CHAPTER 11

Sectional Curvature Comparison II

In the first section we explain how one can find generalized gradients for distance functions in situations where the function might not be smooth. This critical point technique is used in the proofs of all the big theorems in this chapter. The other important technique comes from Toponogov's theorem, which we prove in the next section. The first applications of these new ideas are to sphere theorems. We then prove the soul theorem of Cheeger and Gromoll. Next, we discuss Gromov's finiteness theorem for bounds on Betti numbers and generators for the fundamental group. Finally, we show that these techniques can be adapted to prove the Grove-Petersen homotopy finiteness theorem.

Toponogov's theorem is a very useful refinement of Gauss's early realization that curvature and angle excess of triangles are related. The fact that Toponogov's theorem can be used to get information about the topology of a space seems to originate with Berger's proof of the quarter pinched sphere theorem. Toponogov himself proved these theorems in order to establish the splitting theorem for manifolds with nonnegative sectional curvature and the maximal diameter theorem for manifolds with a positive lower bound for the sectional curvature. As we saw in chapter 9, these results now hold in the Ricci curvature setting. The next use of Toponogov was to the soul theorem of Cheeger-Gromoll-Meyer. However, Toponogov's theorem is not truly needed for any of the results mentioned so far. With little effort one can actually establish these theorems with more basic comparison techniques. Still, it is convenient to have a workhorse theorem of universal use. It wasn't until Grove and Shiohama developed critical point theory to prove their diameter sphere theorem that Toponogov's theorem was put to serious use. Shortly after that, Gromov put these two ideas to even more nontrivial use, with his Betti number estimate for manifolds with nonnegative sectional curvature. After that, it became clear that in working with manifolds that have lower sectional curvature bounds, the two key techniques are Toponogov's theorem and the critical point theory of Grove-Shiohama. These two very geometric techniques are still being used to prove many interesting and nontrivial results.

1. Critical Point Theory

In the particular generalized critical point theory developed here, the object is to define generalized gradients of continuous functions and then use these gradients to conclude that certain regions of a manifold have no topology. The motivating basic lemma is the following:

LEMMA 54. *Let (M, g) be a Riemannian manifold and $f : M \to \mathbb{R}$ a proper smooth function. If f has no critical values in the closed interval $[a, b]$, then the preimages $f^{-1}([-\infty, b])$ and $f^{-1}([-\infty, a])$ are diffeomorphic. Furthermore, there*

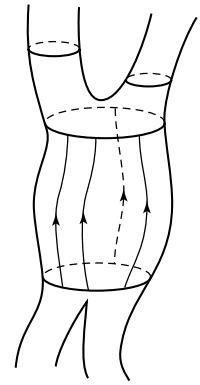

Figure 11.1

is a deformation retraction of $f^{-1}([-\infty,b])$ *onto* $f^{-1}([-\infty,a])$, *in particular, the inclusion*

$$f^{-1}([-\infty,a]) \hookrightarrow f^{-1}([-\infty,b])$$

is a homotopy equivalence.

PROOF. The idea is simply to move the level sets via the gradient of f. Since there are no critical points for f the gradient ∇f is nonzero everywhere on $f^{-1}([a,b])$. We then construct a bump function $\psi : M \to [0,1]$ that is 1 on the compact set $f^{-1}([a,b])$ and zero outside some compact neighborhood of $f^{-1}([a,b])$. Finally consider the vector field

$$X = \psi \cdot \frac{\nabla f}{|\nabla f|^2}$$

This vector field has compact support and must therefore be complete (integral curves are defined for all time). Let F^t denote the flow for this vector field. (See Figure 11.1)

For fixed $q \in M$ consider the function $t \to f(F^t(q))$. The derivative of this function is $g(X, \nabla f)$, so as long as the integral curve $t \to F^t(q)$ remains in $f^{-1}([a,b])$, the function $t \to f(F^t(q))$ is linear with derivative 1. In particular, the diffeomorphism $F^{b-a} : M \to M$ must carry $f^{-1}([-\infty,a])$ diffeomorphically onto $f^{-1}([-\infty,b])$.

Moreover, by flowing backwards we can define the desired retraction:

$$\begin{aligned} r_t &: f^{-1}([-\infty,b]) \to f^{-1}([-\infty,b]), \\ r_t(p) &= \begin{cases} p & \text{if } f(p) \le a, \\ F^{t(a-f(p))}(p) & \text{if } a \le f(p) \le b. \end{cases} \end{aligned}$$

Then $r_0 = id$, and r_1 maps $f^{-1}([-\infty,b])$ diffeomorphically onto $f^{-1}([-\infty,a])$. □

Notice that we used in an essential way that the function is proper to conclude that the vector field is complete. In fact, if we delete a single point from the region $f^{-1}([a,b])$, then the function still won't have any critical values, but clearly the conclusion of the lemma is false.

We shall now try to generalize this lemma to functions that are not even C^1. To minimize technicalities we shall work exclusively with distance functions. Suppose

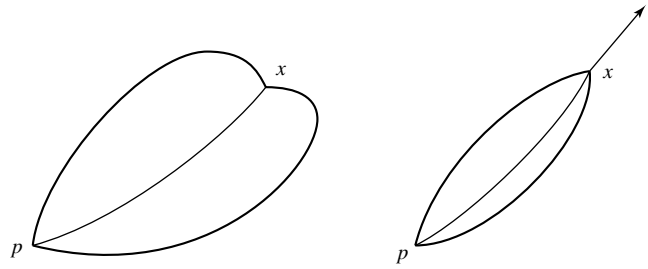

Figure 11.2

(M, g) is complete and $K \subset M$ a compact subset. Then the distance function

$$r(x) = d(x, K) = \min\{d(x, p) : p \in K\}$$

is proper. Wherever this function is smooth, we know that it has unit gradient and must therefore be noncritical at such points. However, it might also have local maxima, and at such points we certainly wouldn't want the function to be noncritical. To define the generalized gradient for such functions, let us list all the possible values it could have. Define $\Gamma(x, K)$, or simply $\Gamma(x)$, as the set of unit vectors in T_xM that are tangent to a segment from K to x. That is, $v \in \Gamma(x, K) \subset T_xM$ if there is a unit speed segment $\sigma : [0, \ell] \to M$ such that $\sigma(0) \in K$, $\sigma(\ell) = x$, and $v = \dot{\sigma}(\ell)$. Note that σ is chosen such that no shorter curve from x to K exists. There might, however, be several such segments. In the case where r is smooth at x, we clearly have that $\{\nabla r\} = \Gamma(x, K)$. At other points, $\Gamma(x, K)$ might contain more vectors. We say that r is *regular, or noncritical,* at x if the set $\Gamma(x, K)$ is contained in an open hemisphere of the unit sphere in T_xM. The center of such a hemisphere is then a possible averaged direction for the gradient of r at x. Stated differently, we have that r is regular at x iff there is a vector $v \in T_xM$ such that the angles $\angle(v, w) < \pi/2$ for all $w \in \Gamma(x, K)$. If v is a unit vector, then it will be the center of the desired hemisphere. We can quantify being regular by saying that r is α-regular at x if there exist $v \in T_xM$ such that $\angle(v, w) < \alpha$ for all $w \in \Gamma(x, K)$. Thus, r is regular at x iff it is $\pi/2$-regular. The set of vectors v that can be used in the definition of α-regularity is denoted by $G_\alpha f(x)$, where G stands for *generalized* gradient.

Evidently, a point x is critical for $d(\cdot, p)$ if the segments from p to x spread out at x, while it is regular if they more or less point in the same direction. (See Figure 11.2) It was Berger who first realized and showed that a local maximum must be critical in the above sense. Berger's result is a consequence of the next proposition.

PROPOSITION 47. *Suppose (M, g) and $r = d(\cdot, K)$ are as above. Then:*
(1) $\Gamma(x, K)$ is closed and therefore compact for all x.
(2) The set of α-regular points is open in M.
(3) $G_\alpha r(x)$ is convex for all $\alpha \leq \frac{\pi}{2}$.
(4) If U is an open set of α-regular points for r, then there is a unit vector field X on U such that $X(x) \in G_\alpha r(x)$ for all $x \in U$. Furthermore, if γ is an integral curve for X and $s < t$, then

$$r(\gamma(t)) - r(\gamma(s)) > \cos(\alpha)(t - s).$$

PROOF. (1) Let $\sigma_i : [0, \ell] \to M$ be a sequence of unit speed segments from K to x with $\dot{\sigma}_i(\ell)$ converging to some unit vector $v \in T_x M$. Clearly,

$$\sigma(t) = \exp_x((\ell - t)\, v)$$

is the limit of the segments σ_i and must therefore be a segment itself. Furthermore, since K is closed $\sigma(0) \in K$.

(2) Suppose $x_i \to x$, and x_i is not α-regular. We shall show that x is not α-regular. This means that for each $v \in T_x M$, we can find $w \in \Gamma(x, K)$ such that $\angle(v, w) \geq \alpha$. Now, for some fixed $v \in T_x M$, choose a sequence $v_i \in T_{x_i} M$ converging to v. For each i we can, by assumption, find $w_i \in \Gamma(x_i, K)$ with $\angle(v_i, w_i) \geq \alpha$. The sequence of unit vectors w_i must now subconverge to a vector $w \in T_x M$. Furthermore, the sequence of segments σ_i that generate w_i must also subconverge to a segment that is tangent to w. Thus, $w \in \Gamma(x, K)$.

