$

for any vector field, X, along γ, with $(\varphi_* X)(t) = d\varphi_{\gamma(t)} Y(t)$, for all t.

(2) Parallel translation along a curve. If P_γ denotes parallel transport along the curve γ and if $P_{\varphi\circ\gamma}$ denotes parallel transport along the curve $\varphi \circ \gamma$, then

$$d\varphi_{\gamma(1)} \circ P_\gamma = P_{\varphi\circ\gamma} \circ d\varphi_{\gamma(0)}.$$

(3) Geodesics. If γ is a geodesic in M, then $\varphi \circ \gamma$ is a geodesic in N. Thus, if γ_v is the unique geodesic with $\gamma(0) = p$ and $\gamma_v'(0) = v$, then

$$\varphi \circ \gamma_v = \gamma_{d\varphi_p v},$$

wherever both sides are defined. Note that the domain of $\gamma_{d\varphi_p v}$ may be strictly larger than the domain of γ_v. For example, consider the inclusion of an open disc into $\mathbb{R}^2$.

(4) Exponential maps. We have

$$\varphi \circ \exp_p = \exp_{\varphi(p)} \circ d\varphi_p,$$

wherever both sides are defined.

(5) Riemannian curvature tensor. We have

$$d\varphi_p R(x,y)z = R(d\varphi_p x, d\varphi_p y) d\varphi_p z, \qquad \text{for all } x, y, z \in T_p M.$$

(6) Sectional, Ricci and Scalar curvature. We have

$$K(d\varphi_p x, d\varphi_p y) = K(x,y)_p,$$

for all linearly independent vectors, $x, y \in T_pM$;

$$\mathrm{Ric}(d\varphi_p x, d\varphi_p y) = \mathrm{Ric}(x, y)_p$$

for all $x, y \in T_pM$;

$$S_M = S_N \circ \varphi.$$

where S_M is the scalar curvature on M and S_N is the scalar curvature on N.

A useful property of local diffeomorphisms is stated below. For a proof, see O'Neill [119] (Chapter 3, Proposition 62):

Proposition 13.9 *Let $\varphi, \psi \colon M \to N$ be two local isometries. If M is connected and if $\varphi(p) = \psi(p)$ and $d\varphi_p = d\psi_p$ for some $p \in M$, then $\varphi = \psi$.*

The idea is to prove that

$$\{p \in M \mid d\varphi_p = d\psi_p\}$$

is both open and closed and for this, to use the preservation of the exponential under local diffeomorphisms.

13.5 Riemannian Covering Maps

The notion of covering map discussed in Section 3.9 can be extended to Riemannian manifolds.

Definition 13.6 If M and N are two Riemannian manifold, then a map, $\pi \colon M \to N$, is a *Riemannian covering* iff the following conditions hold:

(1) The map π is a smooth covering map.

(2) The map π is a local isometry.

Recall from Section 3.9 that a covering map is a local diffeomorphism. A way to obtain a metric on a manifold, M, is to pull-back the metric, g, on a manifold, N, along a local diffeomorphism, $\varphi \colon M \to N$ (see Section 7.4). If φ is a covering map, then it becomes a Riemannian covering map.

Proposition 13.10 *Let $\pi \colon M \to N$ be a smooth covering map. For any Riemannian metric, g, on N, there is a unique metric, $\pi^* g$, on M, so that π is a Riemannian covering.*

Proof. We define the *pull-back metric*, π^*g, on M induced by g as follows: For all $p \in M$, for all $u, v \in T_pM$,

$$(\pi^*g)_p(u,v) = g(d\pi_p(u), d\pi_p(v)).$$

We need to check that $(\pi^*g)_p$ is an inner product, which is very easy since $d\pi_p$ is a linear isomorphism. Our map, π, between the two Riemannian manifolds (M, π^*g) and (N, g) becomes a local isometry. Now, every metric on M making π a local isometry has to satisfy the equation defining, π^*g, so this metric is unique. □

As a corollary of Proposition 13.10 and Theorem 3.35, every connected Riemmanian manifold, M, has a simply connected covering map, $\pi\colon \widetilde{M} \to M$, where π is a Riemannian covering. Furthermore, if $\pi\colon M \to N$ is a Riemannian covering and $\varphi\colon P \to N$ is a local isometry, it is easy to see that its lift, $\widetilde{\varphi}\colon P \to M$, is also a local isometry. In particular, the deck-transformations of a Riemannian covering are isometries.

In general, a local isometry is not a Riemannian covering. However, this is the case when the source space is complete.

Proposition 13.11 *Let $\pi\colon M \to N$ be a local isometry with N connected. If M is a complete manifold, then π is a Riemannian covering map.*

Proof. We follow the proof in Sakai [130] (Chapter III, Theorem 5.4). Because π is a local isometry, geodesics in M can be projected onto geodesics in N and geodesics in N can be lifted back to M. The proof makes heavy use of these facts.

First, we prove that N is complete. Pick any $p \in M$ and let $q = \pi(p)$. For any geodesic, γ_u, of N with initial point, $q \in N$, and initial direction the unit vector, $u \in T_qN$, consider the geodesic, $\widetilde{\gamma}_u$, of M, with initial point p and with $u = d\pi_q^{-1}(v) \in T_pM$. As π is a local isometry, it preserves geodesic, so

$$\gamma_v = \pi \circ \widetilde{\gamma}_u,$$

and since $\widetilde{\gamma}_u$ is defined on $\mathbb{R}$ because M is complete, so if γ_v. As $\exp_q$ is defined on the whole of T_qN, by Hopf-Rinow, N is complete.

Next, we prove that π is surjective. As N is complete, for any $q_1 \in N$, there is a minimal geodesic, $\gamma\colon [0, b] \to N$, joining q to q_1 and for the geodesic, $\widetilde{\gamma}$, in M, emanating from p and with initial direction $d\pi_q^{-1}(\gamma'(0))$, we have $\pi(\widetilde{\gamma}(b)) = \gamma(b) = q_1$, establishing surjectivity.

For any $q \in N$, pick $r > 0$ wih $r < i(q)$, where $i(q)$ denotes the injectivity radius of N at q and consider the open metric ball, $B_r(q) = \exp_q(B(0_q, r))$ (where $B(0_q, r)$ is the open ball of radius r in T_qN). Let

$$\pi^{-1}(q) = \{p_i\}_{i \in I} \subseteq M.$$

We claim that the following properties hold:

(1) Each map, $\pi \upharpoonright B_r(p_i)\colon B_r(p_i) \longrightarrow B_r(q)$, is a diffeomorphism, in fact, an isometry.

(2) $\pi^{-1}(B_r(q)) = \bigcup_{i \in I} B_r(p_i)$.

(3) $B_r(p_i) \cap B_r(p_j) = \emptyset$ whenever $i \neq j$.

It follows from (1), (2) and (3) that $B_r(q)$ is evenly covered by the family of open sets, $\{B_r(p_i)\}_{i\in I}$, so π is a covering map.

(1) Since π is a local isometry, it maps geodesics emanating from p_i to geodesics emanating from q so the following diagram commutes:

$$\begin{array}{ccc} B(0_{p_i}, r) & \xrightarrow{d\pi_{p_i}} & B(0_q, r) \\ {\scriptstyle \exp_{p_i}}\downarrow & & \downarrow{\scriptstyle \exp_q} \\ B_r(p_i) & \xrightarrow[\pi]{} & B_r(q). \end{array}$$

Since $\exp_q \circ d\pi_{p_i}$ is a diffeomorphism, $\pi \restriction B_r(p_i)$ must be injective and since $\exp_{p_i}$ is surjective, so is $\pi \restriction B_r(p_i)$. Then, $\pi \restriction B_r(p_i)$ is a bijection and as π is a local diffeomorphism, $\pi \restriction B_r(p_i)$ is a diffeomorphism.

(2) Obviously, $\bigcup_{i\in I} B_r(p_i) \subseteq \pi^{-1}(B_r(q))$, by (1). Conversely, pick $p_1 \in \pi^{-1}(B_r(q))$. For $q_1 = \pi(p_1)$, we can write $q_1 = \exp_q v$, for some $v \in B(0_q, r)$ and the map $\gamma(t) = \exp_q(1-t)v$, for $t \in [0,1]$, is a geodesic in N joining q_1 to q. Then, we have the geodesic, $\widetilde{\gamma}$, emanating from p_1 with initial direction $d\pi_{q_1}^{-1}(\gamma'(0))$ and as $\pi \circ \widetilde{\gamma}(1) = \gamma(1) = q$, we have $\widetilde{\gamma}(1) = p_i$ for some α. Since γ has length less than r, we get $p_1 \in B_r(p_i)$.

(3) Suppose $p_1 \in B_r(p_i) \cap B_r(p_j)$. We can pick a minimal geodesic, $\widetilde{\gamma}$, in $B_r(p_i)$, (resp $\widetilde{\omega}$ in $B_r(p_j)$) joining p_i to p (resp. joining p_j to p). Then, the geodesics $\pi \circ \widetilde{\gamma}$ and $\pi \circ \widetilde{\omega}$ are geodesics in $B_r(q)$ from q to $\pi(p_1)$ and their length is less than r. Since $r < i(q)$, these geodesics are minimal so they must coincide. Therefore, $\gamma = \omega$, which implies $i = j$. □

13.6 The Second Variation Formula and the Index Form

In Section 12.4, we discovered that the geodesics are exactly the critical paths of the energy functional (Theorem 12.19). For this, we derived the First Variation Formula (Theorem 12.18). It is not too surprising that a deeper understanding is achieved by investigating the second derivative of the energy functional at a critical path (a geodesic). By analogy with the Hessian of a real-valued function on $\mathbb{R}^n$, it is possible to define a bilinear functional,

$$I_\gamma \colon T_\gamma\Omega(p,q) \times T_\gamma\Omega(p,q) \to \mathbb{R},$$

when γ is a critical point of the energy function, E (that is, γ is a geodesic). This bilinear form is usually called the *index form*. Note that Milnor denotes I_γ by E_{**} and refers to it as the *Hessian* of E but this is a bit confusing since I_γ is only defined for critical points, whereas the Hessian is defined for all points, critical or not.

Now, if $f\colon M \to \mathbb{R}$ is a real-valued function on a finite-dimensional manifold, M, and if p is a critical point of f, which means that $df_p = 0$, it is not hard to prove that there is a symmetric bilinear map, $I\colon T_pM \times T_pM \to \mathbb{R}$, such that

$$I(X(p), Y(p)) = X_p(Yf) = Y_p(Xf),$$

for all vector fields, $X, Y \in \mathfrak{X}(M)$. Furthermore, $I(u, v)$ can be computed as follows, for any $u, v \in T_pM$: for any smooth map, $\alpha\colon \mathbb{R}^2 \to \mathbb{R}$, such that

$$\alpha(0,0) = p, \quad \frac{\partial \alpha}{\partial x}(0,0) = u, \quad \frac{\partial \alpha}{\partial y}(0,0) = v,$$

we have

$$I(u, v) = \left.\frac{\partial^2 (f \circ \alpha)(x, y)}{\partial x \partial y}\right|_{(0,0)}.$$

The above suggests that in order to define

$$I_\gamma\colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R},$$

that is, to define $I_\gamma(W_1, W_2)$, where $W_1, W_2 \in T_\gamma\Omega(p, q)$ are vector fields along γ (with $W_1(0) = W_2(0) = 0$ and $W_1(1) = W_2(1) = 0$), we consider 2-parameter variations,

$$\alpha\colon U \times [0, 1] \to M,$$

where U is an open subset of $\mathbb{R}^2$ with $(0, 0) \in U$, such that

$$\alpha(0, 0, t) = \gamma(t), \quad \frac{\partial \alpha}{\partial u_1}(0, 0, t) = W_1(t), \quad \frac{\partial \alpha}{\partial u_2}(0, 0, t) = W_2(t).$$

Then, we set

$$I_\gamma(W_1, W_2) = \left.\frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2}\right|_{(0,0)},$$

where $\widetilde{\alpha} \in \Omega(p, q)$ is the path given by

$$\widetilde{\alpha}(u_1, u_2)(t) = \alpha(u_1, u_2, t).$$

For simplicity of notation, the above derivative if often written as $\frac{\partial^2 E}{\partial u_1 \partial u_2}(0, 0)$.

To prove that $I_\gamma(W_1, W_2)$ is actually well-defined, we need the following result:

Theorem 13.12 *(Second Variation Formula) Let $\alpha\colon U \times [0, 1] \to M$ be a 2-parameter variation of a geodesic, $\gamma \in \Omega(p, q)$, with variation vector fields $W_1, W_2 \in T_\gamma\Omega(p, q)$ given by*

$$W_1(t) = \frac{\partial \alpha}{\partial u_1}(0, 0, t), \quad W_2(t) = \frac{\partial \alpha}{\partial u_2}(0, 0, t).$$

Then, we have the formula

$$\frac{1}{2}\left.\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2}\right|_{(0,0)} = -\sum_t \left\langle W_2(t), \Delta_t\frac{dW_1}{dt}\right\rangle - \int_0^1 \left\langle W_2, \frac{D^2W_1}{dt^2} + R(V,W_1)V\right\rangle dt,$$

where $V(t)=\gamma'(t)$ is the velocity field,

$$\Delta_t\frac{dW_1}{dt} = \frac{dW_1}{dt}(t_+) - \frac{dW_1}{dt}(t_-)$$

is the jump in $\frac{dW_1}{dt}$ at one of its finitely many points of discontinuity in $(0,1)$ and E is the energy function on $\Omega(p,q)$.

Proof. (After Milnor, see [106], Chapter II, Section 13, Theorem 13.1.) By the First Variation Formula (Theorem 12.18), we have

$$\frac{1}{2}\frac{\partial E(\widetilde{\alpha}(u_1,u_2))}{\partial u_2} = -\sum_i \left\langle \frac{\partial\alpha}{\partial u_2}, \Delta_t\frac{\partial\alpha}{\partial t}\right\rangle - \int_0^1 \left\langle \frac{\partial\alpha}{\partial u_2}, \frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt.$$

Thus, we get

$$\begin{aligned}\frac{1}{2}\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2} &= -\sum_i \left\langle \frac{D}{\partial u_1}\frac{\partial\alpha}{\partial u_2}, \Delta_t\frac{\partial\alpha}{\partial t}\right\rangle - \sum_i \left\langle \frac{\partial\alpha}{\partial u_2}, \frac{D}{\partial u_1}\Delta_t\frac{\partial\alpha}{\partial t}\right\rangle \\ &\quad - \int_0^1 \left\langle \frac{D}{\partial u_1}\frac{\partial\alpha}{\partial u_2}, \frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt - \int_0^1 \left\langle \frac{\partial\alpha}{\partial u_2}, \frac{D}{\partial u_1}\frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt.\end{aligned}$$

Let us evaluate this expression for $(u_1,u_2)=(0,0)$. Since $\gamma=\widetilde{\alpha}(0,0)$ is an unbroken geodesic, we have

$$\Delta_t\frac{\partial\alpha}{\partial t}=0, \qquad \frac{D}{\partial t}\frac{\partial\alpha}{\partial t}=0,$$

so that the first and third term are zero. As

$$\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial t} = \frac{D}{\partial t}\frac{\partial\alpha}{\partial u_1},$$

(see the remark just after Proposition 13.3), we can rewrite the second term and we get

$$\frac{1}{2}\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2}(0,0) = -\sum_i \left\langle W_2, \Delta_t\frac{D}{\partial t}W_1\right\rangle - \int_0^1 \left\langle W_2, \frac{D}{\partial u_1}\frac{D}{\partial t}V\right\rangle dt. \qquad (*)$$

In order to interchange the operators $\frac{D}{\partial u_1}$ and $\frac{D}{\partial t}$, we need to bring in the curvature tensor. Indeed, by Proposition 13.3, we have

$$\frac{D}{\partial u_1}\frac{D}{\partial t}V - \frac{D}{\partial t}\frac{D}{\partial u_1}V = R\left(\frac{\partial\alpha}{\partial t}, \frac{\partial\alpha}{\partial u_1}\right)V = R(V,W_1)V.$$

Together with the equation

$$\frac{D}{\partial u_1} V = \frac{D}{\partial u_1}\frac{\partial \alpha}{\partial t} = \frac{D}{\partial t}\frac{\partial \alpha}{\partial u_1} = \frac{D}{\partial t} W_1,$$

this yields

$$\frac{D}{\partial u_1}\frac{D}{\partial t} V = \frac{D^2 W_1}{dt^2} + R(V, W_1)V.$$

Substituting this last expression in $(*)$, we get the Second Variation Formula. □

Theorem 13.12 shows that the expression

$$\left.\frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2}\right|_{(0,0)}$$

only depends on the variation fields W_1 and W_2 and thus, $I_\gamma(W_1, W_2)$ is actually well-defined. If no confusion arises, we write $I(W_1, W_2)$ for $I_\gamma(W_1, W_2)$.

Proposition 13.13 *Given any geodesic, $\gamma \in \Omega(p, q)$, the map, $I \colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$, defined so that, for all $W_1, W_2 \in T_\gamma\Omega(p, q)$,*

$$I(W_1, W_2) = \left.\frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2}\right|_{(0,0)},$$

only depends on W_1 and W_2 and is bilinear and symmetric, where $\alpha \colon U \times [0, 1] \to M$ is any 2-parameter variation, with

$$\alpha(0, 0, t) = \gamma(t), \quad \frac{\partial \alpha}{\partial u_1}(0, 0, t) = W_1(t), \quad \frac{\partial \alpha}{\partial u_2}(0, 0, t) = W_2(t).$$

Proof. We already observed that the Second Variation Formula implies that $I(W_1, W_2)$ is well defined. This formula also shows that I is bilinear. As

$$\frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2} = \frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_2 \partial u_1},$$

I is symmetric (but this is not obvious from the right-handed side of the Second Variation Formula). □

On the diagonal, $I(W, W)$ can be described in terms of a 1-parameter variation of γ. In fact,

$$I(W, W) = \frac{d^2 E(\widetilde{\alpha})}{du^2}(0),$$

where $\widetilde{\alpha} \colon (-\epsilon, \epsilon) \to \Omega(p, q)$ denotes any variation of γ with variation vector field, $\frac{d\widetilde{\alpha}}{du}(0)$ equal to W. To prove this equation it is only necessary to introduce the 2-parameter variation

$$\widetilde{\beta}(u_1, u_2) = \widetilde{\alpha}(u_1 + u_2)$$

and to observe that

$$\frac{\partial\widetilde{\beta}}{\partial u_i} = \frac{d\widetilde{\alpha}}{du}, \qquad \frac{\partial^2(E\circ\widetilde{\beta})}{\partial u_1\partial u_2} = \frac{d^2(E\circ\widetilde{\alpha})}{du^2}.$$

As an application of the above remark we have the following result:

Proposition 13.14 *If $\gamma \in \Omega(p,q)$ is a minimal geodesic, then the bilinear index form, I, is positive semi-definite, which means that $I(W,W) \geq 0$, for all $W \in T_\gamma\Omega(p,q)$.*

Proof. The inequality

$$E(\widetilde{\alpha}(u)) \geq E(\gamma) = E(\widetilde{\alpha}(0))$$

implies that

$$\frac{d^2E(\widetilde{\alpha})}{du^2}\,(0) \geq 0,$$

which is exactly what needs to be proved. □

If we define the *index* of

$$I\colon T_\gamma\Omega(p,q)\times T_\gamma\Omega(p,q)\to\mathbb{R}$$

as the maximum dimension of a subspace of $T_\gamma\Omega(p,q)$ on which I is negative definite, then Proposition 13.14 says that the index of I is zero (for the minimal geodesic γ). It turns out that the index of I is finite for any geodesic, γ (this is a consequence of the *Morse Index Theorem*).

13.7 Jacobi Fields and Conjugate Points

Jacobi fields arise naturally when considering the expression involved under the integral sign in the Second Variation Formula and also when considering the derivative of the exponential.

If $B\colon E\times E\to\mathbb{R}$ is a symmetric bilinear form defined on some vector space, E (possibly infinite dimentional), recall that the *nullspace* of B is the subset, $\mathrm{null}(B)$, of E given by

$$\mathrm{null}(B) = \{u\in E \mid B(u,v) = 0, \quad \text{for all } v\in E\}.$$

The *nullity*, ν, of B is the dimension of its nullspace. The bilinear form, B, is *nondegenerate* iff $\mathrm{null}(B) = (0)$ iff $\nu = 0$. If U is a subset of E, we say that B is *positive definite* (resp. *negative definite*) *on* U iff $B(u,u) > 0$ (resp. $B(u,u) < 0$) for all $u\in U$, with $u\neq 0$. The *index* of B is the maximum dimension of a subspace of E on which B is negative definite. We will determine the nullspace of the symmetric bilinear form,

$$I\colon T_\gamma\Omega(p,q)\times T_\gamma\Omega(p,q)\to\mathbb{R},$$

where γ is a geodesic from p to q in some Riemannian manifold, M. Now, if W is a vector field in $T_\gamma\Omega(p,q)$ and W satisfies the equation

$$\frac{D^2W}{dt^2} + R(V,W)V = 0, \qquad (*)$$

where $V(t) = \gamma'(t)$ is the velocity field of the geodesic, γ, since W is smooth along γ, it is obvious from the Second Variation Formula that

$$I(W, W_2) = 0, \quad \text{for all } W_2 \in T_\gamma\Omega(p,q).$$

Therefore, any vector field in the nullspace of I must satisfy equation $(*)$. Such vector fields are called *Jacobi fields*.

Definition 13.7 Given a geodesic, $\gamma \in \Omega(p,q)$, a vector field, J, along γ is a *Jacobi field* iff it satisfies the *Jacobi differential equation*

$$\frac{D^2J}{dt^2} + R(\gamma', J)\gamma' = 0.$$

The equation of Definition 13.7 is a linear second-order differential equation that can be transformed into a more familiar form by picking some orthonormal parallel vector fields, $X_1, \ldots, X_n$, along γ. To do this, pick any orthonormal basis, $(e_1, \ldots, e_n)$ in T_pM, with $e_1 = \gamma'(0)/\left\|\gamma'(0)\right\|$, and use parallel transport along γ to get $X_1, \ldots, X_n$. Then, we can write $J = \sum_{i=1}^n y_i X_i$, for some smooth functions, y_i, and the Jacobi equation becomes the system of second-order linear ODE's,

$$\frac{d^2y_i}{dt^2} + \sum_{j=1}^n R(\gamma', E_j, \gamma', E_i)y_j = 0, \qquad 1 \leq i \leq n.$$

By the existence and uniqueness theorem for ODE's, for every pair of vectors, $u, v \in T_pM$, there is a unique Jacobi fields, J, so that $J(0) = u$ and $\frac{DJ}{dt}(0) = v$. Since T_pM has dimension n, it follows that the dimension of the space of Jacobi fields along γ is $2n$. If $J(0)$ and $\frac{DJ}{dt}(0)$ are orthogonal to $\gamma'(0)$, then $J(t)$ is orthogonal to $\gamma'(t)$ for all $t \in [0,1]$. Indeed, the ODE for $\frac{d^2y_1}{dt^2}$ yields

$$\frac{d^2y_1}{dt^2} = 0,$$

and as $y_1(0) = 0$ and $\frac{dy_1}{dt}(0) = 0$, we get $y_1(t) = 0$ for all $t \in [0,1]$. Furthermore, if J is orthogonal to γ, which means that $J(t)$ is orthogonal to $\gamma'(t)$, for all $t \in [0,1]$, then $\frac{DJ}{dt}$ is also orthogonal to γ. Indeed, as γ is a geodesic,

$$0 = \frac{d}{dt}\langle J, \gamma'\rangle = \langle \frac{DJ}{dt}, \gamma'\rangle.$$

Therefore, the dimension of the space of Jacobi fields normal to γ is $2n-2$. These facts prove part of the following

Proposition 13.15 *If $\gamma \in \Omega(p,q)$ is a geodesic in a Riemannian manifold of dimension n, then the following properties hold:*

(1) For all $u, v \in T_pM$, there is a unique Jacobi fields, J, so that $J(0) = u$ and $\frac{DJ}{dt}(0) = v$. Consequently, the vector space of Jacobi fields has dimension n.

(2) The subspace of Jacobi fields orthogonal to γ has dimension $2n-2$. The vector fields γ' and $t \mapsto t\gamma'(t)$ are Jacobi fields that form a basis of the subspace of Jacobi fields parallel to γ (that is, such that $J(t)$ is collinear with $\gamma'(t)$, for all $t \in [0,1]$.)

(3) If J is a Jacobi field, then J is orthogonal to γ iff there exist $a, b \in [0,1]$, with $a \neq b$, so that $J(a)$ and $J(b)$ are both orthogonal to γ iff there is some $a \in [0,1]$ so that $J(a)$ and $\frac{DJ}{dt}(a)$ are both orthogonal to γ.

(4) For any two Jacobi fields, X, Y, along γ, the expression $\langle \nabla_{\gamma'} X, Y\rangle - \langle \nabla_{\gamma'} Y, X\rangle$ is a constant and if X and Y vanish at some point on γ, then $\langle \nabla_{\gamma'} X, Y\rangle - \langle \nabla_{\gamma'} Y, X\rangle = 0$.

Proof. We already proved (1) and part of (2). If J is parallel to γ, then $J(t) = f(t)\gamma(t)$ and the Jacobi equation becomes

$$\frac{d^2 f}{dt} = 0.$$

Therefore,

$$J(t) = (\alpha + \beta t)\gamma'(t).$$

It is easily shown that γ' and $t \mapsto t\gamma'(t)$ are linearly independent (as vector fields).

To prove (3), using the Jacobi equation, observe that

$$\frac{d^2}{dt^2}\langle J, \gamma'\rangle = \langle \frac{D^2 J}{dt^2}, \gamma'\rangle = -R(J, \gamma', \gamma', \gamma') = 0.$$

Therefore,

$$\langle J, \gamma'\rangle = \alpha + \beta t$$

and the result follows. We leave (4) as an exercise. □

Following Milnor, we will show that the Jacobi fields in $T_\gamma\Omega(p,q)$ are exactly the vector fields in the nullspace of the index form, I. First, we define the important notion of conjugate points.

Definition 13.8 Let $\gamma \in \Omega(p,q)$ be a geodesic. Two distinct parameter values, $a, b \in [0,1]$, with $a < b$, are *conjugate along* γ iff there is some Jacobi field, J, not identically zero, such that $J(a) = J(b) = 0$. The dimension, k, of the space, $\mathfrak{J}_{a,b}$, consisting of all such Jacobi fields is called the *multiplicity* (or *order of conjugacy)* of a and b as conjugate parameters. We also say that the points $p_1 = \gamma(a)$ and $p_2 = \gamma(b)$ are *conjugate along* γ.

Remark: As remarked by Milnor and others, as γ may have self-intersections, the above definition is ambiguous if we replace a and b by $p_1 = \gamma(a)$ and $p_2 = \gamma(b)$, even though many authors make this slight abuse. Although it makes sense to say that the points p_1 and p_2 are conjugate, the space of Jacobi fields vanishing at p_1 and p_2 is not well defined. Indeed, if $p_1 = \gamma(a)$ for distinct values of a (or $p_2 = \gamma(b)$ for distinct values of b), then we don't know which of the spaces, $\mathfrak{J}_{a,b}$, to pick. We will say that some points p_1 and p_2 on γ are *conjugate* iff there are parameter values, $a < b$, such that $p_1 = \gamma(a)$, $p_2 = \gamma(b)$, and a and b are conjugate along γ.

However, for the endpoints p and q of the geodesic segment γ, we may assume that $p = \gamma(0)$ and $q = \gamma(1)$, so that when we say that p and q are conjugate we consider the space of Jacobi fields vanishing for $t = 0$ and $t = 1$. This is the definition adopted Gallot, Hulin and Lafontaine [60] (Chapter 3, Section 3E).

In view of Proposition 13.15 (3), the Jacobi fields involved in the definition of conjugate points are orthogonal to γ. The dimension of the space of Jacobi fields such that $J(a) = 0$ is obviously n, since the only remaining parameter determining J is $\frac{dJ}{dt}(a)$. Furthermore, the Jacobi field, $t \mapsto (t-a)\gamma'(t)$, vanishes at a but not at b, so the multiplicity of conjugate parameters (points) is at most $n-1$.

For example, if M is a flat manifold, that is, iff its curvature tensor is identically zero, then the Jacobi equation becomes

$$\frac{D^2 J}{dt^2} = 0.$$

It follows that $J \equiv 0$, and thus, there are no conjugate points. More generally, the Jacobi equation can be solved explicitly for spaces of constant curvature.

Theorem 13.16 *Let $\gamma \in \Omega(p,q)$ be a geodesic. A vector field, $W \in T_\gamma\Omega(p,q)$, belongs to the nullspace of the index form, I, iff W is a Jacobi field. Hence, I is degenerate if p and q are conjugate. The nullity of I is equal to the multiplicity of p and q.*

Proof. (After Milnor [106], Theorem 14.1). We already observed that a Jacobi field vanishing at 0 and 1 belong to the nullspace of I.

Conversely, assume that $W_1 \in T_\gamma\Omega(p,q)$ belongs to the nullspace of I. Pick a subdivision, $0 = t_0 < t_1 < \cdots < t_k = 1$ of $[0,1]$ so that $W_1 \restriction [t_i, t_{i+1}]$ is smooth for all $i = 0, \ldots, k-1$ and let $f\colon [0,1] \to [0,1]$ be a smooth function which vanishes for the parameter values $t_0, \ldots, t_k$ and is strictly positive otherwise. Then, if we let

$$W_2(t) = f(t)\left(\frac{D^2 W_1}{dt^2} + R(\gamma', W_1)\gamma'\right)_t,$$

by the Second Variation Formula, we get

$$0 = -\frac{1}{2} I(W_1, W_2) = \sum 0 + \int_0^1 f(t) \left\| \frac{D^2 W_1}{dt^2} + R(\gamma', W_1)\gamma' \right\|^2 dt.$$

Consequently, $W_1 \restriction [t_i, t_{i+1}]$ is a Jacobi field for all $i = 0, \ldots, k-1$.

Now, let $W_2' \in T_\gamma\Omega(p,q)$ be a field such that

$$W_2'(t_i) = \Delta_{t_i}\frac{DW_1}{dt}, \qquad i = 1, \ldots, k-1.$$

We get

$$0 = -\frac{1}{2}I(W_1, W_2') = \sum_{i=1}^{k-1}\left\|\Delta_{t_i}\frac{DW_1}{dt}\right\|^2 + \int_0^1 0\, dt.$$

Hence, $\frac{DW_1}{dt}$ has no jumps. Now, a solution, W_1, of the Jacobi equation is completely determined by the vectors $W_1(t_i)$ and $\frac{DW_1}{dt}(t_i)$, so the k Jacobi fields, $W_1 \restriction [t_i, t_{i+1}]$, fit together to give a Jacobi field, W_1, which is smooth throughout $[0,1]$. □

Theorem 13.16 implies that the nullity of I is finite, since the vector space of Jacobi fields vanishing at 0 and 1 has dimension at most n. In fact, we observed that the dimension of this space is at most $n-1$.