(3) First observe that if $\alpha \leq \pi/2$, then for each $w \in T_x M$, the open cone

$$C_\alpha(w) = \{v \in T_x M : \angle(v, w) < \alpha\}$$

is convex. Then observe that $G_\alpha r(x)$ is the intersection of the cones $C_\alpha(w), w \in \Gamma(x, K)$, and is therefore itself convex.

(4) For each $p \in U$ we can find $v_p \in G_\alpha r(p)$. For each p, extend v_p to a vector field V_p. It now follows from the proof of (2) that $V_p(x) \in G_\alpha r(x)$ for x near p. We can then assume that V_p is defined on a neighborhood U_p on which it is a generalized gradient. We can now select a locally finite collection $\{U_i\}$ of U_p's and a corresponding partition of unity λ_i. Then property (3) tells us that the vector field

$$V = \sum \lambda_i V_i \in G_\alpha r.$$

In particular, it is nonzero and can therefore be normalized to a unit vector field.

The last property is clearly true at points where r is smooth, because in that case the derivative of $t \to r \circ \gamma$ is

$$g(X, \nabla r) = \cos \angle(X, \nabla r) > \cos \alpha.$$

Now observe that since r is Lipschitz continuous, this function is at least absolutely continuous. This implies that $r \circ \gamma$ is differentiable a.e. and is the integral of its derivative. It might, however, happen that $r \circ \gamma$ is differentiable at a point x where ∇r is not defined. To see what happens at such points we select a variation $\bar{\gamma}(s, t)$ such that $t \to \bar{\gamma}(0, t)$ is a segment from K to x, $\bar{\gamma}(s, 0) = \bar{\gamma}(0, 0)$, and $\bar{\gamma}(s, 1) = \gamma(s)$ is the integral curve for X through x. Thus

$$\begin{aligned} \frac{1}{2}(r \circ \gamma)^2 &\leq \frac{1}{2}\left(\int_0^1 \left|\frac{\partial \gamma}{\partial t}\right| dt\right)^2 \\ &\leq \frac{1}{2}\int_0^1 \left|\frac{\partial \gamma}{\partial t}\right|^2 dt \\ &= E(\gamma_s) \end{aligned}$$

with equality holding for $s = 0$. Assuming that $r \circ \gamma$ is differentiable at $s = 0$ we get

$$\begin{aligned} r(\gamma(0)) \frac{dr \circ \gamma}{dt}|_{s=0} &= \frac{dE}{ds}|_{s=0} \\ &= g\left(\frac{\partial \bar{\gamma}}{\partial t}(0,b), \frac{\partial \bar{\gamma}}{\partial s}(0,b)\right) \\ &= g\left(\frac{\partial \bar{\gamma}}{\partial t}(0,b), X\right) \\ &= \left|\frac{\partial \bar{\gamma}}{\partial t}\right| \cos\left(\angle\left(X, \frac{\partial \bar{\gamma}}{\partial t}\right)\right) \\ &> r(\gamma(0)) \cos\alpha. \end{aligned}$$

This proves the desired property. □

We can now generalize the above retraction lemma.

LEMMA 55. *Let (M, g) and $r = d(\cdot, K)$ be as above. Suppose that all points in $r^{-1}([a,b])$ are α-regular for $\alpha < \pi/2$. Then $r^{-1}([-\infty, a])$ is homeomorphic to $r^{-1}([-\infty, b])$, and $r^{-1}([-\infty, b])$ deformation retracts onto $r^{-1}([-\infty, a])$.*

PROOF. The construction is similar to the first lemma but a little more involved. We can construct a compactly supported vector field X such that the flow F^t for X satisfies

$$r\left(F^t(p)\right) - r(p) > t \cdot \cos(\alpha),\ t \geq 0 \text{ if } p, F^t(p) \in r^{-1}([a,b]).$$

For each $p \in r^{-1}(b)$ we can therefore find a first time $t_p \leq \frac{b-a}{\cos\alpha}$ for which $F^{-t_p}(p) \in r^{-1}(a)$. The function $p \to t_p$ is continuous and thus we get the desired retraction

$$\begin{aligned} r_t &: r^{-1}([-\infty, b]) \to r^{-1}([-\infty, b]), \\ r_t(p) &= \begin{cases} p & \text{if } r(p) \leq a \\ F^{-t \cdot t_p}(p) & \text{if } a \leq r(p) \leq b \end{cases}. \end{aligned}$$

□

Note that as the level sets for r are not smooth, we can't expect to get diffeomorphic sublevels. It is now a question of how this can be used. As a very simple result let us mention

COROLLARY 43. *Suppose K is a compact submanifold of a complete Riemannian manifold (M, g) and suppose the distance function $r = d(\cdot, K)$ is regular everywhere on $M - K$. Then M is diffeomorphic to the normal bundle of K in M. In particular, if $K = \{p\}$, then M is diffeomorphic to $\mathbb{R}^n$.*

PROOF. We know that $M - K$ admits a vector field X, such that r is strictly increasing along the integral curves for X. Moreover, near K the distance function is smooth, and therefore X can be assumed to be equal to ∇r near K.

If

$$\nu(K) = \{v \in T_pM : p \in K \text{ and } v \perp T_pK\},$$

then we have the normal exponential map

$$\exp : \nu(K) \to M.$$

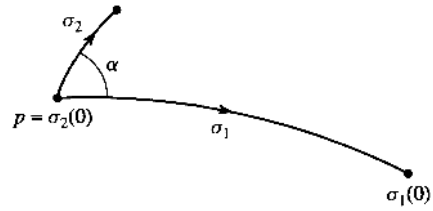

Figure 11.3

On a neighborhood of the zero section in $\nu(K)$ we know that this gives a diffeomorphism onto a neighborhood of K. Also, the curves $t \to \exp(tv)$ are, for small t, integral curves for X. In particular, we have for each $v \in \nu(K)$ a unique integral curve for X denoted $\gamma_v(t) : (0,\infty) \to M$ such that $\lim_{t\to 0} \dot{\gamma}_v(t) = v$. Now define our diffeomorphism $F : \nu(K) \to M$ by

$$\begin{aligned} F(0_p) &= p \text{ for the origin in } \nu_p(K), \\ F(tv) &= \gamma_v(t) \text{ where } |v| = 1. \end{aligned}$$

This clearly defines a differentiable map. For small t this is just the exponential map. The map is one-to-one since integral curves for X can't intersect. It is onto, since r is proper, and therefore integral curves for X are defined for all time and must leave every compact set (since r is increasing along integral curves). Finally, as it is a diffeomorphism onto a neighborhood of K by the normal exponential map and the flow of a vector field always acts by local diffeomorphisms we see that it has nonsingular differential everywhere. □

2. Distance Comparison

In this section we shall introduce the main results that will make it possible to conclude that various distance functions are noncritical. This obviously requires some sort of angle comparison. The most important step in this direction is supplied by the Toponogov theorem (or the *hinge version* of Toponogov's theorem; there are triangle and angle versions as well). The proof we present is probably the simplest available; and is based upon an idea by H. Karcher (see [**28**]).

Some preparations are necessary. Let (M, g) be a Riemannian manifold. We define two very natural geometric objects:

Hinge: A *hinge* consists of two segments σ_1 and σ_2 emanating from a common point p and forming an angle α. We shall always parametrize the geodesics by arc length and assume that

$$\sigma_1(\ell(\sigma_1)) = p = \sigma_2(0).$$

The angle α is then defined as

$$\alpha = \pi - \angle(\dot{\sigma}_1(\ell(\sigma_1)), \dot{\sigma}_2(0)).$$

Thus, the first segment ends at p, while the second begins there. The angle is the *interior* angle. See also Figure 11.3.

Triangle: A *triangle* consists of three segments that meet pairwise at three different points.

In both definitions one could use geodesics. It is then possible to have degenerate triangles where some vertices coincide without the joining geodesics being

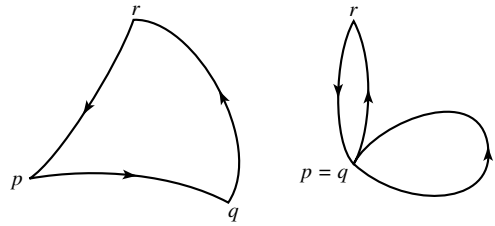

Figure 11.4

trivial. We shall use the more general hinges where σ_1 is a segment and σ_2 merely a geodesic in a few situations. In Figure 11.4 we have depicted a triangle consisting of segments, and a degenerate triangle where one of the sides is a geodesic loop and two of the vertices coincide.

Given a hinge (or a triangle), we can construct *comparison* hinges (or triangles) in the constant-curvature spaces S_k^n.

LEMMA 56. *Suppose* (M, g) *is complete and has* $\sec \geq k$. *Then for each hinge (or triangle) in* M *we can find a comparison hinge (or triangle) in* S_k^n *where the corresponding segments have the same length and the angle is the same (all corresponding segments have the same length).*

PROOF. Suppose we have three points $p, q, r \in M$. First, we know that in case $k > 0$, Myers' theorem implies

$$\text{diam} M \leq \pi/\sqrt{k} = \text{diam} S_k^n.$$

Thus, any segments between these three points have length $\leq \pi/\sqrt{k}$.

The hinge case. Here we have segments from p to q and from q to r forming an angle α at q. In the space form we can first choose $\bar{p}$ and $\bar{q}$ such that $d(\bar{p}, \bar{q}) = d(p, q)$ and then join them by a segment. This is possible because $d(p, q) \leq \pi/\sqrt{k}$. At $\bar{q}$ we can then choose a direction that forms an angle α with the chosen segment. Then we take the unique geodesic going in this direction, and using the arc length parameter we go out distance $d(q, r)$ along this geodesic. This will now be a segment, as $d(q, r) \leq \pi/\sqrt{k}$. We have then found the desired hinge.

The triangle case is similar. First, pick $\bar{p}$ and $\bar{q}$ as above. Then, consider the two distance spheres $\partial B(\bar{p}, d(p, r))$ and $\partial B(\bar{q}, d(q, r))$. Since all possible triangle inequalities between p, q, r hold and $d(q, r), d(p, r) \leq \pi/\sqrt{k}$, these distance spheres are nonempty and intersect. Then, let $\bar{r}$ be any point in the intersection.

To be honest here, we must use Cheng's diameter theorem in case any of the distances is $\pi/\sqrt{k}$. In this case there is nothing to prove as $(M, g) = S_k^n$. □

We can now state the Toponogov comparison theorem.

THEOREM 79. (Toponogov, 1959) *Let* (M, g) *be a complete Riemannian manifold with* $\sec \geq k$.

Hinge Version: *Given any hinge with vertices* $p, q, r \in M$ *forming an angle* α *at* q, *it follows, that for any comparison hinge in* S_k^n *with vertices* $\bar{p}, \bar{q}, \bar{r}$ *we have:* $d(p, r) \leq d(\bar{p}, \bar{r})$.

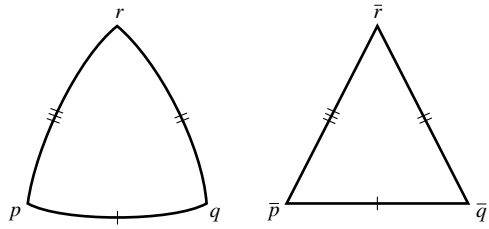

Figure 11.5

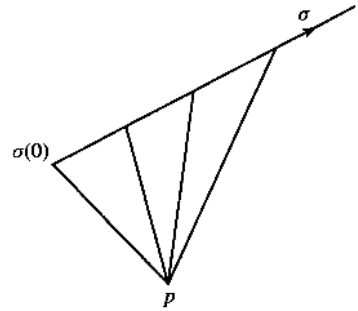

Figure 11.6

Triangle Version: *Given any triangle in M, it follows that the interior angles are larger than the corresponding interior angles for a comparison triangle in S_k^n. See also Figure 11.5*

The proof requires a little preparation. First, we claim that the hinge version implies the triangle version. This follows from the *law of cosines* in constant curvature. This law shows that if we have $p, q, r \in S_k^n$ and increase the distance $d(p, r)$ while keeping $d(p, q)$ and $d(q, r)$ fixed, then the angle at q increases as well. For simplicity, we shall only look at the cases where $k = 1, 0, -1$.

PROPOSITION 48. (Law of Cosines) *Let a triangle be given in S_k^n with side lengths a, b, c. If α denotes the angle opposite to a, then*

$$\begin{array}{ll} k = 0 & a^2 = b^2 + c^2 - 2bc\cos\alpha. \\ k = -1 & \cosh a = \cosh b \cosh c - \sinh b \sinh c \cos\alpha. \\ k = 1 & \cos a = \cos b \cos c + \sin b \sin c \cos\alpha. \end{array}$$

PROOF. The general setup is the same in all cases. Namely, we suppose that a point $p \in S_k^n$ and a unit speed segment $\sigma : [0, c] \to S_k^n$ are given. We then investigate the restriction of the distance function from p to σ. If we denote $r(x) = d(p, x)$, then we are going to study $\varphi(t) = r \circ \sigma(t)$. See also Figure 11.6

Case $k = 0$: Note that $t \to d(p, \sigma(t))$ is not a very nice function, as it is the square root of a quadratic polynomial. This, however, indicates that the function will become more manageable if we square it. Thus, we consider

$$\varphi(t) = \frac{1}{2}(r \circ \sigma(t))^2 = \frac{1}{2}|p - \sigma(t)|^2.$$

We wish to compute the first and second derivatives of this function. This requires that we know the gradient and Hessian.

$$\begin{aligned}\nabla\frac{1}{2}r^2 &= \nabla\frac{1}{2}\left(\left(x^1\right)^2+\cdots+\left(x^n\right)^2\right)\\ &= x^i\partial_i\\ &= r\nabla r;\end{aligned}$$

$$\begin{aligned}\operatorname{Hess}\frac{1}{2}r^2 &= \nabla d\left(\frac{1}{2}r^2\right)\\ &= \nabla\sum x^i dx^i\\ &= \sum dx^i dx^i\end{aligned}$$

As σ is a unit speed geodesic we get

$$\begin{aligned}\varphi'(t) &= g\left(\dot\sigma,\nabla\frac{1}{2}r^2\right),\\ \varphi''(t) &= \operatorname{Hess}\frac{1}{2}r^2(\dot\sigma,\dot\sigma)=1.\end{aligned}$$

So if we define $b=d(p,\sigma(0))$ and let α be the interior angle between σ and the line joining p with $\sigma(0)$, then we have

$$\cos(\pi-\alpha)=-\cos\alpha=g(\dot\sigma(0),\nabla r).$$

After integration of $\varphi''=1$, we get

$$\begin{aligned}\varphi(t) &= \varphi(0)+\varphi'(0)\cdot t+\frac{1}{2}t^2\\ &= \frac{1}{2}b^2-b\cdot\cos\alpha\cdot t+\frac{1}{2}t^2.\end{aligned}$$

Now set $t=c$ and define $a=d(p,\sigma(c))$, then

$$\frac{1}{2}a^2=\frac{1}{2}b^2-b\cdot c\cdot\cos\alpha+\frac{1}{2}c^2,$$

from which the law of cosines follows.

Case $k=-1$: This time we must modify the distance function in a different way. Namely, consider

$$\varphi(t)=\cosh(r\circ\sigma(t))-1.$$

Then

$$\begin{aligned}\varphi'(t) &= \sinh(r\circ\sigma(t))\,g(\nabla r,\dot\sigma),\\ \varphi''(t) &= \cosh(r\circ\sigma(t))=\varphi(t)+1.\end{aligned}$$

As before, we have $b=d(p,\sigma(0))$, and the interior angle satisfies

$$\cos(\pi-\alpha)=-\cos\alpha=g(\dot\sigma(0),\nabla d).$$

Thus, we must solve the initial value problem

$$\begin{aligned}\varphi''-\varphi &= 1,\\ \varphi(0) &= \cosh(b)-1,\\ \varphi'(0) &= -\sinh(b)\cos\alpha.\end{aligned}$$

The general solution is

$$\begin{aligned}\varphi(t) &= C_1\cosh t + C_2\sinh t - 1\\ &= (\varphi(0)+1)\cosh t + \varphi'(0)\sinh t - 1.\end{aligned}$$

So if we let $t = c$ and $a = d(p, \sigma(c))$ as before, we arrive at

$$\cosh a - 1 = \cosh b\cosh c - \sinh b\sinh c\cos\alpha - 1,$$

which implies the law of cosines again.

Case $k = 1$: This case is completely analogous to the case $k = -1$. We set

$$\varphi = 1 - \cos(r\circ\sigma(t))$$

and arrive at the initial value problem

$$\begin{aligned}\varphi'' + \varphi &= 1,\\ \varphi(0) &= 1-\cos(b),\\ \varphi'(0) &= -\sin b\cos\alpha.\end{aligned}$$

Then,

$$\begin{aligned}\varphi(t) &= C_1\cos t + C_2\sin t + 1\\ &= (\varphi(0)-1)\cos t + \varphi'(0)\sin t + 1,\end{aligned}$$

and consequently

$$1 - \cos a = -\cos b\cos c - \sin b\sin c\cos\alpha + 1,$$

which implies the law of cosines. □

The proof of the law of cosines suggests that in working in space forms it is easier to work with a modified distance function, the main advantage being that the Hessian is much simpler. Something similar can be done in variable curvature.

LEMMA 57. *Let (M, g) be a complete Riemannian manifold, $p \in M$, and $r(x) = d(x, p)$. If $\sec M \geq k$, then the Hessian of r satisfies*

$k = 0$: The function $r_0 = \frac{1}{2}r^2$ satisfies $\mathrm{Hess} r_0 \leq g$ in the support sense everywhere.

$k = -1$: The function $r_{-1} = \cosh r - 1$ satisfies $\mathrm{Hess} r_{-1} \leq (\cosh r)\, g = (r_{-1}+1)\, g$ in the support sense everywhere.

$k = 1$: The function $r_1 = 1 - \cos r$ satisfies $\mathrm{Hess} r_1 \leq (\cos r)\, g = (-r_1 + 1)\, g$ in the support sense everywhere.