Corollary 13.17 *The nullity, ν, of I satisfies $0 \leq \nu \leq n-1$, where $n = \dim(M)$.*

Jacobi fields turn out to be induced by certain kinds of variations called *geodesic variations*.

Definition 13.9 Given a geodesic, $\gamma \in \Omega(p,q)$, a *geodesic variation of* γ is a smooth map,

$$\alpha\colon (-\epsilon, \epsilon) \times [0,1] \to M,$$

such that

(1) $\alpha(0,t) = \gamma(t)$, for all $t \in [0,1]$.

(2) For every $u \in (-\epsilon, \epsilon)$, the curve $\widetilde{\alpha}(u)$ is a geodesic, where

$$\widetilde{\alpha}(u)(t) = \alpha(u,t), \qquad t \in [0,1].$$

Note that the geodesics, $\widetilde{\alpha}(u)$, do not necessarily begin at p and end at q and so, a geodesic variation is not a "fixed endpoints" variation.

Proposition 13.18 *If $\alpha\colon (-\epsilon, \epsilon) \times [0,1] \to M$ is a geodesic variation of $\gamma \in \Omega(p,q)$, then the vector field, $W(t) = \frac{\partial \alpha}{\partial u}(0,t)$, is a Jacobi field along γ.*

Proof. As α is a geodesic variation, we have

$$\frac{D}{dt}\frac{\partial \alpha}{\partial t} = 0.$$

Hence, using Proposition 13.3, we have

$$\begin{aligned} 0 &= \frac{D}{\partial u}\frac{D}{\partial t}\frac{\partial \alpha}{\partial t} \\ &= \frac{D}{\partial t}\frac{D}{\partial u}\frac{\partial \alpha}{\partial t} + R\left(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u}\right)\frac{\partial \alpha}{\partial t} \\ &= \frac{D^2}{\partial t^2}\frac{\partial \alpha}{\partial u} + R\left(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u}\right)\frac{\partial \alpha}{\partial t}, \end{aligned}$$

where we used the fact (already used before) that

$$\frac{D}{\partial t}\frac{\partial \alpha}{\partial u} = \frac{D}{\partial u}\frac{\partial \alpha}{\partial t},$$

as the Levi-Civita connection is torsion-free. □

For example, on the sphere, S^n, for any two antipodal points, p and q, rotating the sphere keeping p and q fixed, the variation field along a geodesic, γ, through p and q (a great circle) is a Jacobi field vanishing at p and q. Rotating in $n-1$ different directions one obtains $n-1$ linearly independent Jacobi fields and thus, p and q are conjugate along γ with multiplicity $n-1$.

Interestingly, the converse of Proposition 13.18 holds.

Proposition 13.19 *For every Jacobi field, $W(t)$, along a geodesic, $\gamma \in \Omega(p,q)$, there is some geodesic variation, $\alpha\colon (-\epsilon,\epsilon)\times[0,1] \to M$ of γ, such that $W(t) = \frac{\partial \alpha}{\partial u}(0,t)$. Furthermore, for every point, $\gamma(a)$, there is an open subset, U, containing $\gamma(a)$, such that the Jacobi fields along a geodesic segment in U are uniquely determined by their values at the endpoints of the geodesic.*

Proof. (After Milnor, see [106], Chapter III, Lemma 14.4.) We begin by proving the second assertion. By Proposition 12.4 (1), there is an open subset, U, with $\gamma(0) \in U$, so that any two points of U are joined by a unique minimal geodesic which depends differentially on the endpoints. Suppose that $\gamma(t) \in U$ for $t \in [0,\delta]$. We will construct a Jacobi field, W, along $\gamma \restriction [0,\delta]$ with arbitrarily prescribed values, u, at $t = 0$ and v at $t = \delta$. Choose some curve, $c_0\colon (-\epsilon,\epsilon) \to U$, so that $c_0(0) = \gamma(0)$ and $c_0'(0) = u$ and some curve, $c_\delta\colon (-\epsilon,\epsilon) \to U$, so that $c_\delta(0) = \gamma(\delta)$ and $c_\delta'(0) = v$. Now, define the map,

$$\alpha\colon (-\epsilon,\epsilon)\times[0,\delta] \to M,$$

by letting $\widetilde{\alpha}(u)\colon [0,\delta] \to M$ be the unique minimal geodesic from $c_0(u)$ to $c_\delta(u)$. It is easily checked that α is a geodesic variation of $\gamma \restriction [0,\delta]$ and that

$$J(t) = \frac{\partial \alpha}{\partial u}(0,t)$$

is a Jacobi field such that $J(0) = u$ and $J(\delta) = v$.

We claim that every Jacobi field along $\gamma \restriction [0, \delta]$ can be obtained uniquely in this way. If $\mathfrak{J}_\delta$ denotes the vector space of all Jacobi fields along $\gamma \restriction [0, \delta]$, the map $J \mapsto (J(0), J(\delta))$ defines a linear map

$$\ell \colon \mathfrak{J}_\delta \to T_{\gamma(0)}M \times T_{\gamma(\delta)}M.$$

The above argument shows that ℓ is onto. However, both vector spaces have the same dimension, $2n$, so ℓ is an isomorphism. Therefore, every Jacobi field in $\mathfrak{J}_\delta$ is determined by its values at $\gamma(0)$ and $\gamma(\delta)$.

Now, the above argument can be repeated for every point, $\gamma(a)$, on γ, so we get an open cover, $\{(l_a, r_a)\}$, of $[0, 1]$, such that every Jacobi field along $\gamma \restriction [l_a, r_a]$ is uniquely determined by its endpoints. By compactness of $[0, 1]$, the above cover possesses some finite subcover and we get a geodesic variation, α, defined on the entire interval $[0, 1]$ whose variation field is equal to the original Jacobi field, W. $\square$

Remark: The proof of Proposition 13.19 also shows that there is some open interval $(-\delta, \delta)$, such that if $t \in (-\delta, \delta)$, then $\gamma(t)$ is not conjugate to $\gamma(0)$ along γ. In fact, the Morse Index Theorem implies that for any geodesic segment, $\gamma \colon [0, 1] \to M$, there are only finitely many points which are conjugate to $\gamma(0)$ along γ (see Milnor [106], Part III, Corollary 15.2).

There is also an intimate connection between Jacobi fields and the differential of the exponential map and between conjugate points and critical points of the exponential map.

Recall that if $f \colon M \to N$ is a smooth map between manifolds, a point, $p \in M$, is a *critical* point of f iff the tangent map at p,

$$df_p \colon T_pM \to T_{f(p)}N,$$

is not surjective. If M and N have the same dimension, which will be the case in the sequel, df_p is not surjective iff it is not injective, so p is a critical point of f iff there is some nonzero vector, $u \in T_pM$, such that $df_p(u) = 0$.

If $\exp_p \colon T_pM \to M$ is the exponential map, for any $v \in T_pM$ where $\exp_p(v)$ is defined, we have the derivative of $\exp_p$ at v;

$$(d \exp_p)_v \colon T_v(T_pM) \to T_pM.$$

Since T_pM is a finite-dimensional vector space, $T_v(T_pM)$ is isomorphic to T_pM, so we identify $T_v(T_pM)$ with T_pM.

Proposition 13.20 *Let $\gamma \in \Omega(p, q)$ be a geodesic. The point, $r = \gamma(t)$, with $t \in (0, 1]$, is conjugate to p along γ iff $v = t\gamma'(0)$ is a critical point of $\exp_p$. Furthermore, the multiplicity of p and r as conjugate points is equal to the dimension of the kernel of $(d \exp_p)_v$.*

A proof of Proposition 13.20 can be found in various places, including Do Carmo [50] (Chapter 5, Proposition 3.5), O'Neill [119] (Chapter 10, Proposition 10), or Milnor [106] (Part III, Theorem 18.1).

Using Proposition 13.19 it is easy to characterize conjugate points in terms of geodesic variations.

Proposition 13.21 *If $\gamma \in \Omega(p,q)$ is a geodesic, then q is conjugate to p iff there is a geodesic variation, α, of γ, such that every geodesic, $\widetilde{\alpha}(u)$, starts from p, the Jacobi field, $J(t) = \frac{\partial \alpha}{\partial u}(0,t)$ does not vanish identically, and $J(1) = 0$.*

Jacobi fields can also be used to compute the derivative of the exponential (see Gallot, Hulin and Lafontaine [60], Chapter 3, Corollary 3.46).

Proposition 13.22 *Given any point, $p \in M$, for any vectors $u, v \in T_pM$, if $\exp_p v$ is defined, then*

$$J(t) = (d\exp_p)_{tv}(tu), \qquad 0 \leq t \leq 1,$$

is a Jacobi field such that $\frac{DJ}{dt}(0) = u$.

Remark: If $u, v \in T_pM$ are orthogonal unit vectors, then $R(u,v,u,v) = K(u,v)$, the sectional curvature of the plane spanned by u and v in T_pM, and for t small enough, we have

$$\|J(t)\| = t - \frac{1}{6}\,K(u,v)t^3 + o(t^3).$$

(Here, $o(t^3)$ stands for an expression of the form $t^4R(t)$, such that $\lim_{t \mapsto 0} R(t) = 0$.) Intuitively, this formula tells us how fast the geodesics that start from p and are tangent to the plane spanned by u and v spread apart. Locally, for $K(u,v) > 0$, the radial geodesics spread apart less than the rays in T_pM and for $K(u,v) < 0$, they spread apart more than the rays in T_pM. More details, see Do Carmo [50] (Chapter 5, Section 2).

There is also another version of "Gauss lemma" (see Gallot, Hulin and Lafontaine [60], Chapter 3, Lemma 3.70):

Proposition 13.23 *(Gauss Lemma) Given any point, $p \in M$, for any vectors $u, v \in T_pM$, if $\exp_p v$ is defined, then*

$$\langle d(\exp_p)_{tv}(u), d(\exp_p)_{tv}(v)\rangle = \langle u, v\rangle, \qquad 0 \leq t \leq 1.$$

As our (connected) Riemannian manifold, M, is a metric space, the path space, $\Omega(p,q)$, is also a metric space if we use the metric, d^*, given by

$$d^*(\omega_1, \omega_2) = \max_t(d(\omega_1(t), \omega_2(t))),$$

where d is the metric on M induced by the Riemannian metric.

Remark: The topology induced by d^* turns out to be the compact open topology on $\Omega(p,q)$.

Theorem 13.24 *Let $\gamma \in \Omega(p,q)$ be a geodesic. Then, the following properties hold:*

(1) If there are no conjugate points to p along γ, then there is some open subset, $\mathcal{V}$, of $\Omega(p,q)$, with $\gamma \in \mathcal{V}$, such that

$$L(\omega) \geq L(\gamma) \quad \textit{and} \quad E(\omega) \geq E(\gamma), \qquad \textit{for all } \omega \in \mathcal{V},$$

with strict inequality when $\omega([0,1]) \neq \gamma([0,1])$. We say that γ is a local minimum.

(2) If there is some $t \in (0,1)$ such that p and $\gamma(t)$ are conjugate along γ, then there is a fixed endpoints variation, α, such that

$$L(\widetilde{\alpha}(u)) < L(\gamma) \quad \textit{and} \quad E(\widetilde{\alpha}(u)) < E(\gamma), \qquad \textit{for } u \textit{ small enough.}$$

A proof of Theorem 13.24 can be found in Gallot, Hulin and Lafontaine [60] (Chapter 3, Theorem 3.73) or in O'Neill [119] (Chapter 10, Theorem 17 and Remark 18).

13.8 Convexity, Convexity Radius

Proposition 12.4 shows that if (M,g) is a Riemannian manifold, then for every point, $p \in M$, there is an open subset, $W \subseteq M$, with $p \in W$ and a number $\epsilon > 0$, so that any two points q_1, q_2 of W are joined by a unique geodesic of length $< \epsilon$. However, there is no garantee that this unique geodesic between q_1 and q_2 stays inside W. Intuitively this says that W may not be convex.

The notion of convexity can be generalized to Riemannian manifolds but there are some subtleties. In this short section, we review various definition or convexity found in the literature and state one basic result. Following Sakai [130] (Chapter IV, Section 5), we make the following definition:

Definition 13.10 Let $C \subseteq M$ be a nonempty subset of some Riemannian manifold, M.

(1) The set C is called *strongly convex* iff for any two points, $p, q \in C$, there exists a unique minimal geodesic, γ, from p to q in M and γ is contained in C.

(2) If for every point, $p \in \overline{C}$, there is some $\epsilon(p) > 0$, so that $C \cap B_{\epsilon(p)}(p)$ is strongly convex, then we say that C is *locally convex* (where $B_{\epsilon(p)}(p)$ is the metric ball of center 0 and radius $\epsilon(p)$).

(3) The set C is called *totally convex* iff for any two points, $p, q \in C$, all geodesics from p to q in M are contained in C.

It is clear that if C is strongly convex or totally convex, then C is locally convex. If M is complete and any two points are joined by a unique geodesic, then the three conditions of Definition 13.10 are equivalent. The next Proposition will show that a metric ball with sufficiently small radius is strongly convex.

Definition 13.11 For any $p \in M$, the *convexity radius at* p, denoted, $r(p)$, is the least upper bound of the numbers, $r > 0$, such that for any metric ball, $B_\epsilon(q)$, if $B_\epsilon(q) \subseteq B_r(p)$, then $B_\epsilon(q)$ is strongly convex and every geodesic contained in $B_r(p)$ is a minimal geodesic joining its endpoints. The *convexity radius of* M, $r(M)$, as the greatest lower bound of the set $\{r(p) \mid p \in M\}$.

Note that it is possible that $r(p) = 0$ if M is not compact.

The following proposition is proved in Sakai [130] (Chapter IV, Section 5, Theorem 5.3).

Proposition 13.25 *If M is a Riemannian manifold, then $r(p) > 0$ for every $p \in M$ and the map, $p \mapsto r(p) \in \mathbb{R}_+ \cup \{\infty\}$ is continuous. Furthermore, if $r(p) = \infty$ for some $p \in M$, then $r(q) = \infty$ for all $q \in M$.*

That $r(p) > 0$ is also proved in Do Carmo [50] (Chapter 3, Section 4, Proposition 4.2). More can be said about the structure of connected locally convex subsets of M, see Sakai [130] (Chapter IV, Section 5).

Remark: The following facts are stated in Berger [16] (Chapter 6):

(1) If M is compact, then the convexity radius, $r(M)$, is strictly positive.

(2) $r(M) \leq \frac{1}{2} i(M)$, where $i(M)$ is the injectivity radius of M.

Berger also points out that if M is compact, then the existence of a finite cover by convex balls can used to triangulate M. This method was proposed by Hermann Karcher (see Berger [16], Chapter 3, Note 3.4.5.3).

13.9 Applications of Jacobi Fields and Conjugate Points

Jacobi fields and conjugate points are basic tools that can be used to prove many global results of Riemannian geometry. The flavor of these results is that certain constraints on curvature (sectional, Ricci, sectional) have a significant impact on the topology. One may want consider the effect of non-positive curvature, constant curvature, curvature bounded from below by a positive constant, *etc*. This is a vast subject and we highly recommend Berger's Panorama of Riemannian Geometry [16] for a masterly survey. We will content ourselves with three results:

(1) Hadamard and Cartan's Theorem about complete manifolds of non-positive sectional curvature.

(2) Myers' Theorem about complete manifolds of Ricci curvature bounded from below by a positive number.

(3) The Morse Index Theorem.

First, on the way to Hadamard and Cartan we begin with a proposition.

Proposition 13.26 *Let M be a complete Riemannian manifold with non-positive curvature, $K \leq 0$. Then, for every geodesic, $\gamma \in \Omega(p,q)$, there are no conjugate points to p along γ. Consequently, the exponential map, $\exp_p \colon T_pM \to M$, is a local diffeomorphism for all $p \in M$.*

Proof. Let J be a Jacobi field along γ. Then,

$$\frac{D^2J}{dt^2} + R(\gamma', J)\gamma' = 0$$

so that, by the definition of the sectional curvature,

$$\left\langle \frac{D^2J}{dt^2}, J \right\rangle = -\langle R(\gamma', J)\gamma', J) = -R(\gamma', J, \gamma', J) \geq 0.$$

It follows that

$$\frac{d}{dt}\left\langle \frac{DJ}{dt}, J \right\rangle = \left\langle \frac{D^2J}{dt^2}, J \right\rangle + \left\| \frac{DJ}{dt} \right\|^2 \geq 0.$$

Thus, the function, $t \mapsto \left\langle \frac{DJ}{dt}, J \right\rangle$ is monotonic increasing and, strictly so if $\frac{DJ}{dt} \neq 0$. If J vanishes at both 0 and t, for any given $t \in (0,1]$, then so does $\left\langle \frac{DJ}{dt}, J \right\rangle$, and hence $\left\langle \frac{DJ}{dt}, J \right\rangle$ must vanish throughout the interval $[0,t]$. This implies

$$J(0) = \frac{DJ}{dt}(0) = 0,$$

so that J is identically zero. Therefore, t is not conjugate to 0 along γ. □

Theorem 13.27 *(Hadamard–Cartan) Let M be a complete Riemannian manifold. If M has non-positive sectional curvature, $K \leq 0$, then the following hold:*

(1) For every $p \in M$, the map, $\exp_p \colon T_pM \to M$, is a Riemannian covering.

(2) If M is simply connected then M is diffeomorphic to $\mathbb{R}^n$, where $n = \dim(M)$; more precisely, $\exp_p \colon T_pM \to M$ is a diffeomorphism for all $p \in M$. Furthermore, any two points on M are joined by a unique minimal geodesic.

Proof. We follow the proof in Sakai [130] (Chapter V, Theorem 4.1).

(1) By Proposition 13.26, the exponential map, $\exp_p \colon T_pM \to M$, is a local diffeomorphism for all $p \in M$. Let $\widetilde{g}$ be the pullback metric, $\widetilde{g} = (\exp_p)^* g$, on T_pM (where g denotes the metric on M). We claim that $(T_pM, \widetilde{g})$ is complete.

This is because, for every nonzero $u \in T_pM$, the line, $t \mapsto tu$, is mapped to the geodesic, $t \mapsto \exp_p(tu)$, in M, which is defined for all $t \in \mathbb{R}$ since M is complete, and thus, this line is a geodedic in $(T_pM, \widetilde{g})$. Since this holds for all $u \in T_pM$, $(T_pM, \widetilde{g})$ is geodesically complete at 0, so by Hopf-Rinow, it is complete. But now, $\exp_p \colon T_pM \to M$ is a local isometry and by Proposition 13.11, it is a Riemannian covering map.

(2) If M is simply connected, then by Proposition 3.38, the covering map $\exp_p \colon T_pM \to M$ is a diffeomorphism (T_pM is connected). Therefore, $\exp_p \colon T_pM \to M$ is a diffeomorphism for all $p \in M$. $\square$

Other proofs of Theorem 13.27 can be found in Do Carmo [50] (Chapter 7, Theorem 3.1), Gallot, Hulin and Lafontaine [60] (Chapter 3, Theorem 3.87), Kobayashi and Nomizu [90] (Chapter VIII, Theorem 8.1) and Milnor [106] (Part III, Theorem 19.2).

Remark: A version of Theorem 13.27 was first proved by Hadamard and then extended by Cartan.

Theorem 13.27 was generalized by Kobayashi, see Kobayashi and Nomizu [90] (Chapter VIII, Remark 2 after Corollary 8.2). Also, it is shown in Milnor [106] that if M is complete, assuming non-positive sectional curvature, then all homotopy groups, $\pi_i(M)$, vanish, for $i > 1$, and that $\pi_1(M)$ has no element of finite order except the identity. Finally, non-positive sectional curvature implies that the exponential map does not decrease distance (Kobayashi and Nomizu [90], Chapter VIII, Section 8, Lemma 3).

We now turn to manifolds with strictly positive curvature bounded away from zero and to Myers' Theorem. The first version of such a theorem was first proved by Bonnet for surfaces with positive sectional curvature bounded away from zero. It was then generalized by Myers in 1941. For these reasons, this theorem is sometimes called the *Bonnet-Myers' Theorem*. The proof of Myers Theorem involves a beautiful "trick".

Given any metric space, X, recall that the *diameter* of X is defined by

$$\operatorname{diam}(X) = \sup\{d(p, q) \mid p, q \in X\}.$$

The diameter of X may be infinite.

Theorem 13.28 *(Myers) Let M be a complete Riemannian manifold of dimension n and assume that*

$$\operatorname{Ric}(u, u) \geq (n-1)/r^2, \qquad \textit{for all unit vectors, } u \in T_pM, \textit{ and for all } p \in M,$$

with $r > 0$. Then,

(1) The diameter of M is bounded by πr and M is compact.

(2) The fundamental group of M is finite.

Proof. (1) Pick any two points $p, q \in M$ and let $d(p,q) = L$. As M is complete, by Hopf and Rinow's Theorem, there is a minimal geodesic, γ, joining p and q and by Proposition 13.14, the bilinear index form, I, associated with γ is positive semi-definite, which means that $I(W,W) \geq 0$, for all vector fields, $W \in T_\gamma\Omega(p,q)$. Pick an orthonormal basis, $(e_1, \ldots, e_n)$, of T_pM, with $e_1 = \gamma'(0)/L$. Using parallel transport, we get a field of orthonormal frames, $(X_1, \ldots, X_n)$, along γ, with $X_1(t) = \gamma'(t)/L$. Now comes Myers' beautiful trick. Define new vector fields, Y_i, along γ, by

$$W_i(t) = \sin(\pi t)X_i(t), \qquad 2 \leq i \leq n.$$

We have

$$\gamma'(t) = LX_1 \quad \text{and} \quad \frac{DX_i}{dt} = 0.$$

Then, by the second variation formula,

$$\begin{aligned} \frac{1}{2}\, I(W_i, W_i) &= -\int_0^1 \left\langle W_i, \frac{D^2W_i}{dt^2} + R(\gamma', W_i)\gamma' \right\rangle dt \\ &= \int_0^1 (\sin(\pi t))^2(\pi^2 - L^2 \langle R(X_1, X_i)X_1, X_i\rangle)dt, \end{aligned}$$

for $i = 2, \ldots, n$. Adding up these equations and using the fact that

$$\mathrm{Ric}(X_1(t), X_1(t)) = \sum_{i=2}^n \langle R(X_1(t), X_i(t))X_1(t), X_i(t)\rangle,$$

we get

$$\frac{1}{2}\sum_{i=2}^n I(W_i, W_i) = \int_0^1 (\sin(\pi t))^2[(n-1)\pi^2 - L^2\,\mathrm{Ric}(X_1(t), X_1(t))]dt.$$

Now, by hypothesis,

$$\mathrm{Ric}(X_1(t), X_1(t)) \geq (n-1)/r^2,$$

so

$$0 \leq \frac{1}{2}\sum_{i=2}^n I(W_i, W_i) \leq \int_0^1 (\sin(\pi t))^2 \left[(n-1)\pi^2 - (n-1)\frac{L^2}{r^2}\right] dt,$$

which implies $\frac{L^2}{r^2} \leq \pi^2$, that is

$$d(p,q) = L \leq \pi r.$$

As the above holds for every pair of points, $p, q \in M$, we conclude that

$$\mathrm{diam}(M) \leq \pi r.$$

Since closed and bounded subsets in a complete manifold are compact, M itself must be compact.

(2) Since the universal covering space, $\widetilde{M}$, of M has the pullback of the metric on M, this metric satisfies the same assumption on its Ricci curvature as that of M. Therefore, $\widetilde{M}$ is also compact, which implies that the fundamental group, $\pi_1(M)$, is finite (see the discussion at the end of Section 3.9). □

Remarks:

(1) The condition on the Ricci curvature cannot be weakened to $\mathrm{Ric}(u, u) > 0$ for all unit vectors. Indeed, the paraboloid of revolution, $z = x^2 + y^2$, satisfies the above condition, yet it is not compact.

(2) Theorem 13.28 also holds under the stronger condition that the sectional curvature $K(u, v)$ satisfies

$$K(u, v) \geq (n-1)/r^2,$$

for all orthonormal vectors, u, v. In this form, it is due to Bonnet (for surfaces).

It would be a pity not to include in this section a beautiful theorem due to Morse.

Theorem 13.29 *(Morse Index Theorem) Given a geodesic, $\gamma \in \Omega(p, q)$, the index, λ, of the index form, $I \colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$, is equal to the number of points, $\gamma(t)$, with $0 \leq t \leq 1$, such that $\gamma(t)$ is conjugate to $p = \gamma(0)$ along γ, each such conjugate point counted with its multiplicity. The index λ is always finite.*

As a corollary of Theorem 13.29, we see that there are only finitely many points which are conjugate to $p = \gamma(0)$ along γ.

A proof of Theorem 13.29 can be found in Milnor [106] (Part III, Section 15) and also in Do Carmo [50] (Chapter 11) or Kobayashi and Nomizu [90] (Chapter VIII, Section 6).

A key ingredient of the proof is that the vector space, $T_\gamma\Omega(p, q)$, can be split into a direct sum of subspaces mutually orthogonal with respect to I, on one of which (denoted T') I is positive definite. Furthermore, the subspace orthogonal to T' is finite-dimensional. This space is obtained as follows: Since for every point, $\gamma(t)$, on γ, there is some open subset, U_t, containing $\gamma(t)$ such that any two points in U_t are joined by a unique minimal geodesic, by compactness of $[0, 1]$, there is a subdivision, $0 = t_0 < t_1 < \cdots < t_k = 1$ of $[0, 1]$ so that $\gamma \restriction [t_i, t_{i+1}]$ lies within an open where it is a minimal geodesic.

Let $T_\gamma\Omega(t_0, \ldots, t_k) \subseteq T_\gamma\Omega(p, q)$ be the vector space consisting of all vector fields, W, along γ such that

(1) $W \restriction [t_i, t_{i+1}]$ is a Jacobi field along $\gamma \restriction [t_i, t_{i+1}]$, for $i = 0, \ldots, k-1$.

(2) $W(0) = W(1) = 0$.

The space $T_\gamma\Omega(t_0,\ldots,t_k) \subseteq T_\gamma\Omega(p,q)$ is a finite-dimensional vector space consisting of broken Jacobi fields. Let $T' \subseteq T_\gamma\Omega(p,q)$ be the vector space consisting of all vector fields, $W \in T_\gamma\Omega(p,q)$, for which

$$W(t_i) = 0, \qquad 0 \leq i \leq k.$$

It is not hard to prove that

$$T_\gamma\Omega(p,q) = T_\gamma\Omega(t_0,\ldots,t_k) \oplus T',$$

that $T_\gamma\Omega(t_0,\ldots,t_k)$ and T' are orthogonal *w.r.t* I and that $I \restriction T'$ is positive definite. The reason why $I(W,W) \geq 0$ for $W \in T'$ is that each segment, $\gamma \restriction [t_i, t_{i+1}]$, is a minimal geodesic, which has smaller energy than any other path between its endpoints.

As a consequence, the index (or nullity) of I is equal to the index (or nullity) of I restricted to the finite dimensional vector space, $T_\gamma\Omega(t_0,\ldots,t_k)$. This shows that the index is always finite.

In the next section, we will use conjugate points to give a more precise characterization of the cut locus.

13.10 Cut Locus and Injectivity Radius: Some Properties

We begin by reviewing the definition of the cut locus from a slightly different point of view. Let M be a complete Riemannian manifold of dimension n. There is a bundle, UM, called the *unit tangent bundle*, such that the fibre at any $p \in M$ is the unit sphere, $S^{n-1} \subseteq T_pM$ (check the details). As usual, we let $\pi\colon UM \to M$ denote the projection map which sends every point in the fibre over p to p. Then, we have the function,

$$\rho\colon UM \to \mathbb{R},$$

defined so that for all $p \in M$, for all $v \in S^{n-1} \subseteq T_pM$,

$$\begin{aligned}\rho(v) &= \sup_{t\in\mathbb{R}\cup\{\infty\}} d(\pi(v), \exp_p(tv)) = t\\ &= \sup\{t \in \mathbb{R}\cup\{\infty\} \mid \text{the geodesic} \quad t \mapsto \exp_p(tv) \quad \text{is minimal on } [0,t]\}.\end{aligned}$$

The number $\rho(v)$ is called the *cut value* of v. It can be shown that ρ is continuous and for every $p \in M$, we let

$$\widetilde{\mathrm{Cut}}(p) = \{\rho(v)v \in T_pM \mid v \in UM \cap T_pM,\ \rho(v) \text{ is finite}\}$$

be the *tangential cut locus of* p and

$$\mathrm{Cut}(p) = \exp_p(\widetilde{\mathrm{Cut}}(p))$$

be the *cut locus of* p. The point, $\exp_p(\rho(v)v)$, in M is called the *cut point* of the geodesic, $t \mapsto \exp_p(vt)$, and so, the cut locus of p is the set of cut points of all the geodesics emanating from p. Also recall from Definition 12.7 that

$$\mathcal{U}_p = \{v \in T_pM \mid \rho(v) > 1\}$$

and that $\mathcal{U}_p$ is open and star-shaped. It can be shown that

$$\widetilde{\mathrm{Cut}}(p) = \partial \mathcal{U}_p$$

and the following property holds:

Theorem 13.30 *If M is a complete Riemannian manifold, then for every $p \in M$, the exponential map, $\exp_p$, is a diffeomorphism between $\mathcal{U}_p$ and its image, $\exp_p(\mathcal{U}_p) = M - \mathrm{Cut}(p)$, in M.*

Proof. The fact that $\exp_p$ is injective on $\mathcal{U}_p$ was shown in Proposition 12.16. Now, for any $v \in \mathcal{U}$, as $t \mapsto \exp_p(tv)$ is a minimal geodesic for $t \in [0,1]$, by Theorem 13.24 (2), the point $\exp_p v$ is not conjugate to p, so $d(\exp_p)_v$ is bijective, which implies that $\exp_p$ is a local diffeomorphism. As $\exp_p$ is also injective, it is a diffeomorphism. □

Theorem 13.30 implies that the cut locus is closed.

Remark: In fact, $M - \mathrm{Cut}(p)$ can be retracted homeomorphically onto a ball around p and $\mathrm{Cut}(p)$ is a deformation retract of $M - \{p\}$.