PROOF. All three proofs are, of course, similar so we concentrate just on the first case. The comparison estimates from chapter 6 imply that whenever r is smooth and w is perpendicular to ∇r, then

$$\mathrm{Hess} r(w, w) \leq \frac{1}{r} g(w, w).$$

For such w one can therefore immediately see that

$$\mathrm{Hess} r_0(w, w) \leq g(w, w).$$

If instead, $w = \nabla r$, then it is trivial that this holds, whence we have established the Hessian estimate at points where r is smooth. At all other points we just use the same trick by which we obtained the Laplacian estimates with lower Ricci curvature bounds in chapter 9. □

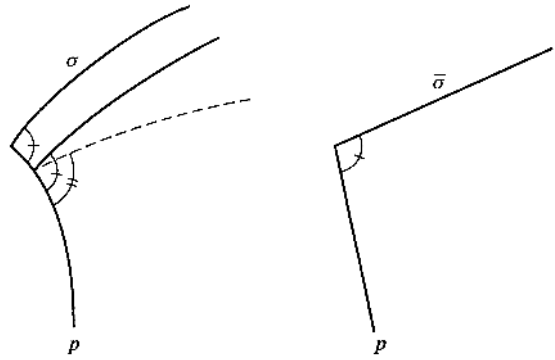

Figure 11.7

We are now ready to prove the hinge version of Toponogov's theorem. The proof is divided into the three cases: $k = 0, -1, 1$. But the setup is the same in all cases. We shall assume that a point $p \in M$ and a *geodesic* $\sigma : [0, \ell] \to M$ are given. Correspondingly, we assume that a point $\bar{p} \in S_k^n$ and segment $\bar{\sigma} : [0, \ell] \to S_k^n$ are given. Given the appropriate initial conditions, we claim that

$$d(p, \sigma(t)) \leq d(\bar{p}, \bar{\sigma}(t)).$$

We shall for simplicity assume that $d(x, p)$ is smooth at $\sigma(0)$. Then the initial conditions are

$$\begin{aligned} d(p, \sigma(0)) &\leq d(\bar{p}, \bar{\sigma}(0)), \\ g(\nabla r, \dot{\sigma}(0)) &\leq g_k\left(\nabla \bar{r}, \frac{d}{dt}\bar{\sigma}(0)\right). \end{aligned}$$

In case r is not smooth at $\sigma(0)$, we can just slide σ down along a segment joining p with $\sigma(0)$ and use a continuity argument. This also shows that we can use the stronger initial condition

$$d(p, \sigma(0)) < d(\bar{p}, \bar{\sigma}(0)).$$

In Figure 11.7 we have shown how σ can be changed by moving it down along a segment joining p and $\sigma(0)$. We have also shown how the angles can be slightly decreased. This will be important in the last part of the proof.

Proof of $k = 0$. We consider the modified functions

$$\begin{aligned} \varphi(t) &= \frac{1}{2}(r \circ \sigma(t))^2, \\ \bar{\varphi}(t) &= \frac{1}{2}(\bar{r} \circ \bar{\sigma}(t))^2. \end{aligned}$$

For small t these functions are smooth and satisfy

$$\begin{aligned} \varphi(0) &< \bar{\varphi}(0), \\ \varphi'(0) &\leq \bar{\varphi}'(0). \end{aligned}$$

Moreover, for the second derivatives we have

$$\begin{aligned} \varphi'' &\leq 1 \text{ in the support sense,} \\ \bar{\varphi}'' &= 1, \end{aligned}$$

whence the difference $\psi(t) = \bar{\varphi}(t) - \varphi(t)$ satisfies

$$\begin{aligned} \psi(0) &> 0, \\ \psi'(0) &\geq 0, \\ \psi''(t) &\geq 0 \text{ in the support sense.} \end{aligned}$$

This shows that ψ is a convex function that is positive and increasing for small t, and hence increasing, and in particular positive, for all t. This proves the hinge version. □

Proof of $k = -1$. Consider

$$\begin{aligned} \varphi(t) &= \cosh r \circ \sigma(t) - 1, \\ \bar{\varphi}(t) &= \cosh \bar{r} \circ \bar{\sigma}(t) - 1. \end{aligned}$$

Then

$$\begin{aligned} \varphi(0) &< \bar{\varphi}(0), \\ \varphi'(0) &\leq \bar{\varphi}'(0), \\ \varphi'' &\leq \varphi + 1 \text{ in the support sense,} \\ \bar{\varphi}'' &= \bar{\varphi} + 1. \end{aligned}$$

Then the difference $\psi = \bar{\varphi} - \varphi$ satisfies

$$\begin{aligned} \psi(0) &> 0, \\ \psi'(0) &\geq 0, \\ \psi''(t) &\geq \psi(t) \text{ in the support sense.} \end{aligned}$$

The first condition again implies that ψ is positive for small t. The last condition shows that as long as ψ is positive, it is also convex. The second condition then shows that ψ is increasing to begin with. It must now follow that ψ keeps increasing. Otherwise, there would be a positive maximum, and that violates convexity at points where ψ is positive. □

Proof of $k = 1$. Case $k = 1$: This case is considerably harder. We begin as before by defining

$$\begin{aligned} \varphi(t) &= 1 - \cos(r \circ \sigma(t)), \\ \bar{\varphi}(t) &= 1 - \cos(\bar{r} \circ \bar{\sigma}(t)) \end{aligned}$$

and then observing that the difference $\psi = \bar{\varphi} - \varphi$ satisfies

$$\begin{aligned} \psi(0) &> 0, \\ \psi'(0) &\geq 0, \\ \psi''(t) &\geq -\psi(t) \text{ in the support sense.} \end{aligned}$$

That, however, doesn't look very promising. Even though the function starts out being positive, the last condition only gives a *negative* lower bound for the second derivative. At this point some people might recall that perhaps Sturm-Liouville theory could save us. But for that to work well it is best to assume $\psi'(0) > 0$. Thus, another little continuity argument is necessary as we need to perturb σ again to decrease the interior angle. If the interior angle is positive, this can clearly be

done, and in the case where this angle is zero the hinge version is trivially true anyway. Now define $\zeta(t)$ by

$$\begin{aligned} \zeta'' &= -(1+\varepsilon)\zeta, \\ \zeta(0) &= \psi(0) = \alpha > 0, \\ \zeta'(0) &= \psi'(0) = \beta > 0, \end{aligned}$$

this means that

$$\zeta(t) = \sqrt{\alpha^2 + \frac{\beta^2}{1+\varepsilon}} \cdot \sin\left(\sqrt{1+\varepsilon}\cdot t + \arctan\left(\frac{\alpha\cdot\sqrt{1+\varepsilon}}{\beta}\right)\right).$$

For small t we have

$$\begin{aligned} (\psi(t) - \zeta(t))'' &\geq -\psi(t) + (1+\varepsilon)\zeta(t) \\ &= \zeta(t) - \psi(t) + \varepsilon\zeta(t) \\ &> 0. \end{aligned}$$

Thus we have $\psi(t) - \zeta(t) \geq 0$ for small t. We now wish to extend this to the interval where $\zeta(t)$ is positive, i.e., for

$$t < \frac{\pi - \arctan\left(\frac{\alpha\cdot\sqrt{1+\varepsilon}}{\beta}\right)}{\sqrt{1+\varepsilon}},$$

To get this to work, consider the quotient

$$h = \frac{\psi}{\zeta}.$$

So far, we know that this function satisfies

$$\begin{aligned} h(0) &= 1, \\ h(t) &\geq 1 \text{ for small } t. \end{aligned}$$

Should it therefore dip below 1 before reaching the end of the interval, then h would have a positive local maximum at some t_0. At this point we can use support functions ψ_δ for ψ from below, and conclude that also $\frac{\psi_\delta}{\zeta}$ has a local maximum at t_0. Thus, we have

$$\begin{aligned} 0 &\geq \frac{d^2}{dt^2}\left(\frac{\psi_\delta}{\zeta}\right)(t_0) \\ &= \frac{\psi_\delta''(t_0)}{\zeta(t_0)} - 2\frac{\zeta'(t_0)}{\zeta(t_0)}\cdot\frac{d}{dt}\left(\frac{\psi_\delta}{\zeta}\right)_{t=t_0} - \frac{\psi_\delta(t_0)}{\zeta^2(t_0)}\zeta''(t_0) \\ &\geq \frac{-\psi_\delta(t_0) - \delta}{\zeta(t_0)} + \frac{\psi_\delta(t_0)}{\zeta(t_0)}(1+\varepsilon) \\ &= \frac{\varepsilon\cdot\psi_\delta(t_0) - \delta}{\zeta(t_0)}. \end{aligned}$$

But this becomes positive as $\delta \to 0$, since we assumed $\psi_\delta(t_0) > 0$, and so we have a contradiction. Next, we can let $\varepsilon \to 0$ and finally, let $\alpha \to 0$ to get the desired estimate for all $t \leq \pi$ using continuity. □

Note that we never really use in the proof that we work with segments. The only thing that must hold is that the geodesics in the space form are segments. For $k \leq 0$ this is of course always true, but when $k = 1$ this means that the geodesic must have length $\leq \pi$. This was precisely the important condition in the last part of the proof.