The following Proposition gives a rather nice characterization of the cut locus in terms of minimizing geodesics and conjugate points:

Proposition 13.31 *Let M be a complete Riemannian manifold. For every pair of points, $p, q \in M$, the point q belongs to the cut locus of p iff one of the two (not mutually exclusive from each other) properties hold:*

(a) There exist two distinct minimizing geodesics from p to q.

(b) There is a minimizing geodesic, γ, from p to q and q is the first conjugate point to p along γ.

A proof of Proposition 13.31 can be found in Do Carmo [50] (Chapter 13, Proposition 2.2) Kobayashi and Nomizu [90] (Chapter VIII, Theorem 7.1) or Klingenberg [88] (Chapter 2, Lemma 2.1.11).

Observe that Proposition 13.31 implies the following symmetry property of the cut locus: $q \in \mathrm{Cut}(p)$ iff $p \in \mathrm{Cut}(q)$. Furthermore, if M is compact, we have

$$p = \bigcap_{q \in \mathrm{Cut}(p)} \mathrm{Cut}(q).$$

Proposition 13.31 admits the following sharpening:

Proposition 13.32 *Let M be a complete Riemannian manifold. For all $p, q \in M$, if $q \in \mathrm{Cut}(p)$, then:*

(a) If among the minimizing geodesics from p to q, there is one, say γ, such that q is not conjugate to p along γ, then there is another minimizing geodesic $\omega \neq \gamma$ from p to q.

(b) Suppose $q \in \mathrm{Cut}(p)$ realizes the distance from p to $\mathrm{Cut}(p)$ (i.e., $d(p,q) = d(p, \mathrm{Cut}(p))$). If there are no minimal geodesics from p to q such that q is conjugate to p along this geodesic, then there are exactly two minimizing geodesics, γ_1 and γ_2, from p to q, with $\gamma_2'(1) = -\gamma_1'(1)$. Moreover, if $d(p,q) = i(M)$ (the injectivity radius), then γ_1 and γ_2 together form a closed geodesic.

Except for the last statement, Proposition 13.32 is proved in Do Carmo [50] (Chapter 13, Proposition 2.12). The last statement is from Klingenberg [88] (Chapter 2, Lemma 2.1.11).

We also have the following characterization of $\widetilde{\mathrm{Cut}}(p)$:

Proposition 13.33 *Let M be a complete Riemannian manifold. For any $p \in M$, the set of vectors, $u \in \widetilde{\mathrm{Cut}}(p)$, such that is some $v \in \widetilde{\mathrm{Cut}}(p)$ with $v \neq u$ and $\exp_p(u) = \exp_p(v)$, is dense in $\widetilde{\mathrm{Cut}}(p)$.*

Proposition 13.33 is proved in Klingenberg [88] (Chapter 2, Theorem 2.1.14).

We conclude this section by stating a classical theorem of Klingenberg about the injectivity radius of a manifold of bounded positive sectional curvature.

Theorem 13.34 *(Klingenberg) Let M be a complete Riemannian manifold and assume that there are some positive constants, $K_{\min}$, $K_{\max}$, such that the sectional curvature of K satisfies*

$$0 < K_{\min} \leq K \leq K_{\max}.$$

Then, M is compact and either

(a) $i(M) \geq \pi/\sqrt{K_{\max}}$, or

(b) There is a closed geodesic, γ, of minimal length among all closed geodesics in M and such that

$$i(M) = \frac{1}{2}\, L(\gamma).$$

The proof of Theorem 13.34 is quite hard. A proof using Rauch's comparison Theorem can be found in Do Carmo [50] (Chapter 13, Proposition 2.13).

Chapter 14

Discrete Curvatures and Geodesics on Polyhedral Surfaces

Chapter 15

The Laplace-Beltrami Operator, Harmonic Forms, The Connection Laplacian and Weitzenböck Formulae

15.1 The Gradient, Hessian and Hodge $*$ Operators on Riemannian Manifolds

The Laplacian is a very important operator because it shows up in many of the equations used in physics to describe natural phenomena such as heat diffusion or wave propagation. Therefore, it is highly desirable to generalize the Laplacian to functions defined on a manifold. Furthermore, in the late 1930's George de Rham (inspired by Élie Cartan) realized that it was fruitful to define a version of the Laplacian operating on differential forms, because of a fundamental and almost miraculous relationship between harmonics forms (those in the kernel of the Laplacian) and the de Rham cohomology groups on a (compact, orientable) smooth manifold. Indeed, as we will see in Section 15.3, for every cohomology group, $H^k_{\mathrm{DR}}(M)$, every cohomology class, $[\omega] \in H^k_{\mathrm{DR}}(M)$, is represented by a *unique harmonic k-form*, ω. This connection between analysis and topology lies deep and has many important consequences. For example, *Poincaré duality* follows as an "easy" consequence of the Hodge Theorem.

Technically, the Laplacian can be defined on differential forms using the Hodge $*$ operator (Section 22.16). On functions, there are alternate definitions of the Laplacian using only the covariant derivative and obtained by generalizing the notions of gradient and divergence to functions on manifolds.

Another version of the Laplacian can be defined in terms of the adjoint of the connection, ∇, on differential forms, viewed as a linear map from $\mathcal{A}^*(M)$ to $\mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$. We obtain the *connection Laplacian* (also called *Bochner Laplacian*), $\nabla^*\nabla$. Then, it is natural to wonder how the Hodge Laplacian, Δ, differs from the connection Laplacian, $\nabla^*\nabla$? Remarkably, there is a formula known as *Weitzenböck's formula* (or *Bochner's formula*) of

the form

$$\Delta = \nabla^*\nabla + C(R_\nabla),$$

where $C(R_\nabla)$ is a contraction of a version of the curvature tensor on differential forms (a fairly complicated term). In the case of one-forms,

$$\Delta = \nabla^*\nabla + \mathrm{Ric},$$

where Ric is a suitable version of the Ricci curvature operating on one-forms.

Weitzenböck-type formulae are at the root of the so-called "Bochner Technique", which consists in exploiting curvature information to deduce topological information. For example, if the Ricci curvature on a compact, orientable Riemannian manifold is strictly positive, then $H^1_{\mathrm{DR}}(M) = (0)$, a theorem due to Bochner.

If $(M, \langle -, - \rangle)$ is a Riemannian manifold of dimension n, then for every $p \in M$, the inner product, $\langle -, - \rangle_p$, on T_pM yields a canonical isomorphism, $\flat \colon T_pM \to T_p^*M$, as explained in Sections 22.1 and 11.5. Namely, for any $u \in T_pM$, $u^\flat = \flat(u)$ is the linear form in T_p^*M defined by

$$u^\flat(v) = \langle u, v \rangle_p, \qquad v \in T_pM.$$

Recall that the inverse of the map $\flat$ is the map $\sharp \colon T_p^*M \to T_pM$. As a consequence, for every smooth function, $f \in C^\infty(M)$, we get smooth vector field, $\mathrm{grad}\, f = (df)^\sharp$, defined so that

$$(\mathrm{grad}\, f)_p = (df_p)^\sharp,$$

that is, we have

$$\langle (\mathrm{grad}\, f)_p, u \rangle_p = df_p(u), \quad \text{for all} \quad u \in T_pM.$$

The vector field, $\mathrm{grad}\, f$, is the *gradient* of the function f.

Conversely, a vector field, $X \in \mathfrak{X}(M)$, yields the one-form, $X^\flat \in \mathcal{A}^1(M)$, given by

$$(X^\flat)_p = (X_p)^\flat.$$

The *Hessian*, $\mathrm{Hess}(f)$, (or $\nabla^2(f)$) of a function, $f \in C^\infty(M)$, is the $(0,2)$-tensor defined by

$$\mathrm{Hess}(f)(X, Y) = X(Y(f)) - (\nabla_X Y)(f) = X(df(Y)) - df(\nabla_X Y),$$

for all vector fields, $X, Y \in \mathfrak{X}(M)$.

Recall from Proposition 11.5 that the covariant derivative, $\nabla_X \theta$, of any one-form, $\theta \in \mathcal{A}^1(M)$, is the one-form given by

$$(\nabla_X \theta)(Y) = X(\theta(Y)) - \theta(\nabla_X Y)$$

Recall from Proposition 11.5 that the covariant derivative, $\nabla_X \theta$, of any one-form, $\theta \in \mathcal{A}^1(M)$, is the one-form given by

$$(\nabla_X \theta)(Y) = X(\theta(Y)) - \theta(\nabla_X Y)$$

so, the Hessian, $\mathrm{Hess}(f)$, is also defined by

$$\mathrm{Hess}(f)(X,Y) = (\nabla_X df)(Y).$$

Since ∇ is torsion-free, we get

$$\mathrm{Hess}(f)(X,Y) = X(Y(f)) - (\nabla_X Y)(f) = Y(X(f)) - (\nabla_Y X)(f) = \mathrm{Hess}(f)(Y,X),$$

which means that the Hessian is a *symmetric* $(0,2)$-tensor. We also have the equation

$$\mathrm{Hess}(f)(X,Y) = \langle \nabla_X \operatorname{grad} f, Y\rangle.$$

Indeed,

$$\begin{aligned} X(Y(f)) &= X(df(Y)) \\ &= X(\langle \operatorname{grad} f, Y\rangle) \\ &= \langle \nabla_X \operatorname{grad} f, Y\rangle + \langle \operatorname{grad} f, \nabla_X Y\rangle \\ &= \langle \nabla_X \operatorname{grad} f, Y\rangle + (\nabla_X Y)(f) \end{aligned}$$

which yields

$$\langle \nabla_X \operatorname{grad} f, Y\rangle = X(Y(f)) - (\nabla_X Y)(f) = \mathrm{Hess}(f)(X,Y).$$

A function, $f \in C^\infty(M)$, is *convex* (resp. *strictly convex*) iff its Hessian, $\mathrm{Hess}(f)$, is positive semi-definite (resp. positive definite).

By the results of Section 22.16, the inner product, $\langle -, -\rangle_p$, on T_pM induces an inner product on $\bigwedge^k T_p^*M$. Therefore, for any two k-forms, $\omega, \eta \in \mathcal{A}^k(M)$, we get the smooth function, $\langle \omega, \eta\rangle$, given by

$$\langle \omega, \eta\rangle(p) = \langle \omega_p, \eta_p\rangle_p.$$

Furthermore, if M is oriented, then we can apply the results of Section 22.16 so the vector bundle, T^*M, is oriented (by giving T_p^*M the orientation induced by the orientation of T_pM, for every $p \in M$) and for every $p \in M$, we get a Hodge $*$-operator,

$$*\colon \bigwedge^k T_p^*M \to \bigwedge^{n-k} T_p^*M.$$

Then, given any k-form, $\omega \in \mathcal{A}^k(M)$, we can define $*\omega$ by

$$(*\omega)_p = *(\omega_p), \qquad p \in M.$$

We have to check that $*\omega$ is indeed a smooth form in $\mathcal{A}^{n-k}(M)$, but this is not hard to do in local coordinates (for help, see Morita [114], Chapter 4, Section 1). Therefore, if M is a Riemannian oriented manifold of dimension n, we have Hodge $*$-operators,

$$*\colon \mathcal{A}^k(M) \to \mathcal{A}^{n-k}(M).$$

Observe that $*1$ is just the volume form, Vol_M, induced by the metric. Indeed, we know from Section 22.1 that in local coordinates, $x_1, \ldots, x_n$, near p, the metric on T_p^*M is given by the inverse, (g^{ij}), of the metric, (g_{ij}), on T_pM and by the results of Section 22.16,

$$\begin{aligned} *(1) &= \frac{1}{\sqrt{\det(g^{ij})}}\, dx_1 \wedge \cdots \wedge dx_n \\ &= \sqrt{\det(g_{ij})}\, dx_1 \wedge \cdots \wedge dx_n = \mathrm{Vol}_M. \end{aligned}$$

Proposition 22.25 yields the following:

Proposition 15.1 *If M is a Riemannian oriented manifold of dimension n, then we have the following properties:*

(i) $*(f\omega + g\eta) = f\, *\omega + g\, *\eta$, *for all* $\omega, \eta \in \mathcal{A}^k(M)$ *and all* $f, g \in C^\infty(M)$.

(ii) $** = (-\mathrm{id})^{k(n-k)}$.

(iii) $\omega \wedge *\eta = \eta \wedge *\omega = \langle \omega, \eta \rangle \,\mathrm{Vol}_M$, *for all* $\omega, \eta \in \mathcal{A}^k(M)$.

(iv) $*(\omega \wedge *\eta) = *(\eta \wedge *\omega) = \langle \omega, \eta \rangle$, *for all* $\omega, \eta \in \mathcal{A}^k(M)$.

(v) $\langle *\omega, *\eta \rangle = \langle \omega, \eta \rangle$, *for all* $\omega, \eta \in \mathcal{A}^k(M)$.

Recall that exterior differentiation, d, is a map, $d\colon \mathcal{A}^k(M) \to \mathcal{A}^{k+1}(M)$. Using the Hodge $*$-operator, we can define an operator, $\delta\colon \mathcal{A}^k(M) \to \mathcal{A}^{k-1}(M)$, that will turn out to be adjoint to d with respect to an inner product on $\mathcal{A}^\bullet(M)$.

Definition 15.1 Let M be an oriented Riemannian manifold of dimension n. For any k, with $1 \leq k \leq n$, let

$$\delta = (-1)^{n(k+1)+1} * d * .$$

Clearly, δ is a map, $\delta\colon \mathcal{A}^k(M) \to \mathcal{A}^{k-1}(M)$, and $\delta = 0$ on $\mathcal{A}^0(M) = C^\infty(M)$. It is easy to see that

$$*\delta = (-1)^k d*, \quad \delta* = (-1)^{k+1} * d, \quad \delta \circ \delta = 0.$$

15.2 The Laplace-Beltrami and Divergence Operators on Riemannian Manifolds

Using d and δ, we can generalize the Laplacian to an operator on differential forms.

Definition 15.2 Let M be an oriented Riemannian manifold of dimension n. The *Laplace-Beltrami operator*, for short, *Laplacian*, is the operator, $\Delta\colon \mathcal{A}^k(M) \to \mathcal{A}^k(M)$, defined by

$$\Delta = d\delta + \delta d.$$

A form, $\omega \in \mathcal{A}^k(M)$, such that $\Delta\omega = 0$ is a *harmonic form*. In particular, a function, $f \in \mathcal{A}^0(M) = C^\infty(M)$, such that $\Delta f = 0$ is called a *harmonic function*.

The Laplacian in Definition 15.2 is also called the *Hodge Laplacian*.

If $M = \mathbb{R}^n$ with the Euclidean metric and f is a smooth function, a laborious computation yields

$$\Delta f = -\sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2},$$

that is, the usual Laplacian *with a negative sign* in front (the computation can be found in Morita [114], Example 4.12 or Jost [83], Chapter 2, Section 2.1). It is also easy to see that Δ commutes with $*$, that is,

$$\Delta * = *\Delta.$$

Given any vector field, $X \in \mathfrak{X}(M)$, its *divergence*, $\operatorname{div} X$, is defined by

$$\operatorname{div} X = \delta X^\flat.$$

Now, for a function, $f \in C^\infty(M)$, we have $\delta f = 0$, so $\Delta f = \delta df$. However,

$$\operatorname{div}(\operatorname{grad} f) = \delta(\operatorname{grad} f)^\flat = \delta((df)^\sharp)^\flat = \delta df,$$

so

$$\Delta f = \operatorname{div}\operatorname{grad} f,$$

as in the case of $\mathbb{R}^n$.

Remark: Since the definition of δ involves two occurrences of the Hodge $*$-operator, δ also makes sense on non-orientable manifolds by using a local definition. Therefore, the Laplacian, Δ, also makes sense on non-orientable manifolds.

In the rest of this section, we assume that M is orientable.

The relationship between δ and d can be made clearer by introducing an inner product on forms with compact support. Recall that $\mathcal{A}_c^k(M)$ denotes the space of k-forms with compact support (an infinite dimensional vector space). For any two k-forms with compact support, $\omega, \eta \in \mathcal{A}_c^k(M)$, set

$$(\omega, \eta) = \int_M \langle \omega, \eta \rangle \operatorname{Vol}_M = \int_M \langle \omega, \eta \rangle \ * (1).$$

Using Proposition 15.1, we have

$$(\omega, \eta) = \int_M \langle \omega, \eta \rangle \operatorname{Vol}_M = \int_M \omega \wedge *\eta = \int_M \eta \wedge *\omega,$$

so it is easy to check that $(-,-)$ is indeed an inner product on k-forms with compact support. We can extend this inner product to forms with compact support in $\mathcal{A}_c^\bullet(M) = \bigoplus_{k=0}^n \mathcal{A}_c^k(M)$ by making $\mathcal{A}_c^h(M)$ and $\mathcal{A}_c^k(M)$ orthogonal if $h \neq k$.

Proposition 15.2 *If M is an orientable Riemannian manifold, then δ is (formally) adjoint to d, that is,*

$$(d\omega, \eta) = (\omega, \delta\eta),$$

for all k-forms, ω, η, with compact support.

Proof. By linearity and orthogonality of the $\mathcal{A}_c^k(M)$ the proof reduces to the case where $\omega \in \mathcal{A}_c^{k-1}(M)$ and $\eta \in \mathcal{A}_c^k(M)$ (both with compact support). By definition of δ and the fact that

$$** = (-\mathrm{id})^{(k-1)(n-k+1)}$$

for $*\colon \mathcal{A}^{k-1}(M) \to \mathcal{A}^{n-k+1}(M)$, we have

$$*\delta = (-1)^k d*,$$

and since

$$\begin{aligned} d(\omega \wedge *\eta) &= d\omega \wedge *\eta + (-1)^{k-1}\omega \wedge d * \eta \\ &= d\omega \wedge *\eta - \omega \wedge *\delta\eta \end{aligned}$$

we get

$$\begin{aligned} \int_M d(\omega \wedge *\eta) &= \int_M d\omega \wedge *\eta - \int_M \omega \wedge *\delta\eta \\ &= (d\omega, \eta) - (\omega, \delta\eta). \end{aligned}$$

However, by Stokes Theorem (Theorem 9.7),

$$\int_M d(\omega \wedge *\eta) = 0,$$

so $(d\omega, \eta) - (\omega, \delta\eta) = 0$, that is, $(d\omega, \eta) = (\omega, \delta\eta)$, as claimed. □

Corollary 15.3 *If M is an orientable Riemannian manifold, then the Laplacian, Δ is self-adjoint that is,*

$$(\Delta\omega, \eta) = (\omega, \Delta\eta),$$

for all k-forms, ω, η, with compact support.

We also obtain the following useful fact:

Proposition 15.4 *If M is an orientable Riemannian manifold, then for every k-form, ω, with compact support, $\Delta\omega = 0$ iff $d\omega = 0$ and $\delta\omega = 0$.*

Proof. Since $\Delta = d\delta + \delta d$, is is obvious that if $d\omega = 0$ and $\delta\omega = 0$, then $\Delta\omega = 0$. Conversely,

$$(\Delta\omega, \omega) = ((d\delta + \delta d)\omega, \omega) = (d\delta\omega, \omega) + (\delta d\omega, \omega) = (\delta\omega, \delta\omega) + (d\omega, d\omega).$$

Thus, if $\Delta\omega = 0$, then $(\delta\omega, \delta\omega) = (d\omega, d\omega) = 0$, which implies $d\omega = 0$ and $\delta\omega = 0$. □

As a consequence of Proposition 15.4, if M is a connected, orientable, compact Riemannian manifold, then every harmonic function on M is a constant.

For practical reasons, we need a formula for the Laplacian of a function, $f \in C^\infty(M)$, in local coordinates. If (U, φ) is a chart near p, as usual, let

$$\frac{\partial f}{\partial x_j}(p) = \frac{\partial (f \circ \varphi^{-1})}{\partial u_j}(\varphi(p)),$$

where $(u_1, \ldots, u_n)$ are the coordinate functions in $\mathbb{R}^n$. Write $|g| = \det(g_{ij})$, where (g_{ij}) is the symmetric, positive definite matrix giving the metric in the chart (U, φ).

Proposition 15.5 *If M is an orientable Riemannian manifold, then for every local chart, (U, φ), for every function, $f \in C^\infty(M)$, we have*

$$\Delta f = -\frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, g^{ij}\, \frac{\partial f}{\partial x_j} \right).$$

Proof. We follow Jost [83], Chapter 2, Section 1. Pick any function, $h \in C^\infty(M)$, with compact support. We have

$$\begin{aligned}
\int_M (\Delta f) h * (1) &= (\Delta f, h) \\
&= (\delta df, h) \\
&= (df, dh) \\
&= \int_M \langle df, dh \rangle * (1) \\
&= \int_M \sum_{ij} g^{ij} \frac{\partial f}{\partial x_i} \frac{\partial h}{\partial x_j} * (1) \\
&= -\int_M \sum_{ij} \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x_j} \left(\sqrt{|g|}\, g^{ij} \frac{\partial f}{\partial x_i} \right) h * (1),
\end{aligned}$$

where we have used integration by parts in the last line. Since the above equation holds for all h, we get our result. □

It turns out that in a Riemannian manifold, the divergence of a vector field and the Laplacian of a function can be given a definition that uses the covariant derivative (see Chapter 11, Section 11.1) instead of the Hodge $*$-operator. For the sake of completeness, we present this alternate definition which is the one used in Gallot, Hulin and Lafontaine [60] (Chapter 4) and O'Neill [119] (Chapter 3). If ∇ is the Levi-Civita connection induced by the Riemannian metric, then the *divergence* of a vector field, $X \in \mathfrak{X}(M)$, is the function, $\operatorname{div} X \colon M \to \mathbb{R}$, defined so that

$$(\operatorname{div} X)(p) = \operatorname{tr}(Y(p) \mapsto (-\nabla_Y X)_p),$$

namely, for every p, $(\operatorname{div} X)(p)$ is the trace of the linear map, $Y(p) \mapsto (-\nabla_Y X)_p$. Of course, for any function, $f \in C^\infty(M)$, we define Δf by

$$\Delta f = \operatorname{div} \operatorname{grad} f.$$

Observe that the above definition of the divergence (and of the Laplacian) makes sense even if M is non-orientable. For orientable manifolds, the equivalence of this new definition of the divergence with our definition is proved in Petersen [121], see Chapter 3, Proposition 31. The main reason is that

$$L_X \operatorname{Vol}_M = -(\operatorname{div} X)\operatorname{Vol}_M$$

and by Cartan's Formula (Proposition 8.15), $L_X = i(X) \circ d + d \circ i(X)$, as $d\operatorname{Vol}_M = 0$, we get

$$(\operatorname{div} X)\operatorname{Vol}_M = -d(i(X)\operatorname{Vol}_M).$$

The above formulae also holds for a local volume form (*i.e.* for a volume form on a local chart).

The operator, $\delta \colon \mathcal{A}^1(M) \to \mathcal{A}^0(M)$, can also be defined in terms of the covariant derivative (see Gallot, Hulin and Lafontaine [60], Chapter 4). For any one-form, $\omega \in \mathcal{A}^1(M)$, recall that

$$(\nabla_X \omega)(Y) = X(\omega(Y)) - \omega(\nabla_X Y).$$

Then, it turns out that

$$\delta\omega = -\operatorname{tr} \nabla\omega,$$

where the trace should be interpreted as the trace of the $\mathbb{R}$-bilinear map, $X, Y \mapsto (\nabla_X \omega)(Y)$, as in Chapter 22, see Proposition 22.2. This means that in any chart, (U, φ),

$$\delta\omega = -\sum_{i=1}^{n} (\nabla_{E_i} \omega)(E_i),$$

for any orthonormal frame field, $(E_1, \ldots, E_n)$ over U. It can be shown that

$$\delta(f df) = f\Delta f - \langle \operatorname{grad} f, \operatorname{grad} f \rangle,$$

and, as a consequence,

$$(\Delta f, f) = \int_M \langle \operatorname{grad} f, \operatorname{grad} f \rangle \operatorname{Vol}_M,$$

for any orientable, compact manifold, M.

Since the proof of the next proposition is quite technical, we omit the proof.

Proposition 15.6 *If M is an orientable and compact Riemannian manifold, then for every vector field, $X \in \mathfrak{X}(M)$, we have*

$$\operatorname{div} X = \delta X^\flat.$$

Consequently, for the Laplacian, we have

$$\Delta f = \delta df = \operatorname{div} \operatorname{grad} f.$$

Remark: Some authors omit the negative sign in the definition of the divergence, that is, they define

$$(\operatorname{div} X)(p) = \operatorname{tr}(Y(p) \mapsto (\nabla_Y X)_p).$$

Here is a frequently used corollary of Proposition 15.6:

Proposition 15.7 *(Green's Formula) If M is an orientable and compact Riemannian manifold without boundary, then for every vector field, $X \in \mathfrak{X}(M)$, we have*

$$\int_M (\operatorname{div} X) \operatorname{Vol}_M = 0.$$

Proofs of proposition 15.7 can be found in Gallot, Hulin and Lafontaine [60] (Chapter 4, Proposition 4.9) and Helgason [72] (Chapter 2, Section 2.4).

There is a generalization of the formula expressing $\delta\omega$ over an orthonormal frame, $E_1, \ldots, E_n$, for a one-form, ω, that applies to any differential form. In fact, there are formulae expressing both d and δ over an orthornormal frame and its coframe and these are often handy in proofs. Recall that for every vector field, $X \in \mathfrak{X}(M)$, the interior product, $i(X)\colon \mathcal{A}^{k+1}(M) \to \mathcal{A}^k(M)$, is defined by

$$(i(X)\omega)(Y_1, \ldots, Y_k) = \omega(X, Y_1, \ldots, Y_k),$$

for all $Y_1, \ldots, Y_k \in \mathfrak{X}(M)$.

Proposition 15.8 *Let M be a compact, orientable, Riemannian manifold. For every $p \in M$, for every local chart, (U, φ), with $p \in M$, if $(E_1, \ldots, E_n)$ is an orthonormal frame over U and $(\theta_1, \ldots, \theta_n)$ is its dual coframe, then for every k-form, $\omega \in \mathcal{A}^k(M)$, we have:*

$$\begin{aligned} d\omega &= \sum_{i=1}^n \theta_i \wedge \nabla_{E_i}\omega \\ \delta\omega &= -\sum_{i=1}^n i(E_i)\nabla_{E_i}\omega. \end{aligned}$$

A proof of Proposition 15.8 can be found in Petersen [121] (Chapter 7, proposition 37) or Jost [83] (Chapter 3, Lemma 3.3.4). When ω is a one-form, $\delta\omega_p$ is just a number and indeed,

$$\delta\omega = -\sum_{i=1}^{n} i(E_i)\nabla_{E_i}\omega = -\sum_{i=1}^{n}(\nabla_{E_i}\omega)(E_i),$$

as stated earlier.

15.3 Harmonic Forms, the Hodge Theorem, Poincaré Duality

Let us now assume that M is orientable and compact.

Definition 15.3 Let M be an orientable and compact Riemannian manifold of dimension n. For every k, with $0 \leq k \leq n$, let

$$\mathbb{H}^k(M) = \{\omega \in \mathcal{A}^k(M) \mid \Delta\omega = 0\},$$

the space of *harmonic k-forms*.

The following proposition is left as an easy exercise:

Proposition 15.9 *Let M be an orientable and compact Riemannian manifold of dimension n. The Laplacian commutes with the Hodge $*$-operator and we have a linear map,*

$$*\colon \mathbb{H}^k(M) \to \mathbb{H}^{n-k}(M).$$

One of the deepest and most important theorems about manifolds is the *Hodge decomposition theorem* which we now state.

Theorem 15.10 *(Hodge Decomposition Theorem) Let M be an orientable and compact Riemannian manifold of dimension n. For every k, with $0 \leq k \leq n$, the space, $\mathbb{H}^k(M)$, is finite dimensional and we have the following orthogonal direct sum decomposition of the space of k-forms:*

$$\mathcal{A}^k(M) = \mathbb{H}^k(M) \oplus d(\mathcal{A}^{k-1}(M)) \oplus \delta(\mathcal{A}^{k+1}(M)).$$

The proof of Theorem 15.10 involves a lot of analysis and it is long and complicated. A complete proof can be found in Warner [147], Chapter 6. Other treatments of Hodge theory can be found in Morita [114] (Chapter 4) and Jost [83] (Chapter 2).

The Hodge Decomposition Theorem has a number of important corollaries, one of which is *Hodge Theorem*:

Theorem 15.11 *(Hodge Theorem) Let M be an orientable and compact Riemannian manifold of dimension n. For every k, with $0 \leq k \leq n$, there is an isomorphism between $\mathbb{H}^k(M)$ and the de Rham cohomology vector space, $H^k_{\mathrm{DR}}(M)$:*

$$H^k_{\mathrm{DR}}(M) \cong \mathbb{H}^k(M).$$

Proof. Since by Proposition 15.4, every harmonic form, $\omega \in \mathbb{H}^k(M)$, is closed, we get a linear map from $\mathbb{H}^k(M)$ to $H^k_{\mathrm{DR}}(M)$ by assigning its cohomology class, $[\omega]$, to ω. This map is injective. Indeed if $[\omega] = 0$ for some $\omega \in \mathbb{H}^k(M)$, then $\omega = d\eta$, for some $\eta \in \mathcal{A}^{k-1}(M)$ so

$$(\omega, \omega) = (d\eta, \omega) = (\eta, \delta\omega).$$

But, as $\omega \in \mathbb{H}^k(M)$ we have $\delta\omega = 0$ by Proposition 15.4, so $(\omega, \omega) = 0$, that is, $\omega = 0$.