3. Sphere Theorems

Our first applications of the Toponogov theorem are to the case of positively curved manifolds. Using scaling, we shall assume throughout this section that we work with a closed Riemannian n-manifold (M, g) with $\sec \geq 1$. For such spaces we have established

(1) $\operatorname{diam}(M, g) \leq \pi$, with equality holding only if $M = S^n(1)$.
(2) If n is odd, then M is orientable.
(3) If n is even and M is orientable, then M is simply connected and $\operatorname{inj}(M) \geq \pi/\sqrt{\max \sec}$.
(4) If n is even and $\max \sec$ is close to 1, then (M, g) is close to a constant curvature metric. In particular, M must be a sphere when it is simply connected.
(5) It has also been mentioned that Klingenberg has shown that if M is simply connected and $\max \sec < 4$, then $\operatorname{inj}(M) \geq \pi/\sqrt{\max \sec}$.
(6) If M is simply connected and $\max \sec < 4$, then M is homotopy equivalent to a sphere.

The penultimate result is quite subtle and is beyond what we can prove here. Gromov (see [**36**]) has a proof of this that in spirit goes as follows: One considers $p \in M$. If the upper curvature bound is $4 - \delta$, then we know that if we pull the metric back to the tangent bundle, then there are no conjugate points on the disc $B\left(0, \pi/\sqrt{4-\delta}\right)$. Consider the modified distance r_1 to the origin in T_pM. This function is smooth on $B\left(0, \pi/\sqrt{4-\delta}\right)$ and satisfies

$$\operatorname{Hess} r_1 \leq (1 - r_1)\, g = (\cos r)\, g.$$

On the region

$$B\left(0, \frac{\pi}{\sqrt{4-\delta}}\right) - \bar{B}(0, \pi/2)$$

this function will therefore have strictly negative Hessian. In particular, the level sets for r or r_1 that lie in that region are strictly concave. Now map these level sets down into M via the exponential map. As this map is nonsingular they will be mapped to strictly concave, possibly immersed, hypersurfaces in M. In the case where M is simply connected, one can prove an analogue to the Hadamard theorem for immersed convex hypersurfaces, namely, that they must be embedded spheres (this also uses that M has nonnegative curvature). However, if these hypersurfaces are embedded, then the exponential map must be an embedding on $B\left(0, \pi/\sqrt{4-\delta}\right)$, and in particular, we obtain the desired injectivity radius estimate.

We can now prove the celebrated Rauch-Berger-Klingenberg sphere theorem, also known as the quarter pinched sphere theorem. Note that the conclusion is stronger than Berger's result mentioned in chapter 6. The part of the proof presented below is also due the Berger.

THEOREM 80. (1951-1961) *If M is a simply connected closed Riemannian manifold with $1 \leq \sec \leq 4-\delta$, then M is homeomorphic to a sphere.*

PROOF. We gave a different proof of this in chapter 6 that used index estimation.

We have shown that the injectivity radius is $\geq \pi/\sqrt{4-\delta}$. Thus, we have large discs around every point in M. Now select two points $p, q \in M$ such that $d(p,q) = \mathrm{diam}M$. Note that

$$\mathrm{diam}M \geq \mathrm{inj}M > \frac{\pi}{2}.$$

We now claim that every point $x \in M$ lies in one of the two balls $B\left(p, \pi/\sqrt{4-\delta}\right)$, or $B\left(q, \pi/\sqrt{4-\delta}\right)$, and thus M is covered by two discs. This certainly makes M look like a sphere as it is the union of two discs. Below we shall construct an explicit homeomorphism to the sphere in a more general setting.

Now take $x \in M$. Let $d = \mathrm{diam}M = d(p,q)$, $a = d(p,x)$, and $b = d(x,q)$. If, for instance, $b > \pi/2$, then we claim that $a < \pi/2$. First, observe that since q is at maximal distance from p, it must follow that q cannot be a regular point for the distance function to p. Therefore, if we select any segment σ_1 from x to q, then we can find a segment σ_2 from p to q that forms an angle $\alpha \leq \pi/2$ with σ_1 at q. Then we can consider the hinge σ_1, σ_2 with angle α. The hinge version of Toponogov's theorem implies

$$\begin{aligned} \cos a &\geq \cos b \cos d + \sin b \sin d \cos \alpha \\ &\geq \cos b \cos d. \end{aligned}$$

Now, both $b, d > \pi/2$, so the left hand side is positive. This implies that $a < \pi/2$, as desired. □

Recall from the last chapter that Micaleff and Moore proved a similar theorem for manifolds that only have positive isotropic curvature.

Note that the theorem does not say anything about the non-simply connected situation. Thus we cannot conclude that such spaces are homeomorphic to spaces of constant curvature. Only that the universal covering is a sphere.

The above proof suggests, perhaps, that the conclusion of the theorem should hold as long as the manifold has large diameter. This is the content of the next theorem. This theorem was first proved by Berger for simply connected manifolds with a different proof and a slightly weaker conclusion. The present version is known as the Grove-Shiohama diameter sphere theorem. It was for the purpose of proving this theorem that Grove and Shiohama introduced critical point theory.

THEOREM 81. (Berger, 1962 and Grove-Shiohama, 1977) *If (M,g) is a closed Riemannian manifold with* $\sec \geq 1$ *and* $\mathrm{diam} > \pi/2$, *then M is homeomorphic to a sphere.*

PROOF. We first give Berger's index estimation proof that follows his index proof of the quarter pinched sphere theorem. The goal is to find $p \in M$ such that all geodesic loops at p have length $> \pi$. The proof from chapter 6, then carries over verbatim. To this end select $p, q \in M$ such that

$$d(p,q) = \mathrm{diam}M = d > \pi/2.$$

We claim that p has the desired property. Supposing otherwise we get a geodesic loop $\gamma : [0,1] \to M$ based at p of length $\leq \pi$. As p is at maximal distance from q,

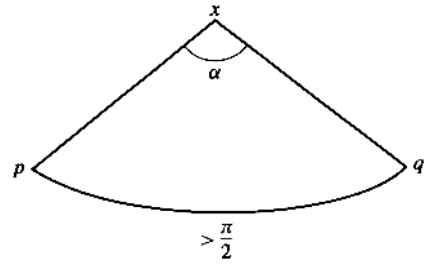

Figure 11.8

we can find a segment σ from q to p such that the hinge spanned by σ and γ has angle $\leq \pi/2$. While γ is not a segment it is sufficiently short that the hinge version of Toponogov's theorem still holds. Thus we must have

$$\begin{aligned} 0 &> \cos(d(p,q)) \\ &= \cos(d(\sigma(0),\gamma(1))) \\ &\geq \cos(d(\sigma(0),\sigma(1)))\cos(\ell(\gamma)) \\ &= \cos(d(p,q))\cos(\ell(\gamma)). \end{aligned}$$

This is clearly not possible unless $\ell(\gamma) = 0$.

We now give the Grove-Shiohama proof. Fix $p, q \in M$ with

$$d(p,q) = \mathrm{diam} M = d > \pi/2.$$

The claim is that the distance function from p has only q as a critical point. To see this, let $x \in M - \{p,q\}$ and let α be the angle between any two geodesics from x to p and q. If we suppose that $\alpha \leq \pi/2$ and set $b = d(p,x)$ and $c = d(x,q)$, then the hinge version of Toponogov's theorem implies

$$\begin{aligned} 0 &> \cos d \geq \cos b \cos c + \sin b \sin c \cos \alpha \\ &\geq \cos b \cos c. \end{aligned}$$

But then $\cos b$ and $\cos c$ have opposite signs. If, for example, $\cos b \in (0,1)$, then we have $\cos d > \cos c$, which implies $c > d = \mathrm{diam} M$. Thus we have arrived at a contradiction, and hence we must have $\alpha > \pi/2$. See also Figure 11.8

We can now construct a vector field X that is the gradient field for $x \to d(x,p)$ near p and the negative of the gradient field for $x \to d(x,q)$ near q. Furthermore, the distance to p increases along integral curves for X. For each $x \in M - \{p,q\}$ there is a unique integral curve $\gamma_x(t)$ for X through x.Suppose that x varies over a small distance sphere $\partial B(p,\varepsilon)$ that is diffeomorphic to S^{n-1}. After time t_x this integral curve will hit the distance sphere $\partial B(q,\varepsilon)$ which can also be assumed to be diffeomorphic to S^{n-1}. The function $x \to t_x$ is continuous and in fact smooth as both distance spheres are smooth submanifolds. Thus we have a diffeomorphism defined by

$$\begin{aligned} \partial B(p,\varepsilon) \times [0,1] &\to M - (B(p,\varepsilon) \cup B(q,\varepsilon)), \\ (x,t) &\to \gamma_x(t \cdot t_x) \end{aligned}$$

Gluing this map together with the two discs $B(p,\varepsilon)$ and $B(q,\varepsilon)$ then yields a continuous bijection $M \to S^n$. Note that the construction does not guarantee smoothness of this map on $\partial B(p,\varepsilon)$ and $\partial B(q,\varepsilon)$. □

Aside from the fact that the conclusions in the above theorems could possibly be strengthened to diffeomorphism, we have optimal results. Complex projective space has curvatures in $[1,4]$ and diameter $\pi/2$ and the real projective space has constant curvature 1 and diameter $\pi/2$. If one relaxes the conditions slightly, it is, however, still possible to say something.

THEOREM 82. *Suppose (M,g) is simply connected of dimension n with $1 \leq \sec \leq 4+\varepsilon$.*

(1) (Berger, 1983) *If n is even, then there is $\varepsilon(n) > 0$ such that M must be homeomorphic to a sphere or diffeomorphic to one of the spaces $\mathbb{C}P^{n/2}$, $\mathbb{H}P^{n/4}$, $\mathbb{O}P^2$.*

(2) (Abresch-Meyer, 1994) *If n is odd, then there is an $\varepsilon > 0$, which can be chosen independently of n, such that M is homeomorphic to a sphere.*

The spaces $\mathbb{C}P^{n/2}$, $\mathbb{H}P^{n/4}$, or $\mathbb{O}P^2$ are known as the compact rank 1 symmetric spaces (CROSS). The complex projective space has already been studied in chapters 3 and 8. The quaternionic projective space is $\mathbb{H}P^n = S^{4n+3}/S^3$, but the octonion plane is a bit more exotic: $F_4/Spin(9) = \mathbb{O}P^2$ (see also chapter 8 for more on these spaces). The proof of (1) uses convergence theory. First, it is shown that if $\varepsilon = 0$, then M is either homeomorphic to a sphere or isometric to one of the CROSSs. Then using the injectivity radius estimate in even dimensions, we can apply the convergence machinery.

For the diameter situation we have

THEOREM 83. (Grove-Gromoll, 1987 and Wilking, 2001) *Suppose (M,g) is closed and satisfies* $\sec \geq 1$, $\operatorname{diam} \geq \frac{\pi}{2}$ *Then one of the following cases holds:*

(1) M is homeomorphic to a sphere.

(2) M is isometric to a finite quotient $S^n(1)/\Gamma$, where the action of Γ is reducible (has an invariant subspace).

(3) M is isometric to one of $\mathbb{C}P^{n/2}$, $\mathbb{H}P^{n/4}$, $\mathbb{C}P^{n/2}/\mathbb{Z}_2$ for $n = 2 \bmod 4$.

(4) M is isometric to $\mathbb{O}P^2$.

Grove and Gromoll settled all but part (4), where they only showed that M had to have the cohomology ring of $\mathbb{O}P^2$. It was Wilking who finally settled this last case (see [**94**]).

4. The Soul Theorem

Let us commence by stating the theorem we are aiming to prove and then slowly work our way through the rather intricate and technical proof.

THEOREM 84. (Cheeger-Gromoll-Meyer, 1969, 1972) *If (M,g) is a complete non-compact Riemannian manifold with* $\sec \geq 0$, *then M contains a* soul $S \subset M$, *which is a closed totally convex submanifold, such that M is diffeomorphic to the normal bundle over S. Moreover, when* $\sec > 0$, *the soul is a point and M is diffeomorphic to $\mathbb{R}^n$.*

The history is briefly that Gromoll-Meyer first showed that if $\sec > 0$, then M is diffeomorphic to $\mathbb{R}^n$. Soon after Cheeger-Gromoll established the full theorem. The Gromoll-Meyer theorem is in itself rather remarkable.

We shall use critical point theory to establish this theorem. The problem lies in finding the soul. When this is done, it will be easy to see that the distance function

to the soul has only regular points, and then we can use the results from the first section.

Before embarking on the proof, it might be instructive to show the following less ambitious result, whose proof will be used in the next section.

LEMMA 58. (Gromov's critical point estimate, 1981) *If (M,g) is a complete open manifold of nonnegative sectional curvature, then for every $p \in M$ the distance function $d(\cdot,p)$ has no critical points outside some ball $B(p,R)$. In particular, M must have the topology of a compact manifold with boundary.*

PROOF. We shall use a contradiction argument. So suppose we have a sequence p_k of critical points for $d(\cdot,p)$, where $d(p_k,p) \to \infty$. After passing to a subsequence we can without loss of generality assume that

$$d(p_{k+1},p) \geq 2d(p_k,p).$$

Now select segments σ_k from p to p_k. The above inequality implies that the angle at p between any two segments is $\geq 1/6$. To see this, suppose σ_k and σ_{k+l} form an angle $< 1/6$ at p. The hinge version of Toponogov's theorem then implies

$$\begin{aligned}(d(p_k,p_{k+l}))^2 &< (d(p,p_{k+l}))^2 + (d(p_k,p))^2 - 2d(p,p_{k+l})\,d(p_k,p)\cos\frac{1}{6}\\ &\leq \left(d(p,p_{k+l}) - \frac{3}{4}d(p_k,p)\right)^2.\end{aligned}$$

Now use that p_k is critical for p to conclude that there are segments from p to p_k and p_{k+l} to p_k that from an angle $\leq \pi/2$ at p_k. Then use the hinge version again to conclude

$$\begin{aligned}(d(p,p_{k+l}))^2 &\leq (d(p_k,p))^2 + (d(p_k,p_{k+l}))^2\\ &\leq (d(p_k,p))^2 + \left(d(p,p_{k+l}) - \frac{3}{4}d(p_k,p)\right)^2\\ &= \frac{25}{16}(d(p_k,p))^2 + (d(p,p_{k+l}))^2 - \frac{3}{2}d(p,p_{k+l})\,d(p_k,p),\end{aligned}$$

which implies

$$d(p,p_{k+l}) \leq \frac{25}{24}d(p_k,p).$$

But this contradicts our assumption that

$$d(p,p_{k+l}) \geq d(p_{k+1},p) \geq 2d(p_k,p).$$

Now that all the unit vectors $\dot\sigma_k(0)$ form angles of at least $1/6$ with each other, we can conclude that there can't be infinitely many such vectors. Hence, there cannot be critical points infinitely far away from p.

Observe that the vectors $\dot\sigma_k(0)$ lie on the unit sphere in T_pM and are distance $1/6$ away form each other. Thus, the balls $B(\dot\sigma_k(0), 1/12)$ are disjoint in the unit sphere and hence there are at most

$$\frac{v(n-1,1,\pi)}{v\left(n-1,1,\frac{1}{12}\right)} \leq 100^n$$

such points. □

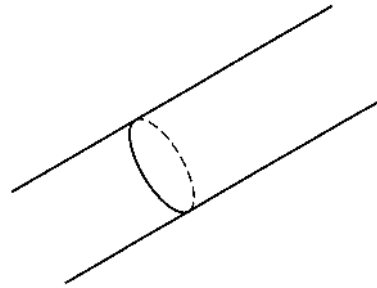

Figure 11.9

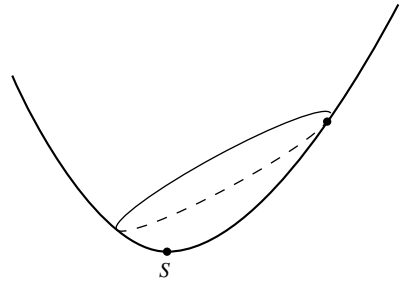

Figure 11.10

We now have to explain what it means for a submanifold, or more generally a subset, to be totally convex. The notion is similar to being totally geodesic. A subset $A \subset M$ of a Riemannian manifold is said to be *totally convex* if any geodesic in M joining two points in A actually lies in A. There are in fact several different kinds of convexity, but as they are not important for any other developments here, we shall confine ourselves to total convexity. The first observation is that this definition agrees with the usual definition for convexity in Euclidean space. Other than that, it is not clear that any totally convex sets exist at all. For example, if $A = \{p\}$, then A is totally convex only if there are no geodesic loops based at p. This means that points will almost never be totally convex. In fact, if M is closed, then M is the only totally convex subset. This is not completely trivial, but using the energy functional as in chapter 6 we note that if $A \subset M$ is totally convex, then $A \subset M$ is k-connected for any k. It is however, not possible for a closed n-manifold to have n-connected nontrivial subsets as this would violate Poincaré duality. On complete manifolds it is sometimes possible to find totally convex sets.

EXAMPLE 59. *Let (M, g) be the flat cylinder $\mathbb{R} \times S^1$. All of the circles $\{p\} \times S^1$ are geodesics and totally convex. This also means that no point in M can be totally convex. In fact, all of those circles are souls. See also Figure 11.9*

EXAMPLE 60. *Let (M, g) be a smooth rotationally symmetric metric on $\mathbb{R}^2$ of the form $dr^2 + \varphi^2(r)\, d\theta^2$, where $\varphi'' < 0$. Thus, (M, g) looks like a parabola of revolution. The radial symmetry implies that all geodesics emanating from the origin $r = 0$ are rays going to infinity. Thus the origin is a soul and totally convex. Most other points, however, will have geodesic loops based there. See also Figure 11.10.*

The way to find totally convex sets is via

LEMMA 59. *If $f : (M, g) \to \mathbb{R}$ is concave, in the sense that the Hessian is weakly nonpositive everywhere, then every superlevel set $A = \{x \in M : f(x) \geq a\}$ is totally convex.*

PROOF. Given a geodesic γ in M, we have that the function $f \circ \gamma$ has nonpositive weak second derivative. Thus, $f \circ \gamma$ is concave as a function on $\mathbb{R}$. In particular, the minimum of this function on any compact interval is obtained at one of the endpoints. This finishes the proof. □

We are now left with the problem of the existence of proper concave functions on complete manifolds with nonnegative sectional curvature.