Our map is also surjective, this is the hard part of Hodge Theorem. By the Hodge Decomposition Theorem, for every closed form, $\omega \in \mathcal{A}^k(M)$, we can write

$$\omega = \omega_H + d\eta + \delta\theta,$$

with $\omega_H \in \mathbb{H}^k(M)$, $\eta \in \mathcal{A}^{k-1}(M)$ and $\theta \in \mathcal{A}^{k+1}(M)$. Since ω is closed and $\omega_H \in \mathbb{H}^k(M)$, we have $d\omega = 0$ and $d\omega_H = 0$, thus

$$d\delta\theta = 0$$

and so

$$0 = (d\delta\theta, \theta) = (\delta\theta, \delta\theta),$$

that is, $\delta\theta = 0$. Therefore, $\omega = \omega_H + d\eta$, which implies $[\omega] = [\omega_H]$, with $\omega_H \in \mathbb{H}^k(M)$, proving the surjectivity of our map. □

The Hodge Theorem also implies the *Poincaré Duality Theorem*. If M is a compact, orientable, n-dimensional smooth manifold, for each k, with $0 \leq k \leq n$, we define a bilinear map,

$$((-,-))\colon H^k_{\mathrm{DR}}(M) \times H^{n-k}_{\mathrm{DR}}(M) \longrightarrow \mathbb{R},$$

by setting

$$(([\omega], [\eta])) = \int_M \omega \wedge \eta.$$

We need to check that this definition does not depend on the choice of closed forms in the cohomology classes $[\omega]$ and $[\eta]$. However, as $d\omega = d\eta = 0$, we have

$$d(\alpha \wedge \eta + (-1)^k \omega \wedge \beta + \alpha \wedge d\beta) = d\alpha \wedge \eta + \omega \wedge d\beta + d\alpha \wedge d\beta,$$

so by Stokes' Theorem,

$$\begin{aligned} \int_M (\omega + d\alpha) \wedge (\eta + d\beta) &= \int_M \omega \wedge \eta + \int_M d(\alpha \wedge \eta + (-1)^k \omega \wedge \beta + \alpha \wedge d\beta) \\ &= \int_M \omega \wedge \eta. \end{aligned}$$

Theorem 15.12 *(Poincaré Duality) If M is a compact, orientable, smooth manifold of dimension n, then the bilinear map*

$$((-,-))\colon H^k_{\mathrm{DR}}(M) \times H^{n-k}_{\mathrm{DR}}(M) \longrightarrow \mathbb{R}$$

defined above is a nondegenerate pairing and hence, yields an isomorphism

$$H^k_{\mathrm{DR}}(M) \cong (H^{n-k}_{\mathrm{DR}}(M))^*.$$

Proof. Pick any Riemannian metric on M. It is enough to show that for every nonzero cohomology class, $[\omega] \in H^k_{\mathrm{DR}}(M)$, there is some $[\eta] \in H^{n-k}_{\mathrm{DR}}(M)$ such that

$$(([\omega],[\eta])) = \int_M \omega \wedge \eta \neq 0.$$

By Hodge Theorem, we may assume that ω is a nonzero harmonic form. By Proposition 15.9, $\eta = *\omega$ is also harmonic and $\eta \in \mathbb{H}^{n-k}(M)$. Then, we get

$$(\omega,\omega) = \int_M \omega \wedge *\omega = (([\omega],[\eta]))$$

and indeed, $(([\omega],[\eta])) \neq 0$, since $\omega \neq 0$. □

15.4 The Connection Laplacian, Weitzenböck Formula and the Bochner Technique

If M is compact, orientable, Riemannian manifold, then the inner product, $\langle -,-\rangle_p$, on T_pM (with $p \in M$) induces an inner product on differential forms, as we explained in Section 15.2. We also get an inner product on vector fields if, for any two vector field, $X, Y \in \mathfrak{X}(M)$, we define (X,Y) by

$$(X,Y) = \int_M \langle X,Y\rangle \,\mathrm{Vol}_M,$$

where $\langle X,Y\rangle$ is the function defined pointwise by

$$\langle X,Y\rangle(p) = \langle X(p), Y(p)\rangle_p.$$

Using Proposition 11.5, we can define the covariant derivative, $\nabla_X\omega$, of any k-form, $\omega \in \mathcal{A}^k(M)$, as the k-form given by

$$(\nabla_X\omega)(Y_1,\ldots,Y_k) = X(\omega(Y_1,\ldots,Y_k)) - \sum_{j=1}^{k} \omega(Y_1,\ldots,\nabla_X Y_j,\ldots,Y_k).$$

We can view ∇ as linear map,

$$\nabla\colon \mathcal{A}^k(M) \to \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M)),$$

where $\nabla\omega$ is the $C^\infty(M)$-linear map, $X \mapsto \nabla_X\omega$. The inner product on $\mathcal{A}^k(M)$ allows us to define the (formal) adjoint, ∇^*, of ∇, as a linear map

$$\nabla^* \colon \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M)) \to \mathcal{A}^k(M).$$

For any linear map, $A \in \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$, let A^* be the adjoint of A defined by

$$(AX, \theta) = (X, A^*\theta),$$

for all vector fields $X \in \mathfrak{X}(M)$ and all k-forms, $\theta \in \mathcal{A}^k(M)$. It can be verified that $A^* \in \mathrm{Hom}_{C^\infty(M)}(\mathcal{A}^k(M), \mathfrak{X}(M))$. Then, given $A, B \in \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$, the expression $\mathrm{tr}(A^*B)$ is a smooth function on M and it can be verified that

$$\langle A, B\rangle = \mathrm{tr}(A^*B)$$

defines a non-degenerate pairing on $\mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$. Using this pairing we obtain the ($\mathbb{R}$-valued) inner product on $\mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ given by

$$(A, B) = \int_M \mathrm{tr}(A^*B)\,\mathrm{Vol}_M.$$

Using all this, the (formal) adjoint, ∇^*, of $\nabla \colon \mathcal{A}^k(M) \to \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ is the linear map, $\nabla^* \colon \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M)) \to \mathcal{A}^k(M)$, defined implicitly by

$$(\nabla^* A, \omega) = (A, \nabla\omega),$$

that is,

$$\int_M \langle \nabla^* A, \omega\rangle\,\mathrm{Vol}_M = \int_M \langle A, \nabla\omega\rangle\,\mathrm{Vol}_M,$$

for all $A \in \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ and all $\omega \in \mathcal{A}^k(M)$.

The notation ∇^* for the adjoint of ∇ should not be confused with the dual connection on T^*M of a connection, ∇, on TM! Here, ∇ denotes the connection on $\mathcal{A}^*(M)$ induced by the orginal connection, ∇, on TM. The argument type (differential form or vector field) should make it clear which ∇ is intended but it might have been better to use a notation such as $\nabla^\top$ instead of ∇^*.

What we just did also applies to $\mathcal{A}^*(M) = \bigoplus_{k=0}^n \mathcal{A}^k(M)$ (where $\dim(M) = n$) and so we can view the connection, ∇, as a linear map, $\nabla \colon \mathcal{A}^*(M) \to \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$ and its adjoint as a linear map, $\nabla^* \colon \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M)) \to \mathcal{A}^*(M)$.

Definition 15.4 Given a compact, orientable, Riemannian manifold, M, the *connection Laplacian* (or *Bochner Laplacian*), $\nabla^*\nabla$, is defined as the composition of the connection, $\nabla \colon \mathcal{A}^*(M) \to \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$, with its adjoint,
$\nabla^* \colon \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M)) \to \mathcal{A}^*(M)$, as defined above.

Observe that

$$(\nabla^*\nabla\omega, \omega) = (\nabla\omega, \nabla\omega) = \int_M \langle \nabla\omega, \nabla\omega\rangle \,\mathrm{Vol}_M,$$

for all $\omega \in \mathcal{A}^k(M)$. Consequently, the "harmonic forms", ω, with respect to $\nabla^*\nabla$ must satisfy

$$\nabla\omega = 0,$$

but this condition is not equivalent to the harmonicity of ω with respect to the Hodge Laplacian. Thus, in general, $\nabla^*\nabla$ and Δ are different operators. The relationship between the two is given by formulae involving contractions of the curvature tensor and known as *Weitzenböck formulae*. We will state such a formula in case of one-forms later on. But first, we can give another definition of the connection Laplacian using *second covariant derivatives* of forms. Given any k-form, $\omega \in \mathcal{A}^k(M)$, for any two vector fields, $X, Y \in \mathfrak{X}(M)$, we define $\nabla^2_{X,Y}\omega$ by

$$\nabla^2_{X,Y}\omega = \nabla_X(\nabla_Y\omega) - \nabla_{\nabla_X Y}\omega.$$

Given any local chart, (U, φ), and given any orthormal frame, $(E_1, \ldots, E_n)$, over U, we can take the trace, $\mathrm{tr}(\nabla^2\omega)$, of $\nabla^2_{X,Y}\omega$, defined by

$$\mathrm{tr}(\nabla^2\omega) = \sum_{i=1}^{n} \nabla^2_{E_i,E_i}\omega.$$

It is easily seen that $\mathrm{tr}(\nabla^2\omega)$ does not depend on the choice of local chart and orthonormal frame.

Proposition 15.13 *If is M a compact, orientable, Riemannian manifold, then the connection Laplacian, $\nabla^*\nabla$, is given by*

$$\nabla^*\nabla\omega = -\mathrm{tr}(\nabla^2\omega),$$

for all differential forms, $\omega \in \mathcal{A}^*(M)$.

The proof of Proposition 15.13, which is quite technical, can be found in Petersen [121] (Chapter 7, Proposition 34).

We are now ready to prove the Weitzenböck formulae for one-forms.

Theorem 15.14 *(Weitzenböck–Bochner Formula) If is M a compact, orientable, Riemannian manifold, then for every one-form, $\omega \in \mathcal{A}^1(M)$, we have*

$$\Delta\omega = \nabla^*\nabla\omega + \mathrm{Ric}(\omega),$$

where $\mathrm{Ric}(\omega)$ *is the one-form given by*

$$\mathrm{Ric}(\omega)(X) = \omega(\mathrm{Ric}^\sharp(X)),$$

where $\mathrm{Ric}^\sharp$ *is the Ricci curvature viewed as a $(1,1)$-tensor (that is, $\langle \mathrm{Ric}^\sharp(u), v\rangle_p = \mathrm{Ric}(u, v)$, for all $u, v \in T_pM$ and all $p \in M$).*

Proof. For any $p \in M$, pick any normal local chart, (U, φ), with $p \in U$, and pick any orthonormal frame, $(E_1, \ldots, E_n)$, over U. Because (U, φ) is a normal chart, at p, we have $(\nabla_{E_j} E_j)_p = 0$ for all i, j. Recall from the discussion at the end of Section 15.2 that for every one-form, ω, we have

$$\delta\omega = -\sum_i \nabla_{E_i}\omega(E_i),$$

and so

$$d\delta\omega = -\sum_i \nabla_X \nabla_{E_i}\omega(E_i).$$

Also recall that

$$d\omega(X, Y) = \nabla_X \omega(Y) - \nabla_Y \omega(X),$$

and using Proposition 15.8 we can show that

$$\delta d\omega(X) = -\sum_i \nabla_{E_i}\nabla_{E_i}\omega(X) + \sum_i \nabla_{E_i}\nabla_X \omega(E_i).$$

Thus, we get

$$\begin{aligned}
\Delta\omega(X) &= -\sum_i \nabla_{E_i}\nabla_{E_i}\omega(X) + \sum_i (\nabla_{E_i}\nabla_X - \nabla_X\nabla_{E_i})\omega(E_i) \\
&= -\sum_i \nabla^2_{E_i,E_i}\omega(X) + \sum_i (\nabla^2_{E_i,X} - \nabla^2_{X,E_i})\omega(E_i) \\
&= \nabla^*\nabla\omega(X) + \sum_i \omega(R(E_i, X)E_i) \\
&= \nabla^*\nabla\omega(X) + \omega(\mathrm{Ric}^\sharp(X)),
\end{aligned}$$

using the fact that $(\nabla_{E_j} E_j)_p = 0$ for all i, j and using Proposition 13.2 and Proposition 15.13. □

For simplicity of notation, we will write $\mathrm{Ric}(u)$ for $\mathrm{Ric}^\sharp(u)$. There should be no confusion since $\mathrm{Ric}(u, v)$ denotes the Ricci curvature, a $(0, 2)$-tensor. There is another way to express $\mathrm{Ric}(\omega)$ which will be useful in the proof of the next theorem. Observe that

$$\begin{aligned}
\mathrm{Ric}(\omega)(Z) &= \omega(\mathrm{Ric}(Z)) \\
&= \langle \omega^\sharp, \mathrm{Ric}(Z)\rangle \\
&= \langle \mathrm{Ric}(Z), \omega^\sharp\rangle \\
&= \mathrm{Ric}(Z, \omega^\sharp) \\
&= \mathrm{Ric}(\omega^\sharp, Z) \\
&= \langle \mathrm{Ric}(\omega^\sharp), Z\rangle \\
&= (\mathrm{Ric}(\omega^\sharp))^\flat(Z),
\end{aligned}$$

and thus,

$$\mathrm{Ric}(\omega)(Z) = (\mathrm{Ric}(\omega^\sharp))^\flat(Z).$$

Consequently the Weitzenböck formula can be written as

$$\Delta\omega = \nabla^*\nabla\omega + (\mathrm{Ric}(\omega^\sharp))^\flat.$$

The Weitzenböck–Bochner Formula implies the following theorem due to Bochner:

Theorem 15.15 *(Bochner) If M is a compact, orientable, connected Riemannian manifold, then the following properties hold:*

(i) If the Ricci curvature is non-negative, that is $\mathrm{Ric}(u,u) \geq 0$ *for all* $p \in M$ *and all* $u \in T_pM$ *and if* $\mathrm{Ric}(u,u) > 0$ *for some* $p \in M$ *and all* $u \in T_pM$*, then* $H^1_{\mathrm{DR}}M = (0)$.

(ii) If the Ricci curvature is non-negative, then $\nabla\omega = 0$ *for all* $\omega \in \mathcal{A}^1(M)$ *and* $\dim H^1_{\mathrm{DR}}M \leq \dim M$.

Proof. (i) Assume $H^1_{\mathrm{DR}}M \neq (0)$. Then, by the Hodge Theorem, there is some nonzero harmonic one-form, ω. The Weitzenböck–Bochner Formula implies that

$$(\Delta\omega,\omega) = (\nabla^*\nabla\omega,\omega) + ((\mathrm{Ric}(\omega^\sharp))^\flat,\omega).$$

Since $\Delta\omega = 0$, we get

$$\begin{aligned} 0 &= (\nabla^*\nabla\omega,\omega) + ((\mathrm{Ric}(\omega^\sharp))^\flat,\omega) \\ &= (\nabla\omega,\nabla\omega) + \int_M \langle(\mathrm{Ric}(\omega^\sharp))^\flat,\omega\rangle \,\mathrm{Vol}_M \\ &= (\nabla\omega,\nabla\omega) + \int_M \langle\mathrm{Ric}(\omega^\sharp),\omega^\sharp\rangle \,\mathrm{Vol}_M \\ &= (\nabla\omega,\nabla\omega) + \int_M \mathrm{Ric}(\omega^\sharp,\omega^\sharp) \,\mathrm{Vol}_M. \end{aligned}$$

However, $(\nabla\omega,\nabla\omega) \geq 0$ and by the assumption on the Ricci curvature, the integrand is nonnegative and strictly positive at some point, so the integral is strictly positive, a contradiction.

(ii) Again, for any one-form, ω, we have

$$(\Delta\omega,\omega) = (\nabla\omega,\nabla\omega) + \int_M \mathrm{Ric}(\omega^\sharp,\omega^\sharp) \,\mathrm{Vol}_M,$$

and so, if the Ricci curvature is non-negative, $\Delta\omega = 0$ iff $\nabla\omega = 0$. This means that ω is invariant by parallel transport (see Section 11.3) and thus, ω is completely determined by its value, ω_p, at some point, $p \in M$, so there is an injection, $\mathbb{H}^1(M) \longrightarrow T^*_pM$, which implies that $\dim H^1_{\mathrm{DR}}M = \dim \mathbb{H}^1(M) \leq \dim M$. □

There is a version of the Weitzenböck formula for p-forms but it involves a more complicated curvature term and its proof is also more complicated. The Bochner technique can

also be generalized in various ways, in particular, to *spin manifolds*, but these considerations are beyond the scope of these notes. Let me just say that Weitzenböck formulae involving the Dirac operator play an important role in physics and 4-manifold geometry. We refer the interested reader to Gallot, Hulin and Lafontaine [60] (Chapter 4) Petersen [121] (Chapter 7), Jost [83] (Chaper 3) and Berger [16] (Section 15.6) for more details on Weitzenböck formulae and the Bochner technique.

Chapter 16

Spherical Harmonics and Linear Representations of Lie Groups

16.1 Introduction, Spherical Harmonics on the Circle

In this chapter, we discuss spherical harmonics and take a glimpse at the linear representation of Lie groups. Spherical harmonics on the sphere, S^2, have interesting applications in computer graphics and computer vision so this material is not only important for theoretical reasons but also for practical reasons.

Joseph Fourier (1768-1830) invented Fourier series in order to solve the heat equation [55]. Using Fourier series, every square-integrable periodic function, f, (of period 2π) can be expressed uniquely as the sum of a power series of the form

$$f(\theta) = a_0 + \sum_{k=1}^{\infty}(a_k \cos k\theta + b_k \cos k\theta),$$

where the *Fourier coefficients*, a_k, b_k, of f are given by the formulae

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)\, d\theta, \quad a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\cos k\theta\, d\theta, \quad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\sin k\theta\, d\theta,$$

for $k \geq 1$. The reader will find the above formulae in Fourier's famous book [55] in Chapter III, Section 233, page 256, essentially using the notation that we use nowdays.

This remarkable discovery has many theoretical and practical applications in physics, signal processing, engineering, *etc.* We can describe Fourier series in a more conceptual manner if we introduce the following inner product on square-integrable functions:

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(\theta)g(\theta)\, d\theta,$$

which we will also denote by

$$\langle f, g\rangle = \int_{S^1} f(\theta)g(\theta)\, d\theta,$$

where S^1 denotes the unit circle. After all, periodic functions of (period 2π) can be viewed as functions on the circle. With this inner product, the space $L^2(S^1)$ is a complete normed vector space, that is, a Hilbert space. Furthermore, if we define the subspaces, $\mathcal{H}_k(S^1)$, of $L^2(S^1)$, so that $\mathcal{H}_0(S^1)\,(=\mathbb{R})$ is the set of constant functions and $\mathcal{H}_k(S^1)$ is the two-dimensional space spanned by the functions $\cos k\theta$ and $\sin k\theta$, then it turns out that we have a Hilbert sum decomposition

$$L^2(S^1) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^1)$$

into pairwise orthogonal subspaces, where $\bigcup_{k=0}^{\infty} \mathcal{H}_k(S^1)$ is dense in $L^2(S^1)$. The functions $\cos k\theta$ and $\sin k\theta$ are also orthogonal in $\mathcal{H}_k(S^1)$.

Now, it turns out that the spaces, $\mathcal{H}_k(S^1)$, arise naturally when we look for homogeneous solutions of the Laplace equation, $\Delta f = 0$, in $\mathbb{R}^2$ (Pierre-Simon Laplace, 1749-1827). Roughly speaking, a homogeneous function in $\mathbb{R}^2$ is a function that can be expressed in polar coordinates, (r, θ), as

$$f(r, \theta) = r^k g(\theta).$$

Recall that the Laplacian on $\mathbb{R}^2$ expressed in cartesian coordinates, (x, y), is given by

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2},$$

where $f\colon \mathbb{R}^2 \to \mathbb{R}$ is a function which is at least of class C^2. In polar coordinates, (r, θ), where $(x, y) = (r\cos\theta, r\sin\theta)$ and $r > 0$, the Laplacian is given by

$$\Delta f = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2}.$$

If we restrict f to the unit circle, S^1, then the Laplacian on S^1 is given by

$$\Delta_{s^1} f = \frac{\partial^2 f}{\partial \theta^2}.$$

It turns out that *the space $\mathcal{H}_k(S^1)$ is the eigenspace of Δ_{S^1} for the eigenvalue $-k^2$.*

To show this, we consider another question, namely, *what are the harmonic functions on* $\mathbb{R}^2$, that is, the functions, f, that are solutions of the *Laplace equation*,

$$\Delta f = 0.$$

Our ancestors had the idea that the above equation can be solved by *separation of variables*. This means that we write $f(r, \theta) = F(r)g(\theta)$, where $F(r)$ and $g(\theta)$ are independent functions.

To make things easier, let us assume that $F(r) = r^k$, for some integer $k \geq 0$, which means that we assume that f is a *homogeneous function of degree* k. Recall that a function, $f\colon \mathbb{R}^2 \to \mathbb{R}$, is *homogeneous of degree* k iff

$$f(tx, ty) = t^k f(x, y) \qquad \text{for all } t > 0.$$

Now, using the Laplacian in polar coordinates, we get

$$\begin{aligned}
\Delta f &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial(r^k g(\theta))}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2(r^k g(\theta))}{\partial \theta^2} \\
&= \frac{1}{r}\frac{\partial}{\partial r}\left(k r^k g\right) + r^{k-2}\frac{\partial^2 g}{\partial \theta^2} \\
&= r^{k-2}k^2 g + r^{k-2}\frac{\partial^2 g}{\partial \theta^2} \\
&= r^{k-2}(k^2 g + \Delta_{S^1} g).
\end{aligned}$$

Thus, we deduce that

$$\Delta f = 0 \quad \text{iff} \quad \Delta_{S^1} g = -k^2 g,$$

that is, g is an eigenfunction of Δ_{S^1} for the eigenvalue $-k^2$. But, the above equation is equivalent to the second-order differential equation

$$\frac{d^2 g}{d\theta^2} + k^2 g = 0,$$

whose general solution is given by

$$g(\theta) = a_n \cos k\theta + b_n \sin k\theta.$$

In summary, we found that the integers, $0, -1, -4, -9, \ldots, -k^2, \ldots$ are eigenvalues of Δ_{S^1} and that the functions $\cos k\theta$ and $\sin k\theta$ are eigenfunctions for the eigenvalue $-k^2$, with $k \geq 0$. So, it looks like the dimension of the eigenspace corresponding to the eigenvalue $-k^2$ is 1 when $k = 0$ and 2 when $k \geq 1$.

It can indeed be shown that Δ_{S^1} has no other eigenvalues and that the dimensions claimed for the eigenspaces are correct. Observe that if we go back to our homogeneous harmonic functions, $f(r, \theta) = r^k g(\theta)$, we see that this space is spanned by the functions

$$u_k = r^k \cos k\theta, \quad v_k = r^k \sin k\theta.$$

Now, $(x + iy)^k = r^k(\cos k\theta + i \sin k\theta)$, and since $\Re(x + iy)^k$ and $\Im(x + iy)^k$ are homogeneous polynomials, we see that u_k and v_k are homogeneous polynomials called *harmonic polynomials*. For example, here is a list of a basis for the harmonic polynomials (in two variables) of degree $k = 0, 1, 2, 3, 4$:

$k = 0$	1
$k = 1$	$x,\ y$
$k = 2$	$x^2 - y^2,\ xy$
$k = 3$	$x^3 - 3xy^2,\ 3x^2y - y^3$
$k = 4$	$x^4 - 6x^2y^2 + y^4,\ x^3y - xy^3.$

Therefore, the eigenfunctions of the Laplacian on S^1 are the restrictions of the harmonic polynomials on $\mathbb{R}^2$ to S^1 and we have a Hilbert sum decomposition, $L^2(S^1) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^1)$. It turns out that this phenomenon generalizes to the sphere $S^n \subseteq \mathbb{R}^{n+1}$ for all $n \geq 1$.

Let us take a look at next case, $n = 2$.

16.2 Spherical Harmonics on the 2-Sphere

The material of section is very classical and can be found in many places, for example Andrews, Askey and Roy [2] (Chapter 9), Sansone [132] (Chapter III), Hochstadt [78] (Chapter 6) and Lebedev [97] (Chapter). We recommend the exposition in Lebedev [97] because we find it particularly clear and uncluttered. We have also borrowed heavily from some lecture notes by Hermann Gluck for a course he offered in 1997-1998.

Our goal is to find the homogeneous solutions of the Laplace equation, $\Delta f = 0$, in $\mathbb{R}^3$, and to show that they correspond to spaces, $\mathcal{H}_k(S^2)$, of eigenfunctions of the Laplacian, Δ_{S^2}, on the 2-sphere,

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

Then, the spaces $\mathcal{H}_k(S^2)$ will give us a Hilbert sum decomposition of the Hilbert space, $L^2(S^2)$, of square-integrable functions on S^2. This is the generalization of Fourier series to the 2-sphere and the functions in the spaces $\mathcal{H}_k(S^2)$ are called *spherical harmonics*.

The Laplacian in $\mathbb{R}^3$ is of course given by

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

If we use spherical coordinates

$$\begin{aligned} x &= r \sin\theta \cos\varphi \\ y &= r \sin\theta \sin\varphi \\ z &= r \cos\theta, \end{aligned}$$

in $\mathbb{R}^3$, where $0 \leq \theta < \pi$, $0 \leq \varphi < 2\pi$ and $r > 0$ (recall that φ is the so-called *azimuthal angle* in the xy-plane originating at the x-axis and θ is the so-called *polar angle* from the z-axis, angle defined in the plane obtained by rotating the xz-plane around the z-axis by the angle φ), then the Laplacian in spherical coordinates is given by

$$\Delta f = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^2} f,$$

where

$$\Delta_{S^2} f = \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2 f}{\partial \varphi^2},$$

is the Laplacian on the sphere, S^2 (for example, see Lebedev [97], Chapter 8 or Section 16.3, where we derive this formula). Let us look for homogeneous harmonic functions, $f(r,\theta,\varphi) = r^k g(\theta,\varphi)$, on $\mathbb{R}^3$, that is, solutions of the Laplace equation

$$\Delta f = 0.$$

We get

$$\begin{aligned}
\Delta f &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial (r^k g)}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^2}(r^k g)\\
&= \frac{1}{r^2}\frac{\partial}{\partial r}\left(k r^{k+1} g\right) + r^{k-2}\Delta_{S^2} g\\
&= r^{k-2}k(k+1)g + r^{k-2}\Delta_{S^2} g\\
&= r^{k-2}(k(k+1)g + \Delta_{S^2} g).
\end{aligned}$$

Therefore,

$$\Delta f = 0 \quad \text{iff} \quad \Delta_{S^2} g = -k(k+1)g,$$

that is, g is an eigenfunction of Δ_{S^2} for the eigenvalue $-k(k+1)$.

We can look for solutions of the above equation using the separation of variables method. If we let $g(\theta,\varphi) = \Theta(\theta)\Phi(\varphi)$, then we get the equation

$$\frac{\Phi}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Theta}{\partial\theta}\right) + \frac{\Theta}{\sin^2\theta}\frac{\partial^2\Phi}{\partial\varphi^2} = -k(k+1)\Theta\Phi,$$

that is, dividing by $\Theta\Phi$ and multiplying by $\sin^2\theta$,

$$\frac{\sin\theta}{\Theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Theta}{\partial\theta}\right) + k(k+1)\sin^2\theta = -\frac{1}{\Phi}\frac{\partial^2\Phi}{\partial\varphi^2}.$$

Since Θ and Φ are independent functions, the above is possible only if both sides are equal to a constant, say μ. This leads to two equations

$$\begin{aligned}
&\frac{\partial^2\Phi}{\partial\varphi^2} + \mu\Phi = 0\\
&\frac{\sin\theta}{\Theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Theta}{\partial\theta}\right) + k(k+1)\sin^2\theta - \mu = 0.
\end{aligned}$$

However, we want Φ to be a periodic in φ since we are considering functions on the sphere, so μ be must of the form $\mu = m^2$, for some non-negative integer, m. Then, we know that the space of solutions of the equation

$$\frac{\partial^2\Phi}{\partial\varphi^2} + m^2\Phi = 0$$

is two-dimensional and is spanned by the two functions

$$\Phi(\varphi) = \cos m\varphi, \qquad \Phi(\varphi) = \sin m\varphi.$$

We still have to solve the equation

$$\sin\theta \frac{\partial}{\partial\theta}\left(\sin\theta \frac{\partial\Theta}{\partial\theta}\right) + (k(k+1)\sin^2\theta - m^2)\Theta = 0,$$

which is equivalent to

$$\sin^2\theta\,\Theta'' + \sin\theta\cos\theta\,\Theta' + (k(k+1)\sin^2\theta - m^2)\Theta = 0.$$

a variant of Legendre's equation. For this, we use the change of variable, $t = \cos\theta$, and we consider the function, u, given by $u(\cos\theta) = \Theta(\theta)$ (recall that $0 \leq \theta < \pi$), so we get the second-order differential equation

$$(1-t^2)u'' - 2tu' + \left(k(k+1) - \frac{m^2}{1-t^2}\right)u = 0$$

sometimes called the *general Legendre equation* (Adrien-Marie Legendre, 1752-1833). The trick to solve this equation is to make the substitution

$$u(t) = (1-t^2)^{\frac{m}{2}} v(t),$$

see Lebedev [97], Chapter 7, Section 7.12. Then, we get

$$(1-t^2)v'' - 2(m+1)tv' + (k(k+1) - m(m+1))v = 0.$$

When $m = 0$, we get the *Legendre equation*:

$$(1-t^2)v'' - 2tv' + k(k+1)v = 0,$$

see Lebedev [97], Chapter 7, Section 7.3. This equation has two fundamental solution, $P_k(t)$ and $Q_k(t)$, called the *Legendre functions of the first and second kinds*. The $P_k(t)$ are actually polynomials and the $Q_k(t)$ are given by power series that diverge for $t = 1$, so we only keep the *Legendre polynomials*, $P_k(t)$. The Legendre polynomials can be defined in various ways. One definition is in terms of *Rodrigues' formula*:

$$P_n(t) = \frac{1}{2^n n!}\frac{d^n}{dt^n}(t^2-1)^n,$$

see Lebedev [97], Chapter 4, Section 4.2. In this version of the Legendre polynomials they are normalized so that $P_n(1) = 1$. There is also the following recurrence relation:

$$\begin{aligned} P_0 &= 1 \\ P_1 &= t \\ (n+1)P_{n+1} &= (2n+1)tP_n - nP_{n-1} \qquad n \geq 1, \end{aligned}$$

see Lebedev [97], Chapter 4, Section 4.3. For example, the first six Legendre polynomials are:

$$\begin{aligned}
&1\\
&t\\
&\frac{1}{2}(3t^2-1)\\
&\frac{1}{2}(5t^3-3t)\\
&\frac{1}{8}(35t^4-30t^2+3)\\
&\frac{1}{8}(63t^5-70t^3+15t).
\end{aligned}$$

Let us now return to our differential equation

$$(1-t^2)v''-2(m+1)tv'+(k(k+1)-m(m+1))v=0. \qquad (*)$$

Observe that if we differentiate with respect to t, we get the equation

$$(1-t^2)v'''-2(m+2)tv''+(k(k+1)-(m+1)(m+2))v'=0.$$

This shows that if v is a solution of our equation $(*)$ for given k and m, then v' is a solution of the same equation for k and $m+1$. Thus, if $P_k(t)$ solves $(*)$ for given k and $m=0$, then $P_k'(t)$ solves $(*)$ for the same k and $m=1$, $P_k''(t)$ solves $(*)$ for the same k and $m=2$, and in general, $d^m/dt^m(P_k(t))$ solves $(*)$ for k and m. Therefore, our original equation,

$$(1-t^2)u''-2tu'+\left(k(k+1)-\frac{m^2}{1-t^2}\right)u=0 \qquad (\dagger)$$

has the solution

$$u(t)=(1-t^2)^{\frac{m}{2}}\frac{d^m}{dt^m}(P_k(t)).$$

The function $u(t)$ is traditionally denoted $P_k^m(t)$ and called an *associated Legendre function*, see Lebedev [97], Chapter 7, Section 7.12. The index k is often called the *band index*. Obviously, $P_k^m(t)\equiv 0$ if $m>k$ and $P_k^0(t)=P_k(t)$, the Legendre polynomial of degree k. An associated Legendre function is not a polynomial in general and because of the factor $(1-t^2)^{\frac{m}{2}}$ it is only defined on the closed interval $[-1,1]$.