LEMMA 60. *Suppose (M, g) is as in the theorem and that $p \in M$. If we take all rays $\{\gamma_\alpha\}$ emanating from p and construct*

$$f = \inf_\alpha b_{\gamma_\alpha},$$

where b_γ denotes the Busemann function, then f is both proper and concave.

PROOF. First we show that in nonnegative sectional curvature all Busemann functions are concave. Using that, we can then show that the given function is concave and proper.

Recall that in nonnegative Ricci curvature Busemann functions are superharmonic. The proof of concavity is almost identical. Instead of the Laplacian estimate for distance functions, we must use a similar Hessian estimate. If $r = d(\cdot, p)$, then we know that Hess$r$ vanishes on radial directions $\partial_r = \nabla r$ and satisfies

$$\text{Hess} r \leq \frac{1}{r} g$$

on vectors perpendicular to the radial direction. In particular, Hess$r \leq \frac{1}{r} g$ at all smooth points. We can then extend this estimate to the points where r isn't smooth as we did for modified distance functions. We can now proceed as in the Ricci curvature case to show that Busemann functions have nonpositive Hessians in the weak sense and are therefore concave.

The infimum of a collection of concave functions is clearly concave. So we must now show that the superlevel sets for f are compact. Suppose, on the contrary, that some superlevel set $A = \{x \in M : f(x) \geq a\}$ is noncompact. If $a > 0$, then also $A = \{x \in M : f(x) \geq 0\}$ is noncompact. So we can assume that $a \leq 0$. As all of the Busemann functions b_{γ_α} are zero at p also $f(p) = 0$. In particular, $p \in A$. Using noncompactness select a sequence $p_n \in A$ that goes to infinity. Then join p_n to p by a segment, and as in the construction of rays, choose a subsequence of these segments converging to a ray emanating from p. As A is totally convex, all of these segments lie in A. Since A is closed the ray must also lie in A and therefore be one of the rays γ_α. But

$$f(\gamma_\alpha(t)) \leq b_{\gamma_\alpha}(\gamma_\alpha(t)) = -t \to -\infty,$$

so we have a contradiction. □

We now need to establish a few fundamental properties of totally convex sets.

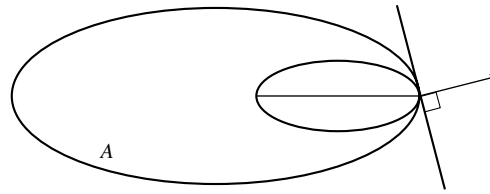

Figure 11.11

LEMMA 61. *If $A \subset (M, g)$ is totally convex, then A has an interior, denoted by* int$A$, *and a boundary ∂A. The interior is a totally convex submanifold of M, and the boundary has the property that for each $x \in \partial A$ there is an inward pointing vector $w \in T_x M$ with the property: If $\gamma(t) : [0, a] \to A$ is a geodesic with $\gamma(0) = x$ and $\gamma(a) \in$ int$A$, then $\angle(w, \dot{\gamma}(0)) < \frac{\pi}{2}$.*

Some comments are in order before the proof. The words *interior* and *boundary*, while describing fairly accurately what the sets look like, are not meant in the topological sense. Most convex sets will, of course, not have any topological interior at all. The property about the boundary is what is often called the *supporting hyperplane property*. Namely, the interior of the convex set is supposed to lie on one side of a hyperplane at any of the boundary points. The vector w is the normal to this hyperplane and can be taken to be tangent to some geodesic that goes into the interior. It is important to note that the supporting hyperplane property shows that the distance function to a subset of int$A$ cannot have any critical points on ∂A. See also Figure 11.11.

PROOF. The convexity radius estimate from chapter 6 will be used in many places. Specifically we shall use that there is a positive function $\varepsilon(p) : M \to (0, \infty)$ such that the distance function $r_p(x) = d(x, p)$ is smooth and strictly convex on $B(p, \varepsilon(p))$.

First, let us identify points in the interior and on the boundary. To make the identifications simpler we assume that A is closed.

Find the maximal integer k such that A contains a k-dimensional submanifold of M. If $k = 0$, then A must be a point. For if A contains two points, then A also contains a segment joining these points and therefore a 1-dimensional submanifold. Now define $N \subset A$ as being the union of all k-dimensional submanifolds in M that are contained in A. We claim that N is a k-dimensional totally convex submanifold whose closure is A. We shall thus identify int$A$ with N and ∂A with $A - N$.

To see that it is a submanifold, pick $p \in N$ and let $N_p \subset A$ be a k-dimensional submanifold of M containing p. By shrinking N_p if necessary, we also assume that it is embedded. We can therefore find $\delta \in (0, \varepsilon(p))$ so that $B(p, \delta) \cap N_p = N_p$. We now claim that also $B(p, \delta) \cap A = N_p$. If this were not true, then we could find

$$q \in A \cap B(p, \delta) - N_p.$$

Now assume that δ is so small that also $\delta < \text{inj}_q$. Then we can join each point in $B(p, \delta) \cap N_p$ to q by a unique segment. The union of these segments will, away from q, form a cone that is a $(k+1)$-dimensional submanifold which is contained in A (see Figure 11.12), thus contradicting maximality of k. In particular, N must be an embedded submanifold as we have $B(p, \delta) \cap N = N_p$.

What we have just proved can easily be modified to show that for points $p \in N$ and $q \in A$ with the property that $d(p,q) < \text{inj}_q$ there is a k-dimensional submanifold $N_p \subset N$ such that $q \in \bar{N}_p$, namely, just take a $(k-1)$-dimensional submanifold through p in N perpendicular to the segment from p to q and consider the cone over this submanifold with vertex q. From this statement we get the property that if $\gamma : [0,a] \to A$ is a geodesic, then $\gamma(0,a) \subset N$ provided that, say, $\gamma(0) \in N$. In particular, N is dense in A.

Having identified the interior and boundary, we now have to establish the supporting hyperplane property. First we note that since N is totally geodesic its tangent spaces T_qN are preserved by parallel translation along curves in N. For $p \in \partial A$ we therefore have a well-defined k-dimensional tangent space $T_pA \subset T_pM$ coming from parallel translating the tangent spaces to N along curves in N that end at p. Next define the tangent cone at $p \in \partial A$

$$C_pA = \left\{ v \in T_pM : \exp_p(tv) \in N \text{ for some } t > 0 \right\}.$$

Note that in fact $\exp_p(tv) \in N$ for all small $t > 0$. This shows that C_pA is a cone. Clearly $C_pA \subset T_pA$ and in fact spans it as we can easily find k linearly independent vectors in C_pA. Finally, we see that C_pA is an open subset of T_pA.

For $\varepsilon > 0$ small, suppose we can select

$$q \in A_\varepsilon = \{x \in A : d(x, \partial A) \geq \varepsilon\}$$

such that $d(q,p) = \varepsilon$. The set of such points is clearly 2ε-dense in ∂A. So the set of points $p \in \partial A$ for which we can find an $\varepsilon > 0$ and $q \in A_\varepsilon$ such that $d(q,p) = \varepsilon$ is dense in ∂A. As the supporting plane property is an open property (this follows from critical point theory), it suffices to prove it for such p. We can also suppose ε is so small that $r_q = d(\cdot, q)$ is smooth and convex on a neighborhood containing p. The claim is that $\angle(-\nabla r_q, v) < \frac{\pi}{2}$ for all $v \in C_pA$. To see this, observe that we have a convex set

$$A' = A \cap \bar{B}(q, \varepsilon),$$

with interior

$$N' = A \cap B(q, \varepsilon) \subset N$$

and $p \in \partial A'$. Thus $C_pA' \subset C_pA$. In addition we see that $T_pA = T_pA'$. The tangent cone of $\bar{B}(q,\varepsilon)$ is given by

$$C_p\bar{B}(q,\varepsilon) = \left\{ v \in T_pM : \angle(v, -\nabla r_q) < \frac{\pi}{2} \right\}$$

as r is smooth at p, thus

$$C_pA' = \left\{ v \in T_pA : \angle(v, -\nabla r_q) < \frac{\pi}{2} \right\}$$

If now

$$C_pA' \subsetneqq C_pA,$$

then openness of C_pA in T_pA implies that there must be a $v \in C_pA$ such that also $-v \in C_pA$. But this implies that $p \in N$, as it becomes a point on a geodesic whose endpoints lie in N. (See Figure 11.13.) □

The last lemma we need is

LEMMA 62. *Let (M,g) have $\sec \geq 0$. If $A \subset M$ is totally convex, then the distance function $r : A \to \mathbb{R}$ defined by $r(x) = d(x, \partial A)$ is concave on A, and strictly concave if $\sec > 0$.*

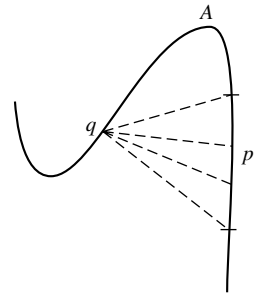

Figure 11.12

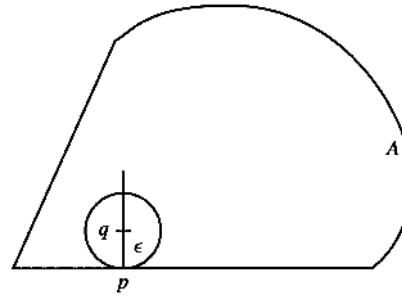

Figure 11.13

PROOF. We shall show that the Hessian is nonpositive in the support sense. Fix $q \in \text{int}A$, and find $p \in \partial A$ so that $d(p,q) = d(q, \partial A)$. Then select $\sigma : [0,a] \to A$ to be a segment from p to q. Using exponential coordinates at p we create a hypersurface H which is the image of the hyperplane perpendicular to $\dot{\sigma}(0)$. This hypersurface is perpendicular to $\dot{\sigma}(0)$, the second fundamental form for H at p is zero, and $H \cap \text{int}A = \emptyset$. (See Figure 11.14.) We have that $f(x) = d(x,H)$ is a support function from above for $d(\cdot, \partial A)$ at $\sigma(t)$ for all $t \in [0,a]$. Moreover f is smooth at $\sigma(t)$ for all $t < a$.