Certain authors add the factor $(-1)^m$ in front of the expression for the associated Legendre function $P_k^m(t)$, as in Lebedev [97], Chapter 7, Section 7.12, see also footnote 29 on page 193. This seems to be common practice in the quantum mechanics literature where it is called the *Condon Shortley phase factor*.

The associated Legendre functions satisfy various recurrence relations that allows us to compute them. For example, for fixed $m \geq 0$, we have (see Lebedev [97], Chapter 7, Section 7.12) the recurrence

$$(k-m+1)P_{k+1}^m(t) = (2k+1)tP_k^m(t) - (k+m)P_{k-1}^m(t), \qquad k \geq 1$$

and for fixed $k \geq 2$ we have

$$P_k^{m+2}(t) = \frac{2(m+1)t}{(t^2-1)^{\frac{1}{2}}}P_k^{m+1}(t) + (k-m)(k+m+1)P_k^m(t), \qquad 0 \leq m \leq k-2$$

which can also be used to compute P_k^m starting from

$$\begin{aligned} P_k^0(t) &= P_k(t) \\ P_k^1(t) &= \frac{kt}{(t^2-1)^{\frac{1}{2}}}P_k(t) - \frac{k}{(t^2-1)^{\frac{1}{2}}}P_{k-1}(t). \end{aligned}$$

Observe that the recurrence relation for m fixed yields the following equation for $k = m$ (as $P_{m-1}^m = 0$):

$$P_{m+1}^m(t) = (2m+1)tP_m^m(t).$$

It it also easy to see that

$$P_m^m(t) = \frac{(2m)!}{2^m m!}(1-t^2)^{\frac{m}{2}}.$$

Observe that

$$\frac{(2m)!}{2^m m!} = (2m-1)(2m-3)\cdots 5 \cdot 3 \cdot 1,$$

an expression that is sometimes denoted $(2m-1)!!$ and called the *double factorial*.

Beware that some papers in computer graphics adopt the definition of associated Legendre functions with the scale factor $(-1)^m$ added so this factor is present in these papers, for example, Green [64].

The equation above allows us to "lift" P_m^m to the higher band $m+1$. The computer graphics community (see Green [64]) uses the following three rules to compute $P_k^m(t)$ where $0 \leq m \leq k$:

(1) Compute

$$P_m^m(t) = \frac{(2m)!}{2^m m!}(1-t^2)^{\frac{m}{2}}.$$

If $m = k$, stop. Otherwise do step 2 once:

(2) Compute $P_{m+1}^m(t) = (2m+1)tP_m^m(t)$. If $k = m+1$, stop. Otherwise, iterate step 3:

(3) Starting from $i = m + 1$, compute

$$(i - m + 1)P_{i+1}^m(t) = (2i + 1)tP_i^m(t) - (i + m)P_{i-1}^m(t)$$

until $i + 1 = k$.

If we recall that equation (†) was obtained from the equation

$$\sin^2\theta\,\Theta'' + \sin\theta\cos\theta\,\Theta' + (k(k+1)\sin^2\theta - m^2)\Theta = 0$$

using the substitution $u(\cos\theta) = \Theta(\theta)$, we see that

$$\Theta(\theta) = P_k^m(\cos\theta)$$

is a solution of the above equation. Putting everything together, as $f(r,\theta,\varphi) = r^k\Theta(\theta)\Phi(\varphi)$, we proved that the homogeneous functions,

$$f(r,\theta,\varphi) = r^k\cos m\varphi\, P_k^m(\cos\theta), \qquad f(r,\theta,\varphi) = r^k\sin m\varphi\, P_k^m(\cos\theta),$$

are solutions of the Laplacian, Δ, in $\mathbb{R}^3$, and that the functions

$$\cos m\varphi\, P_k^m(\cos\theta), \qquad \sin m\varphi\, P_k^m(\cos\theta),$$

are eigenfunctions of the Laplacian, Δ_{S^2}, on the sphere for the eigenvalue $-k(k+1)$. For k fixed, as $0 \leq m \leq k$, we get $2k + 1$ linearly independent functions.

The notation for the above functions varies quite a bit essentially because of the choice of normalization factors used in various fields (such as physics, seismology, geodesy, spectral analysis, magnetics, quantum mechanics *etc.*). We will adopt the notation y_l^m, where l is a nonnegative integer but m is allowed to be negative, with $-l \leq m \leq l$. Thus, we set

$$y_l^m(\theta,\varphi) = \begin{cases} N_l^0 P_l(\cos\theta) & \text{if } m = 0 \\ \sqrt{2}N_l^m\cos m\varphi\, P_l^m(\cos\theta) & \text{if } m > 0 \\ \sqrt{2}N_l^m\sin(-m\varphi)\, P_l^{-m}(\cos\theta) & \text{if } m < 0 \end{cases}$$

for $l = 0, 1, 2, \ldots$, and where the N_l^m are scaling factors. In physics and computer graphics, N_l^m is chosen to be

$$N_l^m = \sqrt{\frac{(2l+1)(l-|m|)!}{4\pi(l+|m|)!}}.$$

The functions y_l^m are called the *real spherical harmonics of degree l and order m*. The index l is called the *band index*.

The functions, y_l^m, have some very nice properties but to explain these we need to recall the Hilbert space structure of the space, $L^2(S^2)$, of square-integrable functions on the sphere. Recall that we have an inner product on $L^2(S^2)$ given by

$$\langle f, g\rangle = \int_{S^2} fg\,\Omega_2 = \int_0^{2\pi}\int_0^{\pi} f(\theta,\varphi)g(\theta,\varphi)\sin\theta d\theta d\varphi,$$

where $f, g \in L^2(S^2)$ and where Ω_2 is the volume form on S^2 (induced by the metric on $\mathbb{R}^3$). With this inner product, $L^2(S^2)$ is a complete normed vector space using the norm, $\|f\| = \sqrt{\langle f, f\rangle}$, associated with this inner product, that is, $L^2(S^2)$ is a *Hilbert space*. Now, it can be shown that the Laplacian, Δ_{S^2}, on the sphere is a self-adjoint linear operator with respect to this inner product. As the functions, $y_{l_1}^{m_1}$ and $y_{l_2}^{m_2}$ with $l_1 \neq l_2$ are eigenfunctions corresponding to distinct eigenvalues ($-l_1(l_1+1)$ and $-l_2(l_2+1)$), they are orthogonal, that is,

$$\langle y_{l_1}^{m_1}, y_{l_2}^{m_2}\rangle = 0, \qquad \text{if} \quad l_1 \neq l_2.$$

It is also not hard to show that for a fixed l,

$$\langle y_l^{m_1}, y_l^{m_2}\rangle = \delta_{m_1,m_2},$$

that is, the functions y_l^m with $-l \leq m \leq l$ form an orthonormal system and we denote by $\mathcal{H}_l(S^2)$ the $(2l+1)$-dimensional space spanned by these functions. It turns out that the functions y_l^m form a basis of the eigenspace, E_l, of Δ_{S^2} associated with the eigenvalue $-l(l+1)$ so that $E_l = \mathcal{H}_l(S^2)$ and that Δ_{S^2} has no other eigenvalues. More is true. It turns out that $L^2(S^2)$ is the orthogonal Hilbert sum of the eigenspaces, $\mathcal{H}_l(S^2)$. This means that the $\mathcal{H}_l(S^2)$ are

(1) mutually orthogonal

(2) closed, and

(3) The space $L^2(S^2)$ is the Hilbert sum, $\bigoplus_{l=0}^{\infty} \mathcal{H}_l(S^2)$, which means that for every function, $f \in L^2(S^2)$, there is a unique sequence of spherical harmonics, $f_j \in \mathcal{H}_l(S^2)$, so that

$$f = \sum_{l=0}^{\infty} f_l,$$

that is, the sequence $\sum_{j=0}^{l} f_j$, converges to f (in the norm on $L^2(S^2)$). Observe that each f_l is a unique linear combination, $f_l = \sum_{m_l} a_{m_l\, l}\, y_l^{m_l}$.

Therefore, (3) gives us a *Fourier decomposition on the sphere* generalizing the familiar Fourier decomposition on the circle. Furthermore, the *Fourier coefficients*, $a_{m_l l}$, can be computed using the fact that the y_l^m form an orthonormal Hilbert basis:

$$a_{m_l\, l} = \langle f, y_l^{m_l}\rangle.$$

We also have the corresponding homogeneous harmonic functions, $H_l^m(r,\theta,\varphi)$, on $\mathbb{R}^3$ given by

$$H_l^m(r,\theta,\varphi) = r^l y_l^m(\theta,\varphi).$$

If one starts computing explicity the H_l^m for small values of l and m, one finds that it is always possible to express these functions in terms of the cartesian coordinates x, y, z as

homogeneous polynomials! This remarkable fact holds in general: The eigenfunctions of the Laplacian, Δ_{S^2}, and thus, the spherical harmonics, are the restrictions of homogeneous harmonic polynomials in $\mathbb{R}^3$. Here is a list of bases of the homogeneous harmonic polynomials of degree k in three variables up to $k = 4$ (thanks to Herman Gluck):

$$
\begin{array}{ll}
k = 0 & 1 \\
k = 1 & x,\ y,\ z \\
k = 2 & x^2 - y^2,\ x^2 - z^2,\ xy,\ xz,\ yz \\
k = 3 & x^3 - 3xy^2,\ 3x^2y - y^3,\ x^3 - 3xz^2,\ 3x^2z - z^3, \\
 & y^3 - 3yz^2,\ 3y^2z - z^3,\ xyz \\
k = 4 & x^4 - 6x^2y^2 + y^4,\ x^4 - 6x^2z^2 + z^4,\ y^4 - 6y^2z^2 + z^4, \\
 & x^3y - xy^3,\ x^3z - xz^3,\ y^3z - yz^3, \\
 & 3x^2yz - yz^3,\ 3xy^2z - xz^3,\ 3xyz^2 - x^3y.
\end{array}
$$

Subsequent sections will be devoted to a proof of the important facts stated earlier.

16.3 The Laplace-Beltrami Operator

In order to define rigorously the Laplacian on the sphere, $S^n \subseteq \mathbb{R}^{n+1}$, and establish its relationship with the Laplacian on $\mathbb{R}^{n+1}$, we need the definition of the Laplacian on a Riemannian manifold, (M, g), the *Laplace-Beltrami operator*, as defined in Section 15.2 (Eugenio Beltrami, 1835-1900). In that section, the Laplace-Beltrami operator is defined as an operator on differential forms but a more direct definition can be given for the Laplacian-Beltrami operator on functions (using the covariant derivative, see the paragraph preceding Proposition 15.6). For the benefit of the reader who may not have read Section 15.2, we present this definition of the divergence again.

Recall that a Riemannian metric, g, on a manifold, M, is a smooth family of inner products, $g = (g_p)$, where g_p is an inner product on the tangent space, T_pM, for every $p \in M$. The inner product, g_p, on T_pM, establishes a canonical duality between T_pM and T_p^*M, namely, we have the isomorphism, $\flat \colon T_pM \to T_p^*M$, defined such that for every $u \in T_pM$, the linear form, $u^\flat \in T_p^*M$, is given by

$$u^\flat(v) = g_p(u, v), \qquad v \in T_pM.$$

The inverse isomorphism, $\sharp \colon T_p^*M \to T_pM$, is defined such that for every $\omega \in T_p^*M$, the vector, $\omega^\sharp$, is the unique vector in T_pM so that

$$g_p(\omega^\sharp, v) = \omega(v), \qquad v \in T_pM.$$

The isomorphisms $\flat$ and $\sharp$ induce isomorphisms between vector fields, $X \in \mathfrak{X}(M)$, and one-forms, $\omega \in \mathcal{A}^1(M)$. In particular, for every smooth function, $f \in C^\infty(M)$, the vector field

corresponding to the one-form, df, is the *gradient*, $\operatorname{grad} f$, of f. The gradient of f is uniquely determined by the condition

$$g_p((\operatorname{grad} f)_p, v) = df_p(v), \qquad v \in T_pM,\ p \in M.$$

If ∇_X is the covariant derivative associated with the Levi-Civita connection induced by the metric, g, then the *divergence* of a vector field, $X \in \mathfrak{X}(M)$, is the function, $\operatorname{div} X \colon M \to \mathbb{R}$, defined so that

$$(\operatorname{div} X)(p) = \operatorname{tr}(Y(p) \mapsto (\nabla_Y X)_p),$$

namely, for every p, $(\operatorname{div} X)(p)$ is the trace of the linear map, $Y(p) \mapsto (\nabla_Y X)_p$. Then, the *Laplace-Beltrami operator*, for short, *Laplacian*, is the linear operator,
$\Delta \colon C^\infty(M) \to C^\infty(M)$, given by

$$\Delta f = \operatorname{div} \operatorname{grad} f.$$

Observe that the definition just given differs from the definition given in Section 15.2 by a negative sign. We adopted this sign convention to conform with most of the literature on spherical harmonics (where the negative sign is omitted). A consequence of this choice is that the eigenvalues of the Laplacian are negative.

For more details on the Laplace-Beltrami operator, we refer the reader to Chapter 15 or to Gallot, Hulin and Lafontaine [60] (Chapter 4) or O'Neill [119] (Chapter 3), Postnikov [125] (Chapter 13), Helgason [72] (Chapter 2) or Warner [147] (Chapters 4 and 6).

All this being rather abstact, it is useful to know how $\operatorname{grad} f$, $\operatorname{div} X$ and Δf are expressed in a chart. If (U, φ) is a chart of M, with $p \in M$ and if, as usual,

$$\left(\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p\right)$$

denotes the basis of T_pM induced by φ, the local expression of the metric g at p is given by the $n \times n$ matrix, $(g_{ij})_p$, with

$$(g_{ij})_p = g_p\left(\left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p\right).$$

The matrix $(g_{ij})_p$ is symmetric, positive definite and its inverse is denoted $(g^{ij})_p$. We also let $|g|_p = \det(g_{ij})_p$. For simplicity of notation we often omit the subscript p. Then, it can be shown that for every function, $f \in C^\infty(M)$, in local coordinates given by the chart (U, φ), we have

$$\operatorname{grad} f = \sum_{ij} g^{ij} \frac{\partial f}{\partial x_j} \frac{\partial}{\partial x_i},$$

where, as usual

$$\frac{\partial f}{\partial x_j}(p) = \left(\frac{\partial}{\partial x_j}\right)_p f = \frac{\partial (f \circ \varphi^{-1})}{\partial u_j}(\varphi(p))$$

and $(u_1, \ldots, u_n)$ are the coordinate functions in $\mathbb{R}^n$. There are formulae for $\operatorname{div} X$ and Δf involving the Christoffel symbols but the following formulae will be more convenient for our purposes: For every vector field, $X \in \mathfrak{X}(M)$, expressed in local coordinates as

$$X = \sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i}$$

we have

$$\operatorname{div} X = \frac{1}{\sqrt{|g|}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, X_i \right)$$

and for every function, $f \in C^\infty(M)$, the Laplacian, Δf, is given by

$$\Delta f = \frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, g^{ij} \frac{\partial f}{\partial x_j} \right).$$

The above formula is proved in Proposition 15.5, assuming M is orientable. A different derivation is given in Postnikov [125] (Chapter 13, Section 5).

One should check that for $M = \mathbb{R}^n$ with its standard coordinates, the Laplacian is given by the familiar formula

$$\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2}.$$

Remark: A different sign convention is also used in defining the divergence, namely,

$$\operatorname{div} X = -\frac{1}{\sqrt{|g|}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, X_i \right).$$

With this convention, which is the one used in Section 15.2, the Laplacian also has a negative sign. This has the advantage that the eigenvalues of the Laplacian are nonnegative.

As an application, let us derive the formula for the Laplacian in spherical coordinates,

$$\begin{aligned} x &= r \sin\theta \cos\varphi \\ y &= r \sin\theta \sin\varphi \\ z &= r \cos\theta. \end{aligned}$$

We have

$$\begin{aligned} \frac{\partial}{\partial r} &= \sin\theta\cos\varphi \frac{\partial}{\partial x} + \sin\theta\sin\varphi \frac{\partial}{\partial y} + \cos\theta \frac{\partial}{\partial z} = \widehat{r} \\ \frac{\partial}{\partial \theta} &= r\left(\cos\theta\cos\varphi \frac{\partial}{\partial x} + \cos\theta\sin\varphi \frac{\partial}{\partial y} - \sin\theta \frac{\partial}{\partial z} \right) = r\widehat{\theta} \\ \frac{\partial}{\partial \varphi} &= r\left(-\sin\theta\sin\varphi \frac{\partial}{\partial x} + \sin\theta\cos\varphi \frac{\partial}{\partial y} \right) = r\widehat{\varphi}. \end{aligned}$$

Observe that $\widehat{r}$, $\widehat{\theta}$ and $\widehat{\varphi}$ are pairwise orthogonal. Therefore, the matrix (g_{ij}) is given by

$$(g_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix}$$

and $|g| = r^4 \sin^2\theta$. The inverse of (g_{ij}) is

$$(g^{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^{-2} & 0 \\ 0 & 0 & r^{-2} \sin^{-2}\theta \end{pmatrix}.$$

We will let the reader finish the computation to verify that we get

$$\Delta f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial f}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 f}{\partial \varphi^2}.$$

Since (θ, φ) are coordinates on the sphere S^2 *via*

$$\begin{aligned} x &= \sin\theta\cos\varphi \\ y &= \sin\theta\sin\varphi \\ z &= \cos\theta, \end{aligned}$$

we see that in these coordinates, the metric, $(\widetilde{g}_{ij})$, on S^2 is given by the matrix

$$(\widetilde{g}_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix},$$

that $|\widetilde{g}| = \sin^2\theta$, and that the inverse of $(\widetilde{g}_{ij})$ is

$$(\widetilde{g}^{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & \sin^{-2}\theta \end{pmatrix}.$$

It follows immediately that

$$\Delta_{S^2} f = \frac{1}{\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial f}{\partial \theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 f}{\partial \varphi^2},$$

so we have verified that

$$\Delta f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^2} f.$$

Let us now generalize the above formula to the Laplacian, Δ, on $\mathbb{R}^{n+1}$ and the Laplacian, Δ_{S^n}, on S^n, where

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\}.$$

Following Morimoto [113] (Chapter 2, Section 2), let us use "polar coordinates". The map from $\mathbb{R}_+ \times S^n$ to $\mathbb{R}^{n+1} - \{0\}$ given by

$$(r, \sigma) \mapsto r\sigma$$

is clearly a diffeomorphism. Thus, for any system of coordinates, $(u_1, \ldots, u_n)$, on S^n, the tuple $(u_1, \ldots, u_n, r)$ is a system of coordinates on $\mathbb{R}^{n+1} - \{0\}$ called *polar coordinates*. Let us establish the relationship between the Laplacian, Δ, on $\mathbb{R}^{n+1} - \{0\}$ in polar coordinates and the Laplacian, Δ_{S^n}, on S^n in local coordinates $(u_1, \ldots, u_n)$.

Proposition 16.1 *If Δ is the Laplacian on $\mathbb{R}^{n+1} - \{0\}$ in polar coordinates $(u_1, \ldots, u_n, r)$ and Δ_{S^n} is the Laplacian on the sphere, S^n, in local coordinates $(u_1, \ldots, u_n)$, then*

$$\Delta f = \frac{1}{r^n}\frac{\partial}{\partial r}\left(r^n \frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^n} f.$$

Proof. Let us compute the $(n+1)\times(n+1)$ matrix, $G = (g_{ij})$, expressing the metric on $\mathbb{R}^{n+1}$ is polar coordinates and the $n \times n$ matrix, $\widetilde{G} = (\widetilde{g}_{ij})$, expressing the metric on S^n. Recall that if $\sigma \in S^n$, then $\sigma \cdot \sigma = 1$ and so,

$$\frac{\partial \sigma}{\partial u_i} \cdot \sigma = 0,$$

as

$$\frac{\partial \sigma}{\partial u_i} \cdot \sigma = \frac{1}{2}\frac{\partial(\sigma \cdot \sigma)}{\partial u_i} = 0.$$

If $x = r\sigma$ with $\sigma \in S^n$, we have

$$\frac{\partial x}{\partial u_i} = r\frac{\partial \sigma}{\partial u_i}, \qquad 1 \leq i \leq n,$$

and

$$\frac{\partial x}{\partial r} = \sigma.$$

It follows that

$$\begin{aligned} g_{ij} &= \frac{\partial x}{\partial u_i} \cdot \frac{\partial x}{\partial u_j} = r^2 \frac{\partial \sigma}{\partial u_i} \cdot \frac{\partial \sigma}{\partial u_j} = r^2 \widetilde{g}_{ij} \\ g_{in+1} &= \frac{\partial x}{\partial u_i} \cdot \frac{\partial x}{\partial r} = r\frac{\partial \sigma}{\partial u_i} \cdot \sigma = 0 \\ g_{n+1n+1} &= \frac{\partial x}{\partial r} \cdot \frac{\partial x}{\partial r} = \sigma \cdot \sigma = 1. \end{aligned}$$

Consequently, we get

$$G = \begin{pmatrix} r^2\widetilde{G} & 0 \\ 0 & 1 \end{pmatrix},$$

$|g| = r^{2n}|\widetilde{g}|$ and

$$G^{-1} = \begin{pmatrix} r^{-2}\widetilde{G}^{-1} & 0 \\ 0 & 1 \end{pmatrix}.$$

Using the above equations and

$$\Delta f = \frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, g^{ij} \frac{\partial f}{\partial x_j} \right),$$

we get

$$\begin{aligned} \Delta f &= \frac{1}{r^n \sqrt{|\widetilde{g}|}} \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left(r^n \sqrt{|\widetilde{g}|}\, \frac{1}{r^2} \widetilde{g}^{ij} \frac{\partial f}{\partial x_j} \right) + \frac{1}{r^n \sqrt{|\widetilde{g}|}} \frac{\partial}{\partial r} \left(r^n \sqrt{|\widetilde{g}|}\, \frac{\partial f}{\partial r} \right) \\ &= \frac{1}{r^2 \sqrt{|\widetilde{g}|}} \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left(\sqrt{|\widetilde{g}|}\, \widetilde{g}^{ij} \frac{\partial f}{\partial x_j} \right) + \frac{1}{r^n} \frac{\partial}{\partial r} \left(r^n \frac{\partial f}{\partial r} \right) \\ &= \frac{1}{r^2} \Delta_{S^n} f + \frac{1}{r^n} \frac{\partial}{\partial r} \left(r^n \frac{\partial f}{\partial r} \right), \end{aligned}$$

as claimed. □

It is also possible to express Δ_{S^n} in terms of $\Delta_{S^{n-1}}$. If $e_{n+1} = (0, \ldots, 0, 1) \in \mathbb{R}^{n+1}$, then we can view S^{n-1} as the intersection of S^n with the hyperplane, $x_{n+1} = 0$, that is, as the set

$$S^{n-1} = \{\sigma \in S^n \mid \sigma \cdot e_{n+1} = 0\}.$$

If $(u_1, \ldots, u_{n-1})$ are local coordinates on S^{n-1}, then $(u_1, \ldots, u_{n-1}, \theta)$ are local coordinates on S^n, by setting

$$\sigma = \sin\theta\, \widetilde{\sigma} + \cos\theta\, e_{n+1},$$

with $\widetilde{\sigma} \in S^{n-1}$ and $0 \leq \theta < \pi$. Using these local coordinate systems, it is a good exercise to find the relationship between Δ_{S^n} and $\Delta_{S^{n-1}}$, namely

$$\Delta_{S^n} f = \frac{1}{\sin^{n-1}\theta} \frac{\partial}{\partial \theta} \left(\sin^{n-1}\theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{\sin^2\theta} \Delta_{S_{n-1}} f.$$

A fundamental property of the divergence is known as *Green's Formula*. There are actually two Greens' Formulae but we will only need the version for an orientable manifold without boundary given in Proposition 15.7. Recall that Green's Formula states that if M is a compact, orientable, Riemannian manifold without boundary, then, for every smooth vector field, $X \in \mathfrak{X}(M)$, we have

$$\int_M (\operatorname{div} X)\, \Omega_M = 0,$$

where Ω_M is the volume form on M induced by the metric.

If M is a compact, orientable Riemannian manifold, then for any two smooth functions, $f, h \in C^\infty(M)$, we define $\langle f, h\rangle$ by

$$\langle f, h\rangle = \int_M fh\,\Omega_M.$$

Then, it is not hard to show that $\langle -, -\rangle$ is an inner product on $\mathcal{C}^\infty(M)$.

An important property of the Laplacian on a compact, orientable Riemannian manifold is that it is a self-adjoint operator. This fact has already been proved in the more general case of an inner product on differential forms in Proposition 15.3 but it might be instructive to give another proof in the special case of functions using Green's Formula.

For this, we prove the following properties: For any two functions, $f, h \in C^\infty(M)$, and any vector field, $X \in \mathcal{C}^\infty(M)$, we have:

$$\begin{aligned} \operatorname{div}(fX) &= f\operatorname{div} X + X(f) = f\operatorname{div} X + g(\operatorname{grad} f, X) \\ \operatorname{grad} f\,(h) &= g(\operatorname{grad} f, \operatorname{grad} h) = \operatorname{grad} h\,(f). \end{aligned}$$

Using these identities, we obtain the following important special case of Proposition 15.3:

Proposition 16.2 *Let M be a compact, orientable, Riemannian manifold without boundary. The Laplacian on M is self-adjoint, that is, for any two functions, $f, h \in C^\infty(M)$, we have*

$$\langle \Delta f, h\rangle = \langle f, \Delta h\rangle$$

or equivalently

$$\int_M f\Delta h\,\Omega_M = \int_M h\Delta f\,\Omega_M.$$

Proof. By the two identities before Proposition 16.2,

$$f\Delta h = f\operatorname{div}\operatorname{grad} h = \operatorname{div}(f\operatorname{grad} h) - g(\operatorname{grad} f, \operatorname{grad} h)$$

and

$$h\Delta f = h\operatorname{div}\operatorname{grad} f = \operatorname{div}(h\operatorname{grad} f) - g(\operatorname{grad} h, \operatorname{grad} f),$$

so we get

$$f\Delta h - h\Delta f = \operatorname{div}(f\operatorname{grad} h - h\operatorname{grad} f).$$

By Green's Formula,

$$\int_M (f\Delta h - h\Delta f)\Omega_M = \int_M \operatorname{div}(f\operatorname{grad} h - h\operatorname{grad} f)\Omega_M = 0,$$

which proves that Δ is self-adjoint. □

The importance of Proposition 16.2 lies in the fact that as $\langle -, -\rangle$ is an inner product on $\mathcal{C}^\infty(M)$, the eigenspaces of Δ for distinct eigenvalues are pairwise orthogonal. We will make heavy use of this property in the next section on harmonic polynomials.

16.4 Harmonic Polynomials, Spherical Harmonics and $L^2(S^n)$

Harmonic homogeneous polynomials and their restrictions to S^n, where

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\},$$

turn out to play a crucial role in understanding the structure of the eigenspaces of the Laplacian on S^n (with $n \geq 1$). The results in this section appear in one form or another in Stein and Weiss [142] (Chapter 4), Morimoto [113] (Chapter 2), Helgason [72] (Introduction, Section 3), Dieudonné [43] (Chapter 7), Axler, Bourdon and Ramey [12] (Chapter 5) and Vilenkin [146] (Chapter IX). Some of these sources assume a fair amount of mathematical background and consequently, uninitiated readers will probably find the exposition rather condensed, especially Helgason. We tried hard to make our presentation more "user-friendly".

Definition 16.1 Let $\mathcal{P}_k(n+1)$ (resp. $\mathcal{P}_k^{\mathbb{C}}(n+1)$) denote the space of homogeneous polynomials of degree k in $n+1$ variables with real coefficients (resp. complex coefficients) and let $\mathcal{P}_k(S^n)$ (resp. $\mathcal{P}_k^{\mathbb{C}}(S^n)$) denote the restrictions of homogeneous polynomials in $\mathcal{P}_k(n+1)$ to S^n (resp. the restrictions of homogeneous polynomials in $\mathcal{P}_k^{\mathbb{C}}(n+1)$ to S^n). Let $\mathcal{H}_k(n+1)$ (resp. $\mathcal{H}_k^{\mathbb{C}}(n+1)$) denote the space of *(real) harmonic polynomials* (resp. *complex harmonic polynomials*), with

$$\mathcal{H}_k(n+1) = \{P \in \mathcal{P}_k(n+1) \mid \Delta P = 0\}$$

and

$$\mathcal{H}_k^{\mathbb{C}}(n+1) = \{P \in \mathcal{P}_k^{\mathbb{C}}(n+1) \mid \Delta P = 0\}.$$

Harmonic polynomials are sometimes called *solid harmonics*. Finally, Let $\mathcal{H}_k(S^n)$ (resp. $\mathcal{H}_k^{\mathbb{C}}(S^n)$) denote the space of *(real) spherical harmonics* (resp. *complex spherical harmonics*) be the set of restrictions of harmonic polynomials in $\mathcal{H}_k(n+1)$ to S^n (resp. restrictions of harmonic polynomials in $\mathcal{H}_k^{\mathbb{C}}(n+1)$ to S^n).

A function, $f\colon \mathbb{R}^n \to \mathbb{R}$ (resp. $f\colon \mathbb{R}^n \to \mathbb{C}$), is *homogeneous of degree k* iff

$$f(tx) = t^k f(x), \qquad \text{for all } x \in \mathbb{R}^n \text{ and } t > 0.$$

The restriction map, $\rho\colon \mathcal{H}_k(n+1) \to \mathcal{H}_k(S^n)$, is a surjective linear map. In fact, it is a bijection. Indeed, if $P \in \mathcal{H}_k(n+1)$, observe that

$$P(x) = \|x\|^k P\left(\frac{x}{\|x\|}\right), \qquad \text{with} \quad \frac{x}{\|x\|} \in S^n,$$

for all $x \neq 0$. Consequently, if $P \restriction S^n = Q \restriction S^n$, that is, $P(\sigma) = Q(\sigma)$ for all $\sigma \in S^n$, then

$$P(x) = \|x\|^k P\left(\frac{x}{\|x\|}\right) = \|x\|^k Q\left(\frac{x}{\|x\|}\right) = Q(x)$$

for all $x \neq 0$, which implies $P = Q$ (as P and Q are polynomials). Therefore, we have a linear isomorphism between $\mathcal{H}_k(n+1)$ and $\mathcal{H}_k(S^n)$ (and between $\mathcal{H}_k^{\mathbb{C}}(n+1)$ and $\mathcal{H}_k^{\mathbb{C}}(S^n)$).

It will be convenient to introduce some notation to deal with homogeneous polynomials. Given $n \geq 1$ variables, $x_1, \ldots, x_n$, and any n-tuple of nonnegative integers, $\alpha = (\alpha_1, \ldots, \alpha_n)$, let $|\alpha| = \alpha_1 + \cdots + \alpha_n$, let $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and let $\alpha! = \alpha_1! \cdots \alpha_n!$. Then, every homogeneous polynomial, P, of degree k in the variables $x_1, \ldots, x_n$ can be written uniquely as

$$P = \sum_{|\alpha|=k} c_\alpha x^\alpha,$$

with $c_\alpha \in \mathbb{R}$ or $c_\alpha \in \mathbb{C}$. It is well known that $\mathcal{P}_k(n)$ is a (real) vector space of dimension

$$d_k = \binom{n+k-1}{k}$$

and $\mathcal{P}_k^{\mathbb{C}}(n)$ is a complex vector space of the same dimension, d_k.

We can define an Hermitian inner product on $\mathcal{P}_k^{\mathbb{C}}(n)$ whose restriction to $\mathcal{P}_k(n)$ is an inner product by viewing a homogeneous polynomial as a differential operator as follows: For every $P = \sum_{|\alpha|=k} c_\alpha x^\alpha \in \mathcal{P}_k^{\mathbb{C}}(n)$, let

$$\partial(P) = \sum_{|\alpha|=k} c_\alpha \frac{\partial^k}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Then, for any two polynomials, $P, Q \in \mathcal{P}_k^{\mathbb{C}}(n)$, let

$$\langle P, Q \rangle = \partial(P)\overline{Q}.$$

A simple computation shows that

$$\left\langle \sum_{|\alpha|=k} a_\alpha x^\alpha, \sum_{|\alpha|=k} b_\alpha x^\alpha \right\rangle = \sum_{|\alpha|=k} \alpha!\, a_\alpha \overline{b}_\alpha.$$

Therefore, $\langle P, Q \rangle$ is indeed an inner product. Also observe that

$$\partial(x_1^2 + \cdots + x_n^2) = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2} = \Delta.$$

Another useful property of our inner product is this:

$$\langle P, QR \rangle = \langle \partial(Q)P, R \rangle.$$

Indeed.

$$\begin{aligned}\langle P, QR\rangle &= \langle QR, P\rangle \\ &= \partial(QR)\overline{P} \\ &= \partial(Q)(\partial(R)\overline{P}) \\ &= \partial(R)(\partial(Q)\overline{P}) \\ &= \langle R, \partial(Q)P\rangle \\ &= \langle \partial(Q)P, R\rangle.\end{aligned}$$

In particular,

$$\langle (x_1^2 + \cdots + x_n^2)P, Q\rangle = \langle P, \partial(x_1^2 + \cdots + x_n^2)Q\rangle = \langle P, \Delta Q\rangle.$$

Let us write $\|x\|^2$ for $x_1^2 + \cdots + x_n^2$. Using our inner product, we can prove the following important theorem:

Theorem 16.3 *The map, $\Delta\colon \mathcal{P}_k(n) \to \mathcal{P}_{k-2}(n)$, is surjective for all $n, k \geq 2$ (and similarly for $\Delta\colon \mathcal{P}_k^{\mathbb{C}}(n) \to \mathcal{P}_{k-2}^{\mathbb{C}}(n)$). Furthermore, we have the following orthogonal direct sum decompositions:*

$$\mathcal{P}_k(n) = \mathcal{H}_k(n) \oplus \|x\|^2 \mathcal{H}_{k-2}(n) \oplus \cdots \oplus \|x\|^{2j} \mathcal{H}_{k-2j}(n) \oplus \cdots \oplus \|x\|^{2[k/2]} \mathcal{H}_{[k/2]}(n)$$

and

$$\mathcal{P}_k^{\mathbb{C}}(n) = \mathcal{H}_k^{\mathbb{C}}(n) \oplus \|x\|^2 \mathcal{H}_{k-2}^{\mathbb{C}}(n) \oplus \cdots \oplus \|x\|^{2j} \mathcal{H}_{k-2j}^{\mathbb{C}}(n) \oplus \cdots \oplus \|x\|^{2[k/2]} \mathcal{H}_{[k/2]}^{\mathbb{C}}(n),$$

with the understanding that only the first term occurs on the right-hand side when $k < 2$.

Proof. If the map $\Delta\colon \mathcal{P}_k^{\mathbb{C}}(n) \to \mathcal{P}_{k-2}^{\mathbb{C}}(n)$ is not surjective, then some nonzero polynomial, $Q \in \mathcal{P}_{k-2}^{\mathbb{C}}(n)$, is orthogonal to the image of Δ. In particular, Q must be orthogonal to ΔP with $P = \|x\|^2 Q \in \mathcal{P}_k^{\mathbb{C}}(n)$. So, using a fact established earlier,

$$0 = \langle Q, \Delta P\rangle = \langle \|x\|^2 Q, P\rangle = \langle P, P\rangle,$$

which implies that $P = \|x\|^2 Q = 0$ and thus, $Q = 0$, a contradiction. The same proof is valid in the real case.

We claim that we have an orthogonal direct sum decomposition,

$$\mathcal{P}_k^{\mathbb{C}}(n) = \mathcal{H}_k^{\mathbb{C}}(n) \oplus \|x\|^2 \mathcal{P}_{k-2}^{\mathbb{C}}(n),$$

and similarly in the real case, with the understanding that the second term is missing if $k < 2$. If $k = 0, 1$, then $\mathcal{P}_k^{\mathbb{C}}(n) = \mathcal{H}_k^{\mathbb{C}}(n)$ so this case is trivial. Assume $k \geq 2$. Since $\operatorname{Ker}\Delta = \mathcal{H}_k^{\mathbb{C}}(n)$ and Δ is surjective, $\dim(\mathcal{P}_k^{\mathbb{C}}(n)) = \dim(\mathcal{H}_k^{\mathbb{C}}(n)) + \dim(\mathcal{P}_{k-2}^{\mathbb{C}}(n))$, so it is

sufficient to prove that $\mathcal{H}_k^{\mathbb{C}}(n)$ is orthogonal to $\|x\|^2 \mathcal{P}_{k-2}^{\mathbb{C}}(n)$. Now, if $H \in \mathcal{H}_k^{\mathbb{C}}(n)$ and $P = \|x\|^2 Q \in \|x\|^2 \mathcal{P}_{k-2}^{\mathbb{C}}(n)$, we have

$$\langle \|x\|^2 Q, H\rangle = \langle Q, \Delta H\rangle = 0,$$

so $\mathcal{H}_k^{\mathbb{C}}(n)$ and $\|x\|^2 \mathcal{P}_{k-2}^{\mathbb{C}}(n)$ are indeed orthogonal. Using induction, we immediately get the orthogonal direct sum decomposition

$$\mathcal{P}_k^{\mathbb{C}}(n) = \mathcal{H}_k^{\mathbb{C}}(n) \oplus \|x\|^2 \mathcal{H}_{k-2}^{\mathbb{C}}(n) \oplus \cdots \oplus \|x\|^{2j} \mathcal{H}_{k-2j}^{\mathbb{C}}(n) \oplus \cdots \oplus \|x\|^{2[k/2]} \mathcal{H}_{[k/2]}^{\mathbb{C}}(n)$$

and the corresponding real version. □

Remark: Theorem 16.3 also holds for $n = 1$.

Theorem 16.3 has some important corollaries. Since every polynomial in $n+1$ variables is the sum of homogeneous polynomials, we get:

Corollary 16.4 *The restriction to S^n of every polynomial (resp. complex polynomial) in $n+1 \geq 2$ variables is a sum of restrictions to S^n of harmonic polynomials (resp. complex harmonic polynomials).*

We can also derive a formula for the dimension of $\mathcal{H}_k(n)$ (and $\mathcal{H}_k^{\mathbb{C}}(n)$).

Corollary 16.5 *The dimension, $a_{k,n}$, of the space of harmonic polynomials, $\mathcal{H}_k(n)$, is given by the formula*

$$a_{k,n} = \binom{n+k-1}{k} - \binom{n+k-3}{k-2}$$

if $n, k \geq 2$, with $a_{0,n} = 1$ and $a_{1,n} = n$, and similarly for $\mathcal{H}_k^{\mathbb{C}}(n)$. As $\mathcal{H}_k(n+1)$ is isomorphic to $\mathcal{H}_k(S^n)$ (and $\mathcal{H}_k^{\mathbb{C}}(n+1)$ is isomorphic to $\mathcal{H}_k^{\mathbb{C}}(S^n)$) we have

$$\dim(\mathcal{H}_k^{\mathbb{C}}(S^n)) = \dim(\mathcal{H}_k(S^n)) = a_{k,n+1} = \binom{n+k}{k} - \binom{n+k-2}{k-2}.$$

Proof. The cases $k = 0$ and $k = 1$ are trivial since in this case $\mathcal{H}_k(n) = \mathcal{P}_k(n)$. For $k \geq 2$, the result follows from the direct sum decomposition

$$\mathcal{P}_k(n) = \mathcal{H}_k(n) \oplus \|x\|^2 \mathcal{P}_{k-2}(n)$$

proved earlier. The proof is identical in the complex case. □

Observe that when $n = 2$, we get $a_{k,2} = 2$ for $k \geq 1$ and when $n = 3$, we get $a_{k,3} = 2k+1$ for all $k \geq 0$, which we already knew from Section 16.2. The formula even applies for $n = 1$ and yields $a_{k,1} = 0$ for $k \geq 2$.

Remark: It is easy to show that

$$a_{k,n+1} = \binom{n+k-1}{n-1} + \binom{n+k-2}{n-1}$$

for $k \geq 2$, see Morimoto [113] (Chapter 2, Theorem 2.4) or Dieudonné [43] (Chapter 7, formula 99), where a different proof technique is used.

Let $L^2(S^n)$ be the space of (real) square-integrable functions on the sphere, S^n. We have an inner product on $L^2(S^n)$ given by

$$\langle f, g\rangle = \int_{S^n} f g \, \Omega_n,$$

where $f, g \in L^2(S^n)$ and where Ω_n is the volume form on S^n (induced by the metric on $\mathbb{R}^{n+1}$). With this inner product, $L^2(S^n)$ is a complete normed vector space using the norm, $\|f\| = \|f\|_2 = \sqrt{\langle f, f\rangle}$, associated with this inner product, that is, $L^2(S^n)$ is a *Hilbert space*. In the case of complex-valued functions, we use the Hermitian inner product

$$\langle f, g\rangle = \int_{S^n} f \, \overline{g} \, \Omega_n$$

and we get the complex Hilbert space, $L^2_{\mathbb{C}}(S^n)$. We also denote by $C(S^n)$ the space of continuous (real) functions on S^n with the L^∞ norm, that is,

$$\|f\|_\infty = \sup\{|f(x)|\}_{x \in S^n}$$

and by $C_{\mathbb{C}}(S^n)$ the space of continuous complex-valued functions on S^n also with the L^∞ norm. Recall that $C(S^n)$ is dense in $L^2(S^n)$ (and $C_{\mathbb{C}}(S^n)$ is dense in $L^2_{\mathbb{C}}(S^n)$). The following proposition shows why the spherical harmonics play an important role:

Proposition 16.6 *The set of all finite linear combinations of elements in $\bigcup_{k=0}^\infty \mathcal{H}_k(S^n)$ (resp. $\bigcup_{k=0}^\infty \mathcal{H}_k^{\mathbb{C}}(S^n)$) is*

(i) dense in $C(S^n)$ (resp. in $C_{\mathbb{C}}(S^n)$) with respect to the L^∞-norm;

(ii) dense in $L^2(S^n)$ (resp. dense in $L^2_{\mathbb{C}}(S^n)$).

Proof. (i) As S^n is compact, by the Stone-Weierstrass approximation theorem (Lang [93], Chapter III, Corollary 1.3), if g is continuous on S^n, then it can be approximated uniformly by polynomials, P_j, restricted to S^n. By Corollary 16.4, the restriction of each P_j to S^n is a linear combination of elements in $\bigcup_{k=0}^\infty \mathcal{H}_k(S^n)$.

(ii) We use the fact that $C(S^n)$ is dense in $L^2(S^n)$. Given $f \in L^2(S^n)$, for every $\epsilon > 0$, we can choose a continuous function, g, so that $\|f - g\|_2 < \epsilon/2$. By (i), we can find a linear

combination, h, of elements in $\bigcup_{k=0}^{\infty} \mathcal{H}_k(S^n)$ so that $\|g - h\|_\infty < \epsilon/(2\sqrt{\mathrm{vol}(S^n)})$, where $\mathrm{vol}(S^n)$ is the volume of S^n (really, area). Thus, we get

$$\|f - h\|_2 \leq \|f - g\|_2 + \|g - h\|_2 < \epsilon/2 + \sqrt{\mathrm{vol}(S^n)}\,\|g - h\|_\infty < \epsilon/2 + \epsilon/2 = \epsilon,$$

which proves (ii). The proof in the complex case is identical. □

We need one more proposition before showing that the spaces $\mathcal{H}_k(S^n)$ constitute an orthogonal Hilbert space decomposition of $L^2(S^n)$.

Proposition 16.7 *For every harmonic polynomial, $P \in \mathcal{H}_k(n+1)$ (resp. $P \in \mathcal{H}_k^{\mathbb{C}}(n+1)$), the restriction, $H \in \mathcal{H}_k(S^n)$ (resp. $H \in \mathcal{H}_k^{\mathbb{C}}(S^n)$), of P to S^n is an eigenfunction of Δ_{S^n} for the eigenvalue $-k(n+k-1)$.*

Proof. We have

$$P(r\sigma) = r^k H(\sigma), \qquad r > 0,\ \sigma \in S^n,$$

and by Proposition 16.1, for any $f \in C^\infty(\mathbb{R}^{n+1})$, we have

$$\Delta f = \frac{1}{r^n}\frac{\partial}{\partial r}\left(r^n \frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^n} f.$$

Consequently,

$$\begin{aligned}
\Delta P = \Delta(r^k H) &= \frac{1}{r^n}\frac{\partial}{\partial r}\left(r^n \frac{\partial (r^k H)}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^n}(r^k H) \\
&= \frac{1}{r^n}\frac{\partial}{\partial r}\left(k r^{n+k-1} H\right) + r^{k-2}\Delta_{S^n} H \\
&= \frac{1}{r^n} k(n+k-1) r^{n+k-2} H + r^{k-2}\Delta_{S^n} H \\
&= r^{k-2}(k(n+k-1)H + \Delta_{S^n} H).
\end{aligned}$$

Thus,

$$\Delta P = 0 \quad \text{iff} \quad \Delta_{S^n} H = -k(n+k-1)H,$$

as claimed. □

From Proposition 16.7, we deduce that the space $\mathcal{H}_k(S^n)$ is a subspace of the eigenspace, E_k, of Δ_{S^n}, associated with the eigenvalue $-k(n+k-1)$ (and similarly for $\mathcal{H}_k^{\mathbb{C}}(S^n)$). Remarkably, $E_k = \mathcal{H}_k(S^n)$ but it will take more work to prove this.

What we can deduce immediately is that $\mathcal{H}_k(S^n)$ *and* $\mathcal{H}_l(S^n)$ *are pairwise orthogonal whenever* $k \neq l$. This is because, by Proposition 16.2, the Laplacian is self-adjoint and thus, any two eigenspaces, E_k and E_l are pairwise orthogonal whenever $k \neq l$ and as $\mathcal{H}_k(S^n) \subseteq E_k$ and $\mathcal{H}_l(S^n) \subseteq E_l$, our claim is indeed true. Furthermore, by Proposition 16.5, each $\mathcal{H}_k(S^n)$ is finite-dimensional and thus, closed. Finally, we know from Proposition 16.6 that $\bigcup_{k=0}^{\infty} \mathcal{H}_k(S^n)$ is dense in $L^2(S^n)$. But then, we can apply a standard result from Hilbert space theory (for example, see Lang [93], Chapter V, Proposition 1.9) to deduce the following important result:

Theorem 16.8 *The family of spaces, $\mathcal{H}_k(S^n)$ (resp. $\mathcal{H}_k^{\mathbb{C}}(S^n)$) yields a Hilbert space direct sum decomposition*

$$L^2(S^n) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^n) \qquad (resp. \quad L^2_{\mathbb{C}}(S^n) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k^{\mathbb{C}}(S^n)),$$

which means that the summands are closed, pairwise orthogonal, and that every $f \in L^2(S^n)$ (resp. $f \in L^2_{\mathbb{C}}(S^n)$) is the sum of a converging series

$$f = \sum_{k=0}^{\infty} f_k,$$

in the L^2-norm, where the $f_k \in \mathcal{H}_k(S^n)$ (resp. $f_k \in \mathcal{H}_k^{\mathbb{C}}(S^n)$) are uniquely determined functions. Furthermore, given any orthonormal basis, $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$, of $\mathcal{H}_k(S^n)$, we have

$$f_k = \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_k^{m_k}, \qquad with \quad c_{k,m_k} = \langle f, Y_k^{m_k} \rangle.$$

The coefficients c_{k,m_k} are "generalized" *Fourier coefficients* with respect to the Hilbert basis $\{Y_k^{m_k} \mid 1 \leq m_k \leq a_{k,n+1}, k \geq 0\}$. We can finally prove the main theorem of this section.

Theorem 16.9

(1) The eigenspaces (resp. complex eigenspaces) of the Laplacian, Δ_{S^n}, on S^n are the spaces of spherical harmonics,

$$E_k = \mathcal{H}_k(S^n) \qquad (resp. \quad E_k = \mathcal{H}_k^{\mathbb{C}}(S^n))$$

and E_k corresponds to the eigenvalue $-k(n+k-1)$.

(2) We have the Hilbert space direct sum decompositions

$$L^2(S^n) = \bigoplus_{k=0}^{\infty} E_k \qquad (resp. \quad L^2_{\mathbb{C}}(S^n) = \bigoplus_{k=0}^{\infty} E_k).$$

(3) The complex polynomials of the form $(c_1 x_1 + \cdots + c_{n+1} x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$, span the space $\mathcal{H}_k^{\mathbb{C}}(n+1)$, for $k \geq 1$.

Proof. We follow essentially the proof in Helgason [72] (Introduction, Theorem 3.1). In (1) and (2) we only deal with the real case, the proof in the complex case being identical.

(1) We already know that the integers $-k(n+k-1)$ are eigenvalues of Δ_{S^n} and that $\mathcal{H}_k(S^n) \subseteq E_k$. We will prove that Δ_{S^n} has no other eigenvalues and no other eigenvectors

using the Hilbert basis, $\{Y_k^{m_k} \mid 1 \leq m_k \leq a_{k,n+1},\ k \geq 0\}$, given by Theorem 16.8. Let λ be any eigenvalue of Δ_{S^n} and let $f \in L^2(S^n)$ be any eigenfunction associated with λ so that

$$\Delta f = \lambda f.$$

We have a unique series expansion

$$f = \sum_{k=0}^{\infty} \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_k^{m_k},$$

with $c_{k,m_k} = \langle f, Y_k^{m_k} \rangle$. Now, as Δ_{S^n} is self-adjoint and $\Delta Y_k^{m_k} = -k(n+k-1)Y_k^{m_k}$, the Fourier coefficients, d_{k,m_k}, of Δf are given by

$$d_{k,m_k} = \langle \Delta f, Y_k^{m_k} \rangle = \langle f, \Delta Y_k^{m_k} \rangle = -k(n+k-1)\langle f, Y_k^{m_k} \rangle = -k(n+k-1)c_{k,m_k}.$$

On the other hand, as $\Delta f = \lambda f$, the Fourier coefficients of Δf are given by

$$d_{k,m_k} = \lambda c_{k,m_k}.$$

By uniqueness of the Fourier expansion, we must have

$$\lambda c_{k,m_k} = -k(n+k-1)c_{k,m_k} \qquad \text{for all } k \geq 0.$$

Since $f \neq 0$, there some k such that $c_{k,m_k} \neq 0$ and we must have

$$\lambda = -k(n+k-1)$$

for any such k. However, the function $k \mapsto -k(n+k-1)$ reaches its maximum for $k = -\frac{n-1}{2}$ and as $n \geq 1$, it is strictly decreasing for $k \geq 0$, which implies that k is unique and that

$$c_{j,m_j} = 0 \qquad \text{for all } j \neq k.$$

Therefore, $f \in \mathcal{H}_k(S^n)$ and the eigenvalues of Δ_{S^n} are exactly the integers $-k(n+k-1)$ so $E_k = \mathcal{H}_k(S^n)$, as claimed.

Since we just proved that $E_k = \mathcal{H}_k(S^n)$, (2) follows immediately from the Hilbert decomposition given by Theorem 16.8.

(3) If $H = (c_1 x_1 + \cdots + c_{n+1}x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$, then for $k \leq 1$ is is obvious that $\Delta H = 0$ and for $k \geq 2$ we have

$$\Delta H = k(k-1)(c_1^2 + \cdots + c_{n+1}^2)(c_1 x_1 + \cdots + c_{n+1}x_{n+1})^{k-2} = 0,$$

so $H \in \mathcal{H}_k^{\mathbb{C}}(n+1)$. A simple computation shows that for every $Q \in \mathcal{P}_k^{\mathbb{C}}(n+1)$, if $c = (c_1, \ldots, c_{n+1})$, then we have

$$\partial(Q)(c_1 x_1 + \cdots + c_{n+1}x_{n+1})^m = m(m-1)\cdots(m-k+1)Q(c)(c_1 x_1 + \cdots + c_{n+1}x_{n+1})^{m-k},$$

for all $m \geq k \geq 1$.

Assume that $\mathcal{H}_k^{\mathbb{C}}(n+1)$ is not spanned by the complex polynomials of the form $(c_1x_1 + \cdots + c_{n+1}x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$, for $k \geq 1$. Then, some $Q \in \mathcal{H}_k^{\mathbb{C}}(n+1)$ is orthogonal to all polynomials of the form $H = (c_1x_1 + \cdots + c_{n+1}x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$. Recall that

$$\langle P, \partial(Q)H\rangle = \langle QP, H\rangle$$

and apply this equation to $P = Q(c)$, H and Q. Since

$$\partial(Q)H = \partial(Q)(c_1x_1 + \cdots + c_{n+1}x_{n+1})^k = k!Q(c),$$

as Q is orthogonal to H, we get

$$k!\langle Q(c), Q(c)\rangle = \langle Q(c), k!Q(c)\rangle = \langle Q(c), \partial(Q)H\rangle = \langle Q\,Q(c), H\rangle = Q(c)\langle Q, H\rangle = 0,$$

which implies $Q(c) = 0$. Consequently, $Q(x_1, \ldots, x_{n+1})$ vanishes on the complex algebraic variety,

$$\{(x_1, \ldots, x_{n+1}) \in \mathbb{C}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 0\}.$$

By the Hilbert *Nullstellensatz*, some power, Q^m, belongs to the ideal, $(x_1^2 + \cdots + x_{n+1}^2)$, generated by $x_1^2 + \cdots + x_{n+1}^2$. Now, if $n \geq 2$, it is well-known that the polynomial $x_1^2 + \cdots + x_{n+1}^2$ is irreducible so the ideal $(x_1^2 + \cdots + x_{n+1}^2)$ is a prime ideal and thus, Q is divisible by $x_1^2 + \cdots + x_{n+1}^2$. However, we know from the proof of Theorem 16.3 that we have an orthogonal direct sum

$$\mathcal{P}_k^{\mathbb{C}}(n+1) = \mathcal{H}_k^{\mathbb{C}}(n+1) \oplus \|x\|^2\,\mathcal{P}_{k-2}^{\mathbb{C}}(n+1).$$

Since $Q \in \mathcal{H}_k^{\mathbb{C}}(n+1)$ and Q is divisible by $x_1^2 + \cdots + x_{n+1}^2$, we must have $Q = 0$. Therefore, if $n \geq 2$, we proved (3). However, when $n = 1$, we know from Section 16.1 that the complex harmonic homogeneous polynomials in two variables, $P(x, y)$, are spanned by the real and imaginary parts, U_k, V_k of the polynomial $(x + iy)^k = U_k + iV_k$. Since $(x - iy)^k = U_k - iV_k$ we see that

$$U_k = \frac{1}{2}\left((x + iy)^k + (x - iy)^k\right), \qquad V_k = \frac{1}{2i}\left((x + iy)^k - (x - iy)^k\right),$$

and as $1 + i^2 = 1 + (-i)^2 = 0$, the space $\mathcal{H}_k^{\mathbb{C}}(\mathbb{R}^2)$ is spanned by $(x + iy)^k$ and $(x - iy)^k$ (for $k \geq 1$), so (3) holds for $n = 1$ as well. □

As an illustration of part (3) of Theorem 16.9, the polynomials $(x_1 + i\cos\theta x_2 + i\sin\theta x_3)^k$ are harmonic. Of course, the real and imaginary part of a complex harmonic polynomial $(c_1x_1 + \cdots + c_{n+1}x_{n+1})^k$ are real harmonic polynomials.

In the next section, we try to show how spherical harmonics fit into the broader framework of linear respresentations of (Lie) groups.

16.5 Spherical Functions and Linear Representations of Lie Groups; A Glimpse

In this section, we indicate briefly how Theorem 16.9 (except part (3)) can be viewed as a special case of a famous theorem known as the *Peter-Weyl Theorem* about unitary representations of compact Lie groups (Herman, Klauss, Hugo Weyl, 1885-1955). First, we review the notion of a linear representation of a group. A good and easy-going introduction to representations of Lie groups can be found in Hall [70]. We begin with finite-dimensional representations.

Definition 16.2 Given a Lie group, G, and a vector space, V, of dimension n, a *linear representation* of G of *dimension* (or *degree* n) is a group homomorphism, $U\colon G \to \mathbf{GL}(V)$, such that the map, $g \mapsto U(g)(u)$, is continuous for every $u \in V$ and where $\mathbf{GL}(V)$ denotes the group of invertible linear maps from V to itself. The space, V, called the *representation space* may be a real or a complex vector space. If V has a Hermitian (resp Euclidean) inner product, $\langle -, - \rangle$, we say that $U\colon G \to \mathbf{GL}(V)$ is a *unitary representation* iff

$$\langle U(g)(u), U(g)(v)\rangle = \langle u, v\rangle, \qquad \text{for all } g \in G \text{ and all } u, v \in V.$$

Thus, a linear representation of G is a map, $U\colon G \to \mathbf{GL}(V)$, satisfying the properties:

$$\begin{aligned} U(gh) &= U(g)U(h) \\ U(g^{-1}) &= U(g)^{-1} \\ U(1) &= I. \end{aligned}$$

For simplicity of language, we usually abbreviate *linear representation* as *representation*. The representation space, V, is also called a *G-module* since the representation, $U\colon G \to \mathbf{GL}(V)$, is equivalent to the left action, $\cdot\colon G \times V \to V$, with $g \cdot v = U(g)(v)$. The representation such that $U(g) = I$ for all $g \in G$ is called the *trivial representation*.

As an example, we describe a class of representations of $\mathbf{SL}(2, \mathbb{C})$, the group of complex matrices with determinant $+1$,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad ad - bc = 1.$$

Recall that $\mathcal{P}_k^{\mathbb{C}}(2)$ denotes the vector space of complex homogeneous polynomials of degree k in two variables, (z_1, z_2). For every matrix, $A \in \mathbf{SL}(2, \mathbb{C})$, with

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

for every homogeneous polynomial, $Q \in \mathcal{P}_k^{\mathbb{C}}(2)$, we define $U_k(A)(Q(z_1, z_2))$ by

$$U_k(A)(Q(z_1, z_2)) = Q(dz_1 - bz_2, -cz_1 + az_2).$$

If we think of the homogeneous polynomial, $Q(z_1, z_2)$, as a function, $Q\binom{z_1}{z_2}$, of the vector $\binom{z_1}{z_2}$, then

$$U_k(A)\left(Q\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right) = QA^{-1}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = Q\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

The expression above makes it clear that

$$U_k(AB) = U_k(A)U_k(B)$$

for any two matrices, $A, B \in \mathbf{SL}(2, \mathbb{C})$, so U_k is indeed a representation of $\mathbf{SL}(2, \mathbb{C})$ into $\mathcal{P}_k^{\mathbb{C}}(2)$. It can be shown that the representations, U_k, are irreducible and that every representation of $\mathbf{SL}(2, \mathbb{C})$ is equivalent to one of the U_k's (see Bröcker and tom Dieck [25], Chapter 2, Section 5). The representations, U_k, are also representations of $\mathbf{SU}(2)$. Again, they are irreducible representations of $\mathbf{SU}(2)$ and they constitute all of them (up to equivalence). The reader should consult Hall [70] for more examples of representations of Lie groups.

One might wonder why we considered $\mathbf{SL}(2, \mathbb{C})$ rather than $\mathbf{SL}(2, \mathbb{R})$. This is because it can be shown that $\mathbf{SL}(2, \mathbb{R})$ has *no* nontrivial unitary (finite-dimensional) representations! For more on representations of $\mathbf{SL}(2, \mathbb{R})$, see Dieudonné [43] (Chapter 14).

Given any basis, $(e_1, \ldots, e_n)$, of V, each $U(g)$ is represented by an $n \times n$ matrix, $U(g) = (U_{ij}(g))$. We may think of the scalar functions, $g \mapsto U_{ij}(g)$, as *special functions* on G. As explained in Dieudonné [43] (see also Vilenkin [146]), essentially all special functions (Legendre polynomials, ultraspherical polynomials, Bessel functions, *etc.*) arise in this way by choosing some suitable G and V. There is a natural and useful notion of equivalence of representations:

Definition 16.3 Given any two representations, $U_1 \colon G \to \mathbf{GL}(V_1)$ and $U_2 \colon G \to \mathbf{GL}(V_2)$, a *$G$-map* (or *morphism of representations*), $\varphi \colon U_1 \to U_2$, is a linear map, $\varphi \colon V_1 \to V_2$, so that the following diagram commutes for every $g \in G$:

$$\begin{array}{ccc} V_1 & \xrightarrow{U_1(g)} & V_1 \\ {\scriptstyle\varphi}\downarrow & & \downarrow{\scriptstyle\varphi} \\ V_2 & \xrightarrow{U_2(g)} & V_2. \end{array}$$

The space of all G-maps between two representations as above is denoted $\mathrm{Hom}_G(U_1, U_2)$. Two representations $U_1 \colon G \to \mathbf{GL}(V_1)$ and $U_2 \colon G \to \mathbf{GL}(V_2)$ are *equivalent* iff $\varphi \colon V_1 \to V_2$ is an invertible linear map (which implies that $\dim V_1 = \dim V_2$). In terms of matrices, the representations $U_1 \colon G \to \mathbf{GL}(V_1)$ and $U_2 \colon G \to \mathbf{GL}(V_2)$ are equivalent iff there is some invertible $n \times n$ matrix, P, so that

$$U_2(g) = PU_1(g)P^{-1}, \qquad g \in G.$$

If $W \subseteq V$ is a subspace of V, then in some cases, a representation $U\colon G \to \mathbf{GL}(V)$ yields a representation $U\colon G \to \mathbf{GL}(W)$. This is interesting because under certain conditions on G (*e.g.*, G compact) every representation may be decomposed into a "sum" of so-called irreducible representations and thus, the study of all representations of G boils down to the study of irreducible representations of G (for instance, see Knapp [89] (Chapter 4, Corollary 4.7) or Bröcker and tom Dieck [25] (Chapter 2, Proposition 1.9).

Definition 16.4 Let $U\colon G \to \mathbf{GL}(V)$ be a representation of G. If $W \subseteq V$ is a subspace of V, then we say that W is *invariant* (or *stable*) under U iff $U(g)(w) \in W$, for all $g \in G$ and all $w \in W$. If W is invariant under U, then we have a homomorphism, $U\colon G \to \mathbf{GL}(W)$, called a *subrepresentation* of G. A representation, $U\colon G \to \mathbf{GL}(V)$, with $V \neq (0)$ is *irreducible* iff it only has the two subrepresentations, $U\colon G \to \mathbf{GL}(W)$, corresponding to $W = (0)$ or $W = V$.

An easy but crucial lemma about irreducible representations is "Schur's Lemma".

Lemma 16.10 *(Schur's Lemma) Let $U_1\colon G \to \mathbf{GL}(V)$ and $U_2\colon G \to \mathbf{GL}(W)$ be any two real or complex representations of a group, G. If U_1 and U_2 are irreducible, then the following properties hold:*

(i) Every G-map, $\varphi\colon U_1 \to U_2$, is either the zero map or an isomorphism.

(ii) If U_1 is a complex representation, then every G-map, $\varphi\colon U_1 \to U_1$, is of the form, $\varphi = \lambda \mathrm{id}$, for some $\lambda \in \mathbb{C}$.

Proof. (i) Observe that the kernel, $\mathrm{Ker}\,\varphi \subseteq V$, of φ is invariant under U_1. Indeed, for every $v \in \mathrm{Ker}\,\varphi$ and every $g \in G$, we have

$$\varphi(U_1(g)(v)) = U_2(g)(\varphi(v)) = U_2(g)(0) = 0,$$

so $U_1(g)(v) \in \mathrm{Ker}\,\varphi$. Thus, $U_1\colon G \to \mathbf{GL}(\mathrm{Ker}\,\varphi)$ is a subrepresentation of U_1 and as U_1 is irreducible, either $\mathrm{Ker}\,\varphi = (0)$ or $\mathrm{Ker}\,\varphi = V$. In the second case, $\varphi = 0$. If $\mathrm{Ker}\,\varphi = (0)$, then φ is injective. However, $\varphi(V) \subseteq W$ is invariant under U_2 since for every $v \in V$ and every $g \in G$,

$$U_2(g)(\varphi(v)) = \varphi(U_1(g)(v)) \in \varphi(V),$$

and as $\varphi(V) \neq (0)$ (as $V \neq (0)$ since U_1 is irreducible) and U_2 is irreducible, we must have $\varphi(V) = W$, that is, φ is an isomorphism.

(ii) Since V is a complex vector space, the linear map, φ, has some eigenvalue, $\lambda \in \mathbb{C}$. Let $E_\lambda \subseteq V$ be the eigenspace associated with λ. The subspace E_λ is invariant under U_1 since for every $u \in E_\lambda$ and every $g \in G$, we have

$$\varphi(U_1(g)(u)) = U_1(g)(\varphi(u)) = U_1(g)(\lambda u) = \lambda U_1(g)(u),$$

so $U_1 \colon G \to \mathbf{GL}(E_\lambda)$ is a subrepresentation of U_1 and as U_1 is irreducible and $E_\lambda \neq (0)$, we must have $E_\lambda = V$. $\square$

An interesting corollary of Schur's Lemma is that every complex irreducible representtaion of a commutative group is one-dimensional.

Let us now restrict our attention to compact Lie groups. If G is a compact Lie group, then it is known that it has a left and right-invariant volume form, ω_G, so we can define the integral of a (real or complex) continuous function, f, defined on G by

$$\int_G f = \int_G f\,\omega_G,$$

also denoted $\int_G f\,d\mu_G$ or simply $\int_G f(t)\,dt$, with ω_G normalized so that $\int_G \omega_G = 1$. (See Section 9.4, or Knapp [89], Chapter 8, or Warner [147], Chapters 4 and 6.) Because G is compact, the *Haar measure*, μ_G, induced by ω_G is both left and right-invariant (G is a *unimodular group*) and our integral has the following invariance properties:

$$\int_G f(t)\,dt = \int_G f(st)\,dt = \int_G f(tu)\,dt = \int_G f(t^{-1})\,dt,$$

for all $s, u \in G$ (see Section 9.4).

Since G is a compact Lie group, we can use an "averaging trick" to show that every (finite-dimensional) representation is equivalent to a unitary representation (see Bröcker and tom Dieck [25] (Chapter 2, Theorem 1.7) or Knapp [89] (Chapter 4, Proposition 4.6).

If we define the Hermitian inner product,

$$\langle f, g\rangle = \int_G f\,\overline{g}\,\omega_G,$$

then, with this inner product, the space of square-integrable functions, $L^2_{\mathbb{C}}(G)$, is a *Hilbert space*. We can also define the *convolution*, $f * g$, of two functions, $f, g \in L^2_{\mathbb{C}}(G)$, by

$$(f * g)(x) = \int_G f(xt^{-1})g(t)dt = \int_G f(t)g(t^{-1}x)dt$$

In general, $f * g \neq g * f$ unless G is commutative. With the convolution product, $L^2_{\mathbb{C}}(G)$ becomes an associative algebra (non-commutative in general).

This leads us to consider unitary representations of G into the infinite-dimensional vector space, $L^2_{\mathbb{C}}(G)$. The definition is the same as in Definition 16.2, except that $\mathbf{GL}(L^2_{\mathbb{C}}(G))$ is the group of automorphisms (unitary operators), $\mathrm{Aut}(L^2_{\mathbb{C}}(G))$, of the Hilbert space, $L^2_{\mathbb{C}}(G)$ and

$$\langle U(g)(u), U(g)(v)\rangle = \langle u, v\rangle$$

with respect to the inner product on $L^2_{\mathbb{C}}(G)$. Also, in the definition of an irreducible representation, $U \colon G \to V$, we require that the only *closed* subrepresentations, $U \colon G \to W$, of the representation, $U \colon G \to V$, correspond to $W = (0)$ or $W = V$.

The *Peter Weyl Theorem* gives a decomposition of $L^2_{\mathbb{C}}(G)$ as a Hilbert sum of spaces that correspond to irreducible unitary representations of G. We present a version of the Peter Weyl Theorem found in Dieudonné [43] (Chapters 3-8) and Dieudonné [44] (Chapter XXI, Sections 1-4), which contains complete proofs. Other versions can be found in Bröcker and tom Dieck [25] (Chapter 3), Knapp [89] (Chapter 4) or Duistermaat and Kolk [53] (Chapter 4). A good preparation for these fairly advanced books is Deitmar [40].

Theorem 16.11 *(Peter-Weyl (1927)) Given a compact Lie group, G, there is a decomposition of $L^2_{\mathbb{C}}(G)$ as a Hilbert sum,*

$$L^2_{\mathbb{C}}(G) = \bigoplus_{\rho} \mathfrak{a}_\rho,$$

of countably many two-sided ideals, $\mathfrak{a}_\rho$, where each $\mathfrak{a}_\rho$ is isomorphic to a finite-dimensional algebra of $n_\rho \times n_\rho$ complex matrices. More precisely, there is a basis of $\mathfrak{a}_\rho$ consisting of smooth pairwise orthogonal functions, $m^{(\rho)}_{ij}$, satisfying various properties, including

$$\langle m^{(\rho)}_{ij}, m^{(\rho)}_{ij} \rangle = n_\rho,$$

and if we form the matrix, $M_\rho(g) = (\frac{1}{n_\rho} m^{(\rho)}_{ij}(g))$, then the map, $g \mapsto M_\rho(g)$ is an **irreducible unitary representation** *of G in the vector space $\mathbb{C}^{n_\rho}$. Furthermore, every irreducible representation of G is equivalent to some M_ρ, so the set of indices, ρ, corresponds to the set of equivalence classes of irreducible unitary representations of G. The function, u_ρ, given by*

$$u_\rho(g) = \sum_{j=1}^{n_\rho} m^{(\rho)}_{jj}(g) = n_\rho \mathrm{tr}(M_\rho(g))$$

is the unit of the algebra $\mathfrak{a}_\rho$ and the orthogonal projection of $L^2_{\mathbb{C}}(G)$ onto $\mathfrak{a}_\rho$ is the map

$$f \mapsto u_\rho * f,$$

that is, convolution with u_ρ.

Remark: The function, $\chi_\rho = \frac{1}{n_\rho} u_\rho = \mathrm{tr}(M_\rho)$, is the *character* of G associated with the representation of G into M_ρ. The functions, χ_ρ, form an orthogonal system. Beware that they are *not* homomorphisms of G into $\mathbb{C}$ unless G is commutative. The characters of G are the key to the definition of the Fourier transform on a (compact) group, G.

A complete proof of Theorem 16.11 is given in Dieudonné [44], Chapter XXI, Section 2, but see also Sections 3 and 4.

There is more to the Peter Weyl Theorem: It gives a description of all unitary representations of G into a separable Hilbert space (see Dieudonné [44], Chapter XXI, Section 4). If $V \colon G \to \mathrm{Aut}(E)$ is such a representation, then for every ρ as above, the map

$$x \mapsto V(u_\rho)(x) = \int_G (V(s)(x)) u_\rho(s)\, ds$$

is an orthogonal projection of E onto a closed subspace, E_ρ. Then, E is the Hilbert sum, $E = \bigoplus_\rho E_\rho$, of those E_ρ such that $E_\rho \neq (0)$ and each such E_ρ is itself a (countable) Hilbert sum of closed spaces invariant under V. The subrepresentations of V corresponding to these subspaces of E_ρ are all equivalent to $M_{\overline{\rho}} = \overline{M_\rho}$ and hence, irreducible. This is why *every* (unitary) representation of G is equivalent to some representation of the form M_ρ.

An interesting special case is the case of the so-called regular representation of G in $L^2_{\mathbb{C}}(G)$ itself. The (left) *regular representation*, $\mathbf{R}$, of G in $L^2_{\mathbb{C}}(G)$ is defined by

$$(\mathbf{R}_s(f))(t) = \lambda_s(f)(t) = f(s^{-1}t), \qquad f \in L^2_{\mathbb{C}}(G),\ s, t \in G.$$

It turns out that we also get the same Hilbert sum,

$$L^2_{\mathbb{C}}(G) = \bigoplus_\rho \mathfrak{a}_\rho,$$

but this time, the $\mathfrak{a}_\rho$ generally do not correspond to irreducible subrepresentations. However, $\mathfrak{a}_\rho$ splits into n_ρ left ideals, $\mathfrak{b}_j^{(\rho)}$, where $\mathfrak{b}_j^{(\rho)}$ corresponds to the jth columm of M_ρ and all the subrepresentations of G in $\mathfrak{b}_j^{(\rho)}$ are equivalent to $M_{\overline{\rho}}$ and thus, are irreducible (see Dieudonné [43], Chapter 3).

Finally, assume that besides the compact Lie group, G, we also have a closed subgroup, K, of G. Then, we know that $M = G/K$ is a manifold called a *homogeneous space* and G acts on M on the left. For example, if $G = \mathbf{SO}(n+1)$ and $K = \mathbf{SO}(n)$, then $S^n = \mathbf{SO}(n+1)/\mathbf{SO}(n)$ (for instance, see Warner [147], Chapter 3). The subspace of $L^2_{\mathbb{C}}(G)$ consisting of the functions $f \in L^2_{\mathbb{C}}(G)$ that are right-invariant under the action of K, that is, such that

$$f(su) = f(s) \qquad \text{for all } s \in G \text{ and all } u \in K$$

form a closed subspace of $L^2_{\mathbb{C}}(G)$ denoted $L^2_{\mathbb{C}}(G/K)$. For example, if $G = \mathbf{SO}(n+1)$ and $K = \mathbf{SO}(n)$, then $L^2_{\mathbb{C}}(G/K) = L^2_{\mathbb{C}}(S^n)$.

It turns out that $L^2_{\mathbb{C}}(G/K)$ is invariant under the regular representation, $\mathbf{R}$, of G in $L^2_{\mathbb{C}}(G)$, so we get a subrepresentation (of the regular representation) of G in $L^2_{\mathbb{C}}(G/K)$. Again, the Peter-Weyl gives us a Hilbert sum decomposition of $L^2_{\mathbb{C}}(G/K)$ of the form

$$L^2_{\mathbb{C}}(G/K) = \bigoplus_\rho L_\rho = L^2_{\mathbb{C}}(G/K) \cap \mathfrak{a}_\rho,$$

for the same ρ's as before. However, these subrepresentations of $\mathbf{R}$ in L_ρ are not necessarily irreducible. What happens is that there is some d_ρ with $0 \leq d_\rho \leq n_\rho$ so that if $d_\rho \geq 1$, then L_σ is the direct sum of the first d_ρ columns of M_ρ (see Dieudonné [43], Chapter 6 and Dieudonné [45], Chapter XXII, Sections 4-5).

We can also consider the subspace of $L^2_{\mathbb{C}}(G)$ consisting of the functions, $f \in L^2_{\mathbb{C}}(G)$, that are left-invariant under the action of K, that is, such that

$$f(ts) = f(s) \qquad \text{for all } s \in G \text{ and all } t \in K.$$

This is a closed subspace of $L^2_{\mathbb{C}}(G)$ denoted $L^2_{\mathbb{C}}(K\backslash G)$. Then, we get a Hilbert sum decomposition of $L^2_{\mathbb{C}}(K\backslash G)$ of the form

$$L^2_{\mathbb{C}}(K\backslash G) = \bigoplus_{\rho} L'_\rho = L^2_{\mathbb{C}}(K\backslash G) \cap \mathfrak{a}_\rho,$$

and for the same d_ρ as before, L'_σ is the direct sum of the first d_ρ rows of M_ρ. We can also consider

$$\begin{aligned} L^2_{\mathbb{C}}(K\backslash G/K) &= L^2_{\mathbb{C}}(G/K) \cap L^2_{\mathbb{C}}(K\backslash G) \\ &= \{f \in L^2_{\mathbb{C}}(G) \mid f(tsu) = f(s)\} \qquad \text{for all } s \in G \text{ and all } t, u \in K. \end{aligned}$$

From our previous discussion, we see that we have a Hilbert sum decomposition

$$L^2_{\mathbb{C}}(K\backslash G/K) = \bigoplus_{\rho} L_\rho \cap L'_\rho$$

and each $L_\rho \cap L'_\rho$ for which $d_\rho \geq 1$ is a matrix algebra of dimension d_ρ^2. As a consequence, *the algebra $L^2_{\mathbb{C}}(K\backslash G/K)$ is commutative iff $d_\rho \leq 1$ for all ρ.*

If the algebra $L^2_{\mathbb{C}}(K\backslash G/K)$ is commutative (for the convolution product), we say that (G, K) is a *Gelfand pair* (see Dieudonné [43], Chapter 8 and Dieudonné [45], Chapter XXII, Sections 6-7). In this case, the L_ρ in the Hilbert sum decomposition of $L^2_{\mathbb{C}}(G/K)$ are nontrivial of dimension n_ρ iff $d_\rho = 1$ and the subrepresentation, $\mathbf{U}$, (of the regular representation) of G into L_ρ is irreducible and equivalent to $M_{\overline{\rho}}$. The space L_ρ is generated by the functions, $m^{(\rho)}_{1,1}, \ldots, m^{(\rho)}_{n_\rho,1}$, but the function

$$\omega_\rho(s) = \frac{1}{n_\rho}\, m^{(\rho)}_{1,1}(s)$$

plays a special role. This function called a *zonal spherical function* has some interesting properties. First, $\omega_\rho(e) = 1$ (where e is the identity element of the group, G) and

$$\omega_\rho(ust) = \omega_\rho(s) \qquad \text{for all } s \in G \text{ and all } u, t \in K.$$

In addition, ω_ρ is of positive type. A function, $f\colon G \to \mathbb{C}$, is of *positive type* iff

$$\sum_{j,k=1}^{n} f(s_j^{-1}s_k) z_j \overline{z}_k \geq 0,$$

for every finite set, $\{s_1, \ldots, s_n\}$, of elements of G and every finite tuple, $(z_1, \ldots, z_n) \in \mathbb{C}^n$. Because the subrepresentation of G into L_ρ is irreducible, the function ω_ρ generates L_ρ under left translation. This means the following: If we recall that for any function, f, on G,

$$\lambda_s(f)(t) = f(s^{-1}t), \qquad s, t \in G,$$

then, L_ρ is generated by the functions $\lambda_s(\omega_\rho)$, as s varies in G. The function ω_ρ also satisfies the following property:

$$\omega_\rho(s) = \langle \mathbf{U}(s)(\omega_\rho), \omega_\rho \rangle.$$

The set of zonal spherical functions on G/K is denoted $S(G/K)$. It is a countable set.

The notion of Gelfand pair also applies to locally-compact unimodular groups that are not necessary compact but we will not discuss this notion here. Curious readers may consult Dieudonné [43] (Chapters 8 and 9) and Dieudonné [45] (Chapter XXII, Sections 6-9).

It turns out that $G = \mathbf{SO}(n+1)$ and $K = \mathbf{SO}(n)$ form a Gelfand pair (see Dieudonné [43], Chapters 7-8 and Dieudonné [46], Chapter XXIII, Section 38). In this particular case, $\rho = k$ is any nonnegative integer and $L_\rho = E_k$, the eigenspace of the Laplacian on S^n corresponding to the eigenvalue $-k(n+k-1)$. Therefore, the regular representation of $\mathbf{SO}(n)$ into $E_k = \mathcal{H}_k^{\mathbb{C}}(S^n)$ is irreducible. This can be proved more directly, for example, see Helgason [72] (Introduction, Theorem 3.1) or Bröcker and tom Dieck [25] (Chapter 2, Proposition 5.10).

The zonal spherical harmonics, ω_k, can be expressed in terms of the *ultraspherical polynomials* (also called *Gegenbauer polynomials*), $P_k^{(n-1)/2}$ (up to a constant factor), see Stein and Weiss [142] (Chapter 4), Morimoto [113] (Chapter 2) and Dieudonné [43] (Chapter 7). For $n = 2$, $P_k^{\frac{1}{2}}$ is just the ordinary Legendre polynomial (up to a constant factor). We will say more about the zonal spherical harmonics and the ultraspherical polynomials in the next two sections.

The material in this section belongs to the overlapping areas of *representation theory* and *noncommutative harmonic analysis*. These are deep and vast areas. Besides the references cited earlier, for noncommutative harmonic analysis, the reader may consult Folland [54] or Taylor [144], but they may find the pace rather rapid. Another great survey on both topics is Kirillov [87], although it is not geared for the beginner.

16.6 Reproducing Kernel, Zonal Spherical Functions and Gegenbauer Polynomials

We now return to S^n and its spherical harmonics. The previous section suggested that zonal spherical functions play a special role. In this section, we describe the zonal spherical functions on S^n and show that they essentially come from certain polynomials generalizing the Legendre polyomials known as the *Gegenbauer Polynomials*. Most proof will be omitted. We refer the reader to Stein and Weiss [142] (Chapter 4) and Morimoto [113] (Chapter 2) for a complete exposition with proofs.

Recall that the space of spherical harmonics, $\mathcal{H}_k^{\mathbb{C}}(S^n)$, is the image of the space of homogeneous harmonic poynomials, $\mathcal{P}_k^{\mathbb{C}}(n+1)$, under the restriction map. It is a finite-dimensional

space of dimension

$$a_{k,n+1} = \binom{n+k}{k} - \binom{n+k-2}{k-2},$$

if $n \geq 1$ and $k \geq 2$, with $a_{0,n+1} = 1$ and $a_{1,n+1} = n+1$. Let $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ be any orthonormal basis of $\mathcal{H}_k^{\mathbb{C}}(S^n)$ and define $F_k(\sigma, \tau)$ by

$$F_k(\sigma, \tau) = \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\tau)}, \qquad \sigma, \tau \in S^n.$$

The following proposition is easy to prove (see Morimoto [113], Chapter 2, Lemma 1.19 and Lemma 2.20):

Proposition 16.12 *The function F_k is independent of the choice of orthonormal basis. Furthermore, for every orthogonal transformation, $R \in \mathbf{O}(n+1)$, we have*

$$F_k(R\sigma, R\tau) = F_k(\sigma, \tau), \qquad \sigma, \tau \in S^n.$$

Clearly, F_k is a symmetric function. Since we can pick an orthonormal basis of real orthogonal functions for $\mathcal{H}_k^{\mathbb{C}}(S^n)$ (pick a basis of $\mathcal{H}_k(S^n)$), Proposition 16.12 shows that F_k is a real-valued function.

The function F_k satisfies the following property which justifies its name, the *reproducing kernel* for $\mathcal{H}_k^{\mathbb{C}}(S^n)$:

Proposition 16.13 *For every spherical harmonic, $H \in \mathcal{H}_j^{\mathbb{C}}(S^n)$, we have*

$$\int_{S^n} H(\tau)F_k(\sigma, \tau)\, d\tau = \delta_{j\,k} H(\sigma), \qquad \sigma, \tau \in S^n,$$

for all $j, k \geq 0$.

Proof. When $j \neq k$, since $\mathcal{H}_k^{\mathbb{C}}(S^n)$ and $\mathcal{H}_j^{\mathbb{C}}(S^n)$ are orthogonal and since $F_k(\sigma, \tau) = \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\tau)}$, it is clear that the integral in Proposition 16.13 vanishes. When $j = k$, we have

$$\begin{aligned}
\int_{S^n} H(\tau)F_k(\sigma, \tau)\, d\tau &= \int_{S^n} H(\tau) \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\tau)}\, d\tau \\
&= \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma) \int_{S^n} H(\tau)\overline{Y_k^i(\tau)}\, d\tau \\
&= \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\, \langle H, Y_k^i \rangle \\
&= H(\sigma),
\end{aligned}$$

since $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ is an orthonormal basis. $\square$

In Stein and Weiss [142] (Chapter 4), the function $F_k(\sigma, \tau)$ is denoted by $Z_\sigma^{(k)}(\tau)$ and it is called the *zonal harmonic of degree k with pole σ*.

The value, $F_k(\sigma, \tau)$, of the function F_k depends only on $\sigma \cdot \tau$, as stated in Proposition 16.15 which is proved in Morimoto [113] (Chapter 2, Lemma 2.23). The following proposition also proved in Morimoto [113] (Chapter 2, Lemma 2.21) is needed to prove Proposition 16.15:

Proposition 16.14 *For all $\sigma, \tau, \sigma', \tau' \in S^n$, with $n \geq 1$, the following two conditions are equivalent:*

(i) There is some orthogonal transformation, $R \in \mathbf{O}(n+1)$, such that $R(\sigma) = \sigma'$ and $R(\tau) = \tau'$.

(ii) $\sigma \cdot \tau = \sigma' \cdot \tau'$.

Propositions 16.13 and 16.14 immediately yield

Proposition 16.15 *For all $\sigma, \tau, \sigma', \tau' \in S^n$, if $\sigma \cdot \tau = \sigma' \cdot \tau'$, then $F_k(\sigma, \tau) = F_k(\sigma', \tau')$. Consequently, there is some function, $\varphi \colon \mathbb{R} \to \mathbb{R}$, such that $F_k(\omega, \tau) = \varphi(\omega \cdot \tau)$.*

We are now ready to define zonal functions. Remarkably, the function φ in Proposition 16.15 comes from a real polynomial. We need the following proposition which is of independent interest:

Proposition 16.16 *If P is any (complex) polynomial in n variables such that*

$$P(R(x)) = P(x) \qquad \text{for all rotations, } R \in \mathbf{SO}(n), \text{ and all } x \in \mathbb{R}^n,$$

then P is of the form

$$P(x) = \sum_{j=0}^{m} c_j (x_1^2 + \cdots + x_n^2)^j,$$

for some $c_0, \ldots, c_m \in \mathbb{C}$.

Proof. Write P as the sum of its homogeneous pieces, $P = \sum_{l=0}^{k} Q_l$, where Q_l is homogeneous of degree l. Then, for every $\epsilon > 0$ and every rotation, R, we have

$$\sum_{l=0}^{k} \epsilon^l Q_l(x) = P(\epsilon x) = P(R(\epsilon x)) = P(\epsilon R(x)) = \sum_{l=0}^{k} \epsilon^l Q_l(R(x)),$$

which implies that

$$Q_l(R(x)) = Q_l(x), \qquad l = 0, \ldots, k.$$

If we let $F_l(x) = \|x\|^{-l} Q_l(x)$, then F_l is a homogeneous function of degree 0 and F_l is invariant under all rotations. This is only possible if F_l is a constant function, thus $F_l(x) = a_l$ for all $x \in \mathbb{R}^n$. But then, $Q_l(x) = a_l \|x\|^l$. Since Q_l is a polynomial, l must be even whenever $a_l \neq 0$. It follows that

$$P(x) = \sum_{j=0}^{m} c_j \|x\|^{2j}$$

with $c_j = a_{2j}$ for $j = 0, \ldots, m$ and where m is the largest integer $\leq k/2$. □

Proposition 16.16 implies that if a polynomial function on the sphere, S^n, in particular, a spherical harmonic, is invariant under all rotations, then it is a constant. If we relax this condition to invariance under all rotations leaving some given point, $\tau \in S^n$, invariant, then we obtain zonal harmonics.

The following theorem from Morimoto [113] (Chapter 2, Theorem 2.24) gives the relationship between zonal harmonics and the Gegenbauer polynomials:

Theorem 16.17 *Fix any $\tau \in S^n$. For every constant, $c \in \mathbb{C}$, there is a unique homogeneous harmonic polynomial, $Z_k^\tau \in \mathcal{H}_k^{\mathbb{C}}(n+1)$, satisfying the following conditions:*

(1) $Z_k^\tau(\tau) = c$;

(2) For every rotation, $R \in \mathbf{SO}(n+1)$, if $R\tau = \tau$, then $Z_k^\tau(R(x)) = Z_k^\tau(x)$, for all $x \in \mathbb{R}^{n+1}$.

Furthermore, we have

$$Z_k^\tau(x) = c \|x\|^k \, P_{k,n}\left(\frac{x}{\|x\|} \cdot \tau\right),$$

for some polynomial, $P_{k,n}(t)$, of degree k.

Remark: The proof given in Morimoto [113] is essentially the same as the proof of Theorem 2.12 in Stein and Weiss [142] (Chapter 4) but Morimoto makes an implicit use of Proposition 16.16 above. Also, Morimoto states Theorem 16.17 only for $c = 1$ but the proof goes through for any $c \in \mathbb{C}$, including $c = 0$, and we will need this extra generality in the proof of the Funk-Hecke formula.

Proof. Let $e_{n+1} = (0, \ldots, 0, 1) \in \mathbb{R}^{n+1}$ and for any $\tau \in S^n$, let R_τ be some rotation such that $R_\tau(e_{n+1}) = \tau$. Assume $Z \in \mathcal{H}_k^{\mathbb{C}}(n+1)$ satisfies conditions (1) and (2) and let Z' be given by $Z'(x) = Z(R_\tau(x))$. As $R_\tau(e_{n+1}) = \tau$, we have $Z'(e_{n+1}) = Z(\tau) = c$. Furthermore, for any rotation, S, such that $S(e_{n+1}) = e_{n+1}$, observe that

$$R_\tau \circ S \circ R_\tau^{-1}(\tau) = R_\tau \circ S(e_{n+1}) = R_\tau(e_{n+1}) = \tau,$$

and so, as Z satisfies property (2) for the rotation $R_\tau \circ S \circ R_\tau^{-1}$, we get

$$Z'(S(x)) = Z(R_\tau \circ S(x)) = Z(R_\tau \circ S \circ R_\tau^{-1} \circ R_\tau(x)) = Z(R_\tau(x)) = Z'(x),$$

which proves that Z' is a harmonic polynomial satisfying properties (1) and (2) with respect to e_{n+1}. Therefore, we may assume that $\tau = e_{n+1}$.

Write

$$Z(x) = \sum_{j=0}^{k} x_{n+1}^{k-j} P_j(x_1, \ldots, x_n),$$

where $P_j(x_1, \ldots, x_n)$ is a homogeneous polynomial of degree j. Since Z is invariant under every rotation, R, fixing e_{n+1} and since the monomials x_{n+1}^{k-j} are clearly invariant under such a rotation, we deduce that every $P_j(x_1, \ldots, x_n)$ is invariant under all rotations of $\mathbb{R}^n$ (clearly, there is a one-two-one correspondence between the rotations of $\mathbb{R}^{n+1}$ fixing e_{n+1} and the rotations of $\mathbb{R}^n$). By Proposition 16.16, we conclude that

$$P_j(x_1, \ldots, x_n) = c_j(x_1^2 + \cdots + x_n^2)^{\frac{j}{2}},$$

which implies that $P_j = 0$ is j is odd. Thus, we can write

$$Z(x) = \sum_{i=0}^{[k/2]} c_i x_{n+1}^{k-2i}(x_1^2 + \cdots + x_n^2)^i$$

where $[k/2]$ is the greatest integer, m, such that $2m \leq k$. If $k < 2$, then $Z = c_0$, so $c_0 = c$ and Z is uniquely determined. If $k \geq 2$, we know that Z is a harmonic polynomial so we assert that $\Delta Z = 0$. A simple computation shows that

$$\Delta(x_1^2 + \cdots + x_n^2)^i = 2i(n + 2i - 2)(x_1^2 + \cdots + x_n^2)^{i-1}$$

and

$$\begin{aligned}
\Delta x_{n+1}^{k-2i}(x_1^2 + \cdots + x_n^2)^i &= (k-2i)(k-2i-1)x_{n+1}^{k-2i-2}(x_1^2 + \cdots + x_n^2)^i \\
&\quad + x_{n+1}^{k-2i}\Delta(x_1^2 + \cdots + x_n^2)^i \\
&= (k-2i)(k-2i-1)x_{n+1}^{k-2i-2}(x_1^2 + \cdots + x_n^2)^i \\
&\quad + 2i(n+2i-2)x_{n+1}^{k-2i}(x_1^2 + \cdots + x_n^2)^{i-1},
\end{aligned}$$

so we get

$$\Delta Z = \sum_{i=0}^{[k/2]-1} \left((k-2i)(k-2i-1)c_i + 2(i+1)(n+2i)c_{i+1}\right) x_{n+1}^{k-2i-2}(x_1^2 + \cdots + x_n^2)^i.$$

Then, $\Delta Z = 0$ yields the relations

$$2(i+1)(n+2i)c_{i+1} = -(k-2i)(k-2i-1)c_i, \qquad i = 0, \ldots, [k/2] - 1,$$

which shows that Z is uniquely determined up to the constant c_0. Since we are requiring $Z(e_{n+1}) = c$, we get $c_0 = c$ and Z is uniquely determined. Now, on S^n, we have $x_1^2 + \cdots + x_{n+1}^2 = 1$, so if we let $t = x_{n+1}$, for $c_0 = 1$, we get a polynomial in one variable,

$$P_{k,n}(t) = \sum_{i=0}^{[k/2]} c_i t^{k-2i}(1-t^2)^i.$$

Thus, we proved that when $Z(e_{n+1}) = c$, we have

$$Z(x) = c\,\|x\|^k\; P_{k,n}\left(\frac{x_{n+1}}{\|x\|}\right) = c\,\|x\|^k\; P_{k,n}\left(\frac{x}{\|x\|}\cdot e_{n+1}\right).$$

When $Z(\tau) = c$, we write $Z = Z' \circ R_\tau^{-1}$ with $Z' = Z \circ R_\tau$ and where R_τ is a rotation such that $R_\tau(e_{n+1}) = \tau$. Then, as $Z'(e_{n+1}) = c$, using the formula above for Z', we have

$$\begin{aligned} Z(x) = Z'(R_\tau^{-1}(x)) &= c\left\|R_\tau^{-1}(x)\right\|^k\; P_{k,n}\left(\frac{R_\tau^{-1}(x)}{\|R_\tau^{-1}(x)\|}\cdot e_{n+1}\right) \\ &= c\,\|x\|^k\; P_{k,n}\left(\frac{x}{\|x\|}\cdot R_\tau(e_{n+1})\right) \\ &= c\,\|x\|^k\; P_{k,n}\left(\frac{x}{\|x\|}\cdot \tau\right), \end{aligned}$$

since R_τ is an isometry. □

The function, Z_k^τ, is called a *zonal function* and its restriction to S^n is a *zonal spherical function*. The polynomial, $P_{k,n}$, is called the *Gegenbauer polynomial* of degree k and dimension $n+1$ or *ultraspherical polynomial*. By definition, $P_{k,n}(1) = 1$.

The proof of Theorem 16.17 shows that for k even, say $k = 2m$, the polynomial $P_{2m,n}$ is of the form

$$P_{2m,n} = \sum_{j=0}^{m} c_{m-j} t^{2j}(1-t^2)^{m-j}$$

and for k odd, say $k = 2m+1$, the polynomial $P_{2m+1,n}$ is of the form

$$P_{2m+1,n} = \sum_{j=0}^{m} c_{m-j} t^{2j+1}(1-t^2)^{m-j}.$$

Consequently, $P_{k,n}(-t) = (-1)^k P_{k,n}(t)$, for all $k \geq 0$. The proof also shows that the "natural basis" for these polynomials consists of the polynomials, $t^i(1-t^2)^{\frac{k-i}{2}}$, with $k-i$ even. Indeed, with this basis, there are simple recurrence equations for computing the coefficients of $P_{k,n}$.

Remark: Morimoto [113] calls the polynomials, $P_{k,n}$, "Legendre polynomials". For $n = 2$, they are indeed the Legendre polynomials. Stein and Weiss denotes our (and Morimoto's) $P_{k,n}$ by $P_k^{\frac{n-1}{2}}$ (up to a constant factor) and Dieudonné [43] (Chapter 7) by $G_{k,n+1}$.

When $n = 2$, using the notation of Section 16.2, the zonal functions on S^2 are the spherical harmonics, y_l^0, for which $m = 0$, that is (up to a constant factor),

$$y_l^0(\theta, \varphi) = \sqrt{\frac{(2l+1)}{4\pi}}\, P_l(\cos\theta),$$

where P_l is the Legendre polynomial of degree l. For example, for $l = 2$, $P_l(t) = \frac{1}{2}(3t^2 - 1)$.

If we put $Z(r^k\sigma) = r^k F_k(\sigma, \tau)$ for a fixed τ, then by the definition of $F_k(\sigma, \tau)$ it is clear that Z is a homogeneous harmonic polynomial. The value $F_k(\tau, \tau)$ does not depend of τ because by transitivity of the action of $\mathbf{SO}(n+1)$ on S^n, for any other $\sigma \in S^n$, there is some rotation, R, so that $R\tau = \sigma$ and by Proposition 16.12, we have $F_k(\sigma, \sigma) = F_k(R\tau, R\tau) = F_k(\tau, \tau)$. To compute $F_k(\tau, \tau)$, since

$$F_k(\tau, \tau) = \sum_{i=1}^{a_{k,n+1}} \left\| Y_k^i(\tau) \right\|^2,$$

and since $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ is an orthonormal basis of $\mathcal{H}_k^{\mathbb{C}}(S^n)$, observe that

$$\begin{aligned} a_{k,n+1} &= \sum_{i=1}^{a_{k,n+1}} \int_{S^n} \left\| Y_k^i(\tau) \right\|^2 d\tau && (16.1)\\ &= \int_{S^n} \left(\sum_{i=1}^{a_{k,n+1}} \left\| Y_k^i(\tau) \right\|^2 \right) d\tau && (16.2)\\ &= \int_{S^n} F_k(\tau, \tau)\, d\tau = F_k(\tau, \tau)\, \mathrm{vol}(S^n). && (16.3) \end{aligned}$$

Therefore,

$$F_k(\tau, \tau) = \frac{a_{k,n+1}}{\mathrm{vol}(S^n)}.$$

Beware that Morimoto [113] uses the normalized measure on S^n, so the factor involving $\mathrm{vol}(S^n)$ does not appear.

Remark: Recall that

$$\mathrm{vol}(S^{2d}) = \frac{2^{d+1}\pi^d}{1 \cdot 3 \cdots (2d-1)} \quad \text{if} \quad d \geq 1 \quad \text{and} \quad \mathrm{vol}(S^{2d+1}) = \frac{2\pi^{d+1}}{d!} \quad \text{if} \quad d \geq 0.$$

Now, if $R\tau = \tau$, then Proposition 16.12 shows that

$$Z(R(r^k\sigma)) = Z(r^k R(\sigma)) = r^k F_k(R\sigma, \tau) = r^k F_k(R\sigma, R\tau) = r^k F_k(\sigma, \tau) = Z(r^k\sigma).$$

Therefore, the function Z_k^τ satisfies conditions (1) and (2) of Theorem 16.17 with $c = \frac{a_{k,n+1}}{\mathrm{vol}(S^n)}$ and by uniqueness, we get

$$F_k(\sigma, \tau) = \frac{a_{k,n+1}}{\mathrm{vol}(S^n)}\, P_{k,n}(\sigma \cdot \tau).$$

Consequently, we have obtained the so-called *addition formula*:

Proposition 16.18 *(Addition Formula) If* $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ *is any orthonormal basis of* $\mathcal{H}_k^{\mathbb{C}}(S^n)$*, then*

$$P_{k,n}(\sigma \cdot \tau) = \frac{\mathrm{vol}(S^n)}{a_{k,n+1}} \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\tau)}.$$

Again, beware that Morimoto [113] does not have the factor $\mathrm{vol}(S^n)$.

For $n = 1$, we can write $\sigma = (\cos\theta, \sin\theta)$ and $\tau = (\cos\varphi, \sin\varphi)$ and it is easy to see that the addition formula reduces to

$$P_{k,1}(\cos(\theta - \varphi)) = \cos k\theta \cos k\varphi + \sin k\theta \sin k\varphi = \cos k(\theta - \varphi),$$

the standard addition formula for trigonometric functions.

Proposition 16.18 implies that $P_{k,n}$ has real coefficients. Furthermore Proposition 16.13 is reformulated as

$$\frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} P_{k,n}(\sigma \cdot \tau)H(\tau)\, d\tau = \delta_{j\,k} H(\sigma), \tag{rk}$$

showing that the Gengenbauer polynomials are reproducing kernels. A neat application of this formula is a formula for obtaining the kth spherical harmonic component of a function, $f \in L^2_{\mathbb{C}}(S^n)$.

Proposition 16.19 *For every function,* $f \in L^2_{\mathbb{C}}(S^n)$*, if* $f = \sum_{k=0}^{\infty} f_k$ *is the unique decomposition of* f *over the Hilbert sum* $\bigoplus_{k=0}^{\infty} \mathcal{H}_k^{\mathbb{C}}(S^k)$*, then* f_k *is given by*

$$f_k(\sigma) = \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} f(\tau)P_{k,n}(\sigma \cdot \tau)\, d\tau,$$

for all $\sigma \in S^n$*.*

Proof. If we recall that $\mathcal{H}_k^{\mathbb{C}}(S^k)$ and $\mathcal{H}_j^{\mathbb{C}}(S^k)$ are orthogonal for all $j \neq k$, using the formula (rk), we have

$$\begin{aligned}
\frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} f(\tau)P_{k,n}(\sigma \cdot \tau)\, d\tau &= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} \sum_{j=0}^{\infty} f_j(\tau)P_{k,n}(\sigma \cdot \tau)\, d\tau \\
&= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \sum_{j=0}^{\infty} \int_{S^n} f_j(\tau)P_{k,n}(\sigma \cdot \tau)\, d\tau \\
&= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} f_k(\tau)P_{k,n}(\sigma \cdot \tau)\, d\tau \\
&= f_k(\sigma),
\end{aligned}$$

as claimed. □

We know from the previous section that the kth zonal function generates $\mathcal{H}_k^{\mathbb{C}}(S^n)$. Here is an explicit way to prove this fact.

Proposition 16.20 *If $H_1, \ldots, H_m \in \mathcal{H}_k^{\mathbb{C}}(S^n)$ are linearly independent, then there are m points, $\sigma_1, \ldots, \sigma_m$, on S^n, so that the $m \times m$ matrix, $(H_j(\sigma_i))$, is invertible.*

Proof. We proceed by induction on m. The case $m = 1$ is trivial. For the induction step, we may assume that we found m points, $\sigma_1, \ldots, \sigma_m$, on S^n, so that the $m \times m$ matrix, $(H_j(\sigma_i))$, is invertible. Consider the function

$$\sigma \mapsto \begin{vmatrix} H_1(\sigma) & \ldots & H_m(\sigma) & H_{m+1}(\sigma) \\ H_1(\sigma_1) & \ldots & H_m(\sigma_1) & H_{m+1}(\sigma_1) \\ \vdots & \ddots & \vdots & \vdots \\ H_1(\sigma_m) & \ldots & H_m(\sigma_m) & H_{m+1}(\sigma_m). \end{vmatrix}$$

Since $H_1, \ldots, H_{m+1}$ are linearly independent, the above function does not vanish for all σ since otherwise, by expanding this determinant with respect to the first row, we get a linear dependence among the H_j's where the coefficient of H_{m+1} is nonzero. Therefore, we can find σ_{m+1} so that the $(m+1) \times (m+1)$ matrix, $(H_j(\sigma_i))$, is invertible. □

We say that $a_{k,n+1}$ points, $\sigma_1, \ldots, \sigma_{a_{k,n+1}}$ on S^n form a *fundamental system* iff the $a_{k,n+1} \times a_{k,n+1}$ matrix, $(P_{n,k}(\sigma_i \cdot \sigma_j))$, is invertible.

Theorem 16.21 *The following properties hold:*

(i) There is a fundamental system, $\sigma_1, \ldots, \sigma_{a_{k,n+1}}$, for every $k \geq 1$.

(ii) Every spherical harmonic, $H \in \mathcal{H}_k^{\mathbb{C}}(S^n)$, can be written as

$$H(\sigma) = \sum_{j=1}^{a_{k,n+1}} c_j \, P_{k,n}(\sigma_j \cdot \sigma),$$

for some unique $c_j \in \mathbb{C}$.

Proof. (i) By the addition formula,

$$P_{k,n}(\sigma_i \cdot \sigma_j) = \frac{\mathrm{vol}(S^n)}{a_{k,n+1}} \sum_{l=1}^{a_{k,n+1}} Y_k^l(\sigma_i)\overline{Y_k^l(\sigma_j)}$$

for any orthonormal basis, $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$. It follows that the matrix $(P_{k,n}(\sigma_i \cdot \sigma_j))$ can be written as

$$(P_{k,n}(\sigma_i \cdot \sigma_j)) = \frac{\mathrm{vol}(S^n)}{a_{k,n+1}} \, YY^*,$$

where $Y = (Y_k^l(\sigma_i))$, and by Proposition 16.20, we can find $\sigma_1, \ldots, \sigma_{a_{k,n+1}} \in S^n$ so that Y and thus also Y^* are invertible and so, $(P_{n,k}(\sigma_i \cdot \sigma_j))$ is invertible.

(ii) Again, by the addition formula,

$$P_{k,n}(\sigma \cdot \sigma_j) = \frac{\mathrm{vol}(S^n)}{a_{k,n+1}} \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\sigma_j)}.$$

However, as $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ is an orthonormal basis, (i) proved that the matrix Y^* is invertible so the $Y_k^i(\sigma)$ can be expressed uniquely in terms of the $P_{k,n}(\sigma \cdot \sigma_j)$, as claimed. □

A neat geometric characterization of the zonal spherical functions is given in Stein and Weiss [142]. For this, we need to define the notion of a parallel on S^n. A *parallel of S^n orthogonal to a point $\tau \in S^n$* is the intersection of S^n with any (affine) hyperplane orthogonal to the line through the center of S^n and τ. Clearly, any rotation, R, leaving τ fixed leaves every parallel orthogonal to τ globally invariant and for any two points, σ_1 and σ_2, on such a parallel there is a rotation leaving τ fixed that maps σ_1 to σ_2. Consequently, the zonal function, Z_k^τ, defined by τ is constant on the parallels orthogonal to τ. In fact, this property characterizes zonal harmonics, up to a constant.

The theorem below is proved in Stein and Weiss [142] (Chapter 4, Theorem 2.12). The proof uses Proposition 16.16 and it is very similar to the proof of Theorem 16.17 so, to save space, it is omitted.

Theorem 16.22 *Fix any point, $\tau \in S^n$. A spherical harmonic, $Y \in \mathcal{H}_k^{\mathbb{C}}(S^n)$, is constant on parallels orthogonal to τ iff $Y = cZ_k^\tau$, for some constant, $c \in \mathbb{C}$.*

In the next section, we show how the Gegenbauer polynomials can actually be computed.

16.7 More on the Gegenbauer Polynomials

The Gegenbauer polynomials are characterized by a formula generalizing the Rodrigues formula defining the Legendre polynomials (see Section 16.2). The expression

$$\left(k + \frac{n-2}{2}\right)\left(k - 1 + \frac{n-2}{2}\right)\cdots\left(1 + \frac{n-2}{2}\right)$$

can be expressed in terms of the Γ function as

$$\frac{\Gamma\left(k + \frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Recall that the Γ function is a generalization of factorial that satisfies the equation

$$\Gamma(z+1) = z\Gamma(z).$$

For $z = x + iy$ with $x > 0$, $\Gamma(z)$ is given by

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt,$$

where the integral converges absolutely. If n is an integer $n \geq 0$, then $\Gamma(n+1) = n!$.

It is proved in Morimoto [113] (Chapter 2, Theorem 2.35) that

Proposition 16.23 *The Gegenbauer polynomial, $P_{k,n}$, is given by Rodrigues' formula:*

$$P_{k,n}(t) = \frac{(-1)^k}{2^k} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(k + \frac{n}{2}\right)} \frac{1}{(1-t^2)^{\frac{n-2}{2}}} \frac{d^k}{dt^k}(1-t^2)^{k+\frac{n-2}{2}},$$

with $n \geq 2$.

The Gegenbauer polynomials satisfy the following orthogonality properties with respect to the kernel $(1-t^2)^{\frac{n-2}{2}}$ (see Morimoto [113] (Chapter 2, Theorem 2.34):

Proposition 16.24 *The Gegenbauer polynomial, $P_{k,n}$, have the following properties:*

$$\begin{aligned} \int_{-1}^{-1} (P_{k,n}(t))^2 (1-t^2)^{\frac{n-2}{2}}\, dt &= \frac{\mathrm{vol}(S^n)}{a_{k,n+1}\mathrm{vol}(S^{n-1})} \\ \int_{-1}^{-1} P_{k,n}(t) P_{l,n}(t) (1-t^2)^{\frac{n-2}{2}}\, dt &= 0, \qquad k \neq l. \end{aligned}$$

The Gegenbauer polynomials satisfy a second-order differential equation generalizing the Legendre equation from Section 16.2.

Proposition 16.25 *The Gegenbauer polynomial, $P_{k,n}$, are solutions of the differential equation*

$$(1-t^2)P''_{k,n}(t) - ntP'_{k,n}(t) + k(k+n-1)P_{k,n}(t) = 0.$$

Proof. For a fixed τ, the function H given by $H(\sigma) = P_{k,n}(\sigma \cdot \tau) = P_{k,n}(\cos\theta)$, belongs to $\mathcal{H}_k^{\mathbb{C}}(S^n)$, so

$$\Delta_{S^n} H = -k(k+n-1)H.$$

Recall from Section 16.3 that

$$\Delta_{S^n} f = \frac{1}{\sin^{n-1}\theta} \frac{\partial}{\partial\theta}\left(\sin^{n-1}\theta \frac{\partial f}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\Delta_{S_{n-1}} f,$$

in the local coordinates where

$$\sigma = \sin\theta\, \widetilde{\sigma} + \cos\theta\, e_{n+1},$$

with $\widetilde{\sigma} \in S^{n-1}$ and $0 \leq \theta < \pi$. If we make the change of variable $t = \cos\theta$, then it is easy to see that the above formula becomes

$$\Delta_{S^n} f = (1-t^2)\frac{\partial^2 f}{\partial t^2} - nt\frac{\partial f}{\partial t} + \frac{1}{1-t^2}\Delta_{S^{n-1}}$$

(see Morimoto [113], Chapter 2, Theorem 2.9.) But, H being zonal, it only depends on θ, that is, on t, so $\Delta_{S^{n-1}} H = 0$ and thus,

$$-k(k+n-1)P_{k,n}(t) = \Delta_{S^n} P_{k,n}(t) = (1-t^2)\frac{\partial^2 P_{k,n}}{\partial t^2} - nt\frac{\partial P_{k,n}}{\partial t},$$

which yields our equation. □

Note that for $n = 2$, the differential equation of Proposition 16.25 is the Legendre equation from Section 16.2.

The Gegenbauer poynomials also appear as coefficients in some simple generating functions. The following proposition is proved in Morimoto [113] (Chapter 2, Theorem 2.53 and Theorem 2.55):

Proposition 16.26 *For all r and t such that $-1 < r < 1$ and $-1 \leq t \leq 1$, for all $n \geq 1$, we have the following generating formula:*

$$\sum_{k=0}^{\infty} a_{k,n+1}\, r^k P_{k,n}(t) = \frac{1-r^2}{(1-2rt+r^2)^{\frac{n+1}{2}}}.$$

Furthermore, for all r and t such that $0 \leq r < 1$ and $-1 \leq t \leq 1$, if $n = 1$, then

$$\sum_{k=1}^{\infty} \frac{r^k}{k}\, P_{k,1}(t) = -\frac{1}{2}\log(1-2rt+r^2)$$

and if $n \geq 2$, then

$$\sum_{k=0}^{\infty} \frac{n-1}{2k+n-1}\, a_{k,n+1}\, r^k P_{k,n}(t) = \frac{1}{(1-2rt+r^2)^{\frac{n-1}{2}}}.$$

In Stein and Weiss [142] (Chapter 4, Section 2), the polynomials, $P_k^{\lambda}(t)$, where $\lambda > 0$ are defined using the following generating formula:

$$\sum_{k=0}^{\infty} r^k P_k^{\lambda}(t) = \frac{1}{(1-2rt+r^2)^{\lambda}}.$$

Each polynomial, $P_k^{\lambda}(t)$, has degree k and is called an *ultraspherical polynomial of degree k associated with λ*. In view of Proposition 16.26, we see that

$$P_k^{\frac{n-1}{2}}(t) = \frac{n-1}{2k+n-1}\, a_{k,n+1}\, P_{k,n}(t),$$

as we mentionned ealier. There is also an integral formula for the Gegenbauer polynomials known as *Laplace representation*, see Morimoto [113] (Chapter 2, Theorem 2.52).

16.8 The Funk-Hecke Formula

The Funk-Hecke Formula (also known as Hecke-Funk Formula) basically allows one to perform a sort of convolution of a "kernel function" with a spherical function in a convenient way. Given a measurable function, K, on $[-1, 1]$ such that the integral

$$\int_{-1}^{1} |K(t)|(1-t^2)^{\frac{n-2}{2}}\, dt$$

makes sense, given a function $f \in L^2_{\mathbb{C}}(S^n)$, we can view the expression

$$K \star f(\sigma) = \int_{S^n} K(\sigma \cdot \tau) f(\tau)\, d\tau$$

as a sort of *convolution* of K and f. Actually, the use of the term convolution is really unfortunate because in a "true" convolution, $f * g$, either the argument of f or the argument of g should be multiplied by the inverse of the variable of integration, which means that the integration should really be taking place over the group $\mathbf{SO}(n+1)$. We will come back to this point later. For the time being, let us call the expression $K \star f$ defined above a *pseudo-convolution*. Now, if f is expressed in terms of spherical harmonics as

$$f = \sum_{k=0}^{\infty} \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_k^{m_k},$$

then the Funk-Hecke Formula states that

$$K \star Y_k^{m_k}(\sigma) = \alpha_k Y_k^{m_k}(\sigma),$$

for some fixed constant, α_k, and so

$$K \star f = \sum_{k=0}^{\infty} \sum_{m_k=1}^{a_{k,n+1}} \alpha_k c_{k,m_k} Y_k^{m_k}.$$

Thus, if the constants, α_k are known, then it is "cheap" to compute the pseudo-convolution $K \star f$.

This method was used in a ground-breaking paper by Basri and Jacobs [14] to compute the reflectance function, r, from the lighting function, ℓ, as a pseudo-convolution $K \star \ell$ (over S^2) with the *Lambertian kernel*, K, given by

$$K(\sigma \cdot \tau) = \max(\sigma \cdot \tau, 0).$$

Below, we give a proof of the Funk-Hecke formula due to Morimoto [113] (Chapter 2, Theorem 2.39) but see also Andrews, Askey and Roy [2] (Chapter 9). This formula was first published by Funk in 1916 and then by Hecke in 1918.

Theorem 16.27 *(Funk-Hecke Formula) Given any measurable function, K, on $[-1,1]$ such that the integral*

$$\int_{-1}^{1} |K(t)|(1-t^2)^{\frac{n-2}{2}}\, dt$$

makes sense, for every function, $H \in \mathcal{H}_k^{\mathbb{C}}(S^n)$, we have

$$\int_{S^n} K(\sigma\cdot\xi)H(\xi)\, d\xi = \left(\mathrm{vol}(S_{n-1})\int_{-1}^{1} K(t)P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\, dt\right) H(\sigma).$$

Observe that when $n=2$, the term $(1-t^2)^{\frac{n-2}{2}}$ is missing and we are simply requiring that $\int_{-1}^{1}|K(t)|\, dt$ makes sense.

Proof. We first prove the formula in the case where H is a zonal harmonic and then use the fact that the $P_{k,n}$'s are reproducing kernels (formula (rk)).

For all $\sigma,\tau \in S^n$ define H by

$$H(\sigma) = P_{k,n}(\sigma\cdot\tau)$$

and F by

$$F(\sigma,\tau) = \int_{S^n} K(\sigma\cdot\xi)P_{k,n}(\xi\cdot\tau)\, d\xi.$$

Since the volume form on the sphere is invariant under orientation-preserving isometries, for every $R \in \mathbf{SO}(n+1)$, we have

$$F(R\sigma, R\tau) = F(\sigma,\tau).$$

On the other hand, for σ fixed, it is not hard to see that as a function in τ, the function $F(\sigma,-)$ is a spherical harmonic, because $P_{k,n}$ satisfies a differential equation that implies that $\Delta_{S^2}F(\sigma,-) = -k(k+n-1)F(\sigma,-)$. Now, for every rotation, R, that fixes σ,

$$F(\sigma,\tau) = F(R\sigma, R\tau) = F(\sigma, R\tau),$$

which means that $F(\sigma,-)$ satisfies condition (2) of Theorem 16.17. By Theorem 16.17, we get

$$F(\sigma,\tau) = F(\sigma,\sigma)P_{k,n}(\sigma\cdot\tau).$$

If we use local coordinates on S^n where

$$\sigma = \sqrt{1-t^2}\,\widetilde{\sigma} + t\, e_{n+1},$$

with $\widetilde{\sigma} \in S^{n-1}$ and $-1 \le t \le 1$, it is not hard to show that the volume form on S^n is given by

$$d\sigma_{S^n} = (1-t^2)^{\frac{n-2}{2}}\, dt d\sigma_{S^{n-1}}.$$

Using this, we have

$$F(\sigma,\sigma) = \int_{S^n} K(\sigma\cdot\xi)P_{k,n}(\xi\cdot\sigma)\,d\xi = \mathrm{vol}(S^{n-1})\int_{-1}^{1} K(t)P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\,dt,$$

and thus,

$$F(\sigma,\tau) = \left(\mathrm{vol}(S^{n-1})\int_{-1}^{1} K(t)P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\,dt\right)P_{k,n}(\sigma\cdot\tau),$$

which is the Funk-Hecke formula when $H(\sigma) = P_{k,n}(\sigma\cdot\tau)$.

Let us now consider any function, $H \in \mathcal{H}_k^{\mathbb{C}}(S^n)$. Recall that by the reproducing kernel property (rk), we have

$$\frac{a_{k,n+1}}{\mathrm{vol}(S^n)}\int_{S^n} P_{k,n}(\xi\cdot\tau)H(\tau)\,d\tau = H(\xi).$$

Then, we can compute $\int_{S^n} K(\sigma\cdot\xi)H(\xi)\,d\xi$ using Fubini's Theorem and the Funk-Hecke formula in the special case where $H(\sigma) = P_{k,n}(\sigma\cdot\tau)$, as follows:

$$\begin{aligned}
&\int_{S^n} K(\sigma\cdot\xi)H(\xi)\,d\xi\\
&= \int_{S^n} K(\sigma\cdot\xi)\left(\frac{a_{k,n+1}}{\mathrm{vol}(S^n)}\int_{S^n} P_{k,n}(\xi\cdot\tau)H(\tau)\,d\tau\right)d\xi\\
&= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)}\int_{S^n} H(\tau)\left(\int_{S^n} K(\sigma\cdot\xi)P_{k,n}(\xi\cdot\tau)\,d\xi\right)d\tau\\
&= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)}\int_{S^n} H(\tau)\left(\left(\mathrm{vol}(S^{n-1})\int_{-1}^{1} K(t)P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\,dt\right)P_{k,n}(\sigma\cdot\tau)\right)d\tau\\
&= \left(\mathrm{vol}(S^{n-1})\int_{-1}^{1} K(t)P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\,dt\right)\left(\frac{a_{k,n+1}}{\mathrm{vol}(S^n)}\int_{S^n} P_{k,n}(\sigma\cdot\tau)H(\tau)\,d\tau\right)\\
&= \left(\mathrm{vol}(S^{n-1})\int_{-1}^{1} K(t)P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\,dt\right)H(\sigma),
\end{aligned}$$

which proves the Funk-Hecke formula in general. □

The Funk-Hecke formula can be used to derive an "addition theorem" for the ultraspherical polynomials (Gegenbauer polynomials). We omit this topic and we refer the interested reader to Andrews, Askey and Roy [2] (Chapter 9, Section 9.8).

Remark: Oddly, in their computation of $K\star\ell$, Basri and Jacobs [14] first expand K in terms of spherical harmonics as

$$K = \sum_{n=0}^{\infty} k_n Y_n^0,$$

and then use the Funk-Hecke formula to compute $K \star Y_n^m$ and they get (see page 222)

$$K \star Y_n^m = \alpha_n Y_n^m, \quad \text{with} \quad \alpha_n = \sqrt{\frac{4\pi}{2n+1}}\, k_n,$$

for some constant, k_n, given on page 230 of their paper (see below). However, there is no need to expand K as the Funk-Hecke formula yields directly

$$K \star Y_n^m(\sigma) = \int_{S^2} K(\sigma \cdot \xi) Y_n^m(\xi)\, d\xi = \left(\int_{-1}^{1} K(t) P_n(t)\, dt \right) Y_n^m(\sigma),$$

where $P_n(t)$ is the standard Legendre polynomial of degree n since we are in the case of S^2. By the definition of K ($K(t) = \max(t, 0)$) and since $\mathrm{vol}(S^1) = 2\pi$, we get

$$K \star Y_n^m = \left(2\pi \int_0^1 t P_n(t)\, dt \right) Y_n^m,$$

which is equivalent to Basri and Jacobs' formula (14) since their α_n on page 222 is given by

$$\alpha_n = \sqrt{\frac{4\pi}{2n+1}}\, k_n,$$

but from page 230,

$$k_n = \sqrt{(2n+1)\pi} \int_0^1 t P_n(t)\, dt.$$

What remains to be done is to compute $\int_0^1 t P_n(t)\, dt$, which is done by using the Rodrigues Formula and integrating by parts (see Appendix A of Basri and Jacobs [14]).

16.9 Convolution on G/K, for a Gelfand Pair (G, K)