We start by showing that the support function f is concave at $\sigma(t)$ as long as f is smooth at $\sigma(t)$. Note that σ is an integral curve for ∇f. Evaluating the fundamental equation on a parallel field along σ that starts out being tangent to H, i.e., perpendicular to σ, therefore yields:

$$\begin{aligned} \frac{d}{dt}\text{Hess}f(E,E) &= -g(R(E,\dot{\sigma})\dot{\sigma},E) - \text{Hess}^2 f(E,E) \\ &\leq 0 . \end{aligned}$$

Since $\text{Hess}f(E,E) = 0$ at $t = 0$ we see that $\text{Hess}f(E,E) \leq 0$ along σ (and < 0 if $\sec > 0$). This shows that we have a smooth support function for $d(\cdot, \partial A)$ on an open and dense subset in A.

If f is not smooth at $\sigma(a)$ we can for $t < a$ find a hypersurface H_t as above that is perpendicular to $\dot{\sigma}(t)$ at $\sigma(t)$ and has vanishing second fundamental form at $\sigma(t)$. For t close to a we have that $f_t = d(\cdot, H_t)$ is smooth at q and therefore also has nonpositive (negative) Hessian at q. In this case we claim that $t + f_t$ is a support function for $d(\cdot, \partial A)$. Clearly, the functions are equal at q. If x is close to

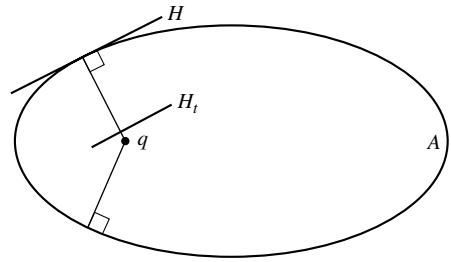

Figure 11.14

q, then we can select $z \in H_t$ so that

$$d(x, \partial A) = d(x, z) + f_t(z).$$

Thus we are reduced to showing that $d(z, \partial A) \leq t$ for each $z \in H_t$.

As f is smooth at $\sigma(t)$ it follows that $d(\cdot, \partial A)$ is concave in a neighborhood of $\sigma(t)$. Now select a short geodesic $\gamma(s)$ from $z \in H_t$ to $\sigma(t)$. By the construction of H_t we can assume that this geodesic is contained in H_t and therefore perpendicular to $\sigma(t)$. Concavity of $s \to d(\gamma(s), \partial A)$ then shows that

$$d(\gamma(s), \partial A) \leq d(\gamma(0), \partial A) = t.$$

This establishes our claim. □

We are now ready to prove the soul theorem. Start with the proper concave function f constructed from the Busemann functions. The maximum level set

$$C_1 = \{x \in M : f(x) = \max f\}$$

is nonempty and convex since f is proper and concave. Moreover, it follows from the previous lemma that C_1 is a point if $\sec > 0$. This is because the superlevel sets

$$A = \{x \in M : f(x) \geq a\}$$

are convex with $\partial A = f^{-1}(a)$, so $f = d(\cdot, \partial A)$ on A. Now, a strictly concave function (Hessian in support sense is negative) must have a unique maximum or no maximum, thus showing that C_1 is a point. If C_1 is a submanifold, then we are also done. In this case $d(\cdot, C_1)$ has no critical points, as any point lies on the boundary of a convex superlevel set. Otherwise, C_1 is a convex set with nonempty boundary. But then $d(\cdot, \partial C_1)$ is concave. The maximum set C_2 is again nonempty, since C_1 is compact and convex. If it is a submanifold, then we again claim that we are done. For the distance function $d(\cdot, C_2)$ has no critical points, as any point lies on the boundary for a superlevel set for either f or $d(\cdot, \partial C_1)$. We can now iterate to get a sequence of convex sets

$$C_1 \supset C_2 \supset \cdots \supset C_k.$$

We claim that in at most $n = \dim M$ steps we arrive at a point or submanifold S, that we call the soul (see Figure 11.15). This is because $\dim C_i > \dim C_{i+1}$. To see this suppose $\dim C_i = \dim C_{i+1}$, then $\operatorname{int} C_{i+1}$ will be an open subset of $\operatorname{int} C_i$. So if $p \in \operatorname{int} C_{i+1}$, then we can find δ such that

$$B(p, \delta) \cap \operatorname{int} C_{i+1} = B(p, \delta) \cap \operatorname{int} C_i.$$

Now choose a segment σ from p to ∂C_i. Clearly $d(\cdot, \partial C_i)$ is strictly increasing along this curve. This curve, however, runs through $B(p, \delta) \cap \operatorname{int} C_i$, thus showing that $d(\cdot, \partial C_i)$ must be constant on the part of the curve close to p.

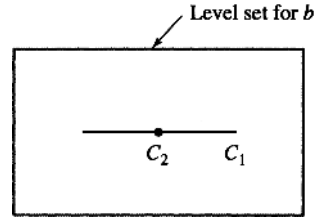

Figure 11.15

Much more can be said about complete manifolds with nonnegative sectional curvature. A rather complete account can be found in Greene's survey in [**50**]. We briefly mention two important results:

THEOREM 85. *Let S be a soul of a complete Riemannian manifold with* $\sec \geq 0$, *arriving from the above construction.*

(1) (Sharafudtinov, 1978) *There is a distance-nonincreasing map $sh : M \to S$ such that $sh|_S = id$. In particular, all souls must be isometric to each other.*

(2) (Perel'man, 1993) *The map $sh : M \to S$ is a submetry. From this it follows that S must be a point if all sectional curvatures based at just one point are positive.*

Having reduced all complete nonnegatively curved manifolds to bundles over closed nonnegatively curved manifolds, it is natural to ask the converse question: Given a closed manifold S with non-negative curvature, which bundles over S admit complete metrics with $\sec \geq 0$? Clearly, the trivial bundles do. When $S = T^2$ Özaydın-Walschap in [**75**] have shown that this is the only 2-dimensional vector bundle that admits such a metric. Still, there doesn't seem to be a satisfactory general answer. If, for instance, we let $S = S^2$, then any 2-dimensional bundle is of the form $\left(S^3 \times \mathbb{C}\right)/S^1$, where S^1 is the Hopf action on S^3 and acts by rotations on $\mathbb{C}$ in the following way: $\omega \times z = \omega^k z$ for some integer k. This integer is the Euler number of the bundle. As we have a complete metric of nonnegative curvature on $S^3 \times \mathbb{C}$, the O'Neill formula from chapter 3 shows that these bundles admit metrics with $\sec \geq 0$.

There are some interesting examples of manifolds with positive and zero Ricci curvature that show how badly the soul theorem fails for such manifolds. In 1978, Gibbons-Hawking in [**43**] constructed Ricci flat metrics on quotients of $\mathbb{C}^2$ blown up at any finite number of points. Thus, one gets a Ricci flat manifold with arbitrarily large second Betti number. About ten years later Sha-Yang showed that the infinite connected sum

$$\left(S^2 \times S^2\right) \sharp \left(S^2 \times S^2\right) \sharp \cdots \sharp \left(S^2 \times S^2\right) \sharp \cdots$$

admits a metric with positive Ricci curvature, thus putting to rest any hopes for general theorems in this direction. Sha-Yang have a very nice survey in [**45**] describing these and other examples. The construction uses doubly warped product metrics on $I \times S^2 \times S^1$ as described in chapter 3.

5. Finiteness of Betti Numbers

The theorem we wish to prove is

THEOREM 86. (Gromov, 1978, 1981) *There is a constant $C(n)$ such that any complete manifold (M, g) with* $\sec \geq 0$ *satisfies*

(1) $\pi_1(M)$ can be generated by $\leq C(n)$ generators.

(2) For any field F of coefficients the Betti numbers are bounded:

$$\sum_{i=0}^{n} b_i(M,F) = \sum_{i=0}^{n} \dim H_i(M,F) \leq C(n).$$

Part (2) of this result is considered one of the deepest and most beautiful results in Riemannian geometry. Before embarking on the proof, let us put it in context. First, we should note that the Gibbons-Hawking and Sha-Yang examples show that a similar result cannot hold for manifolds with nonnegative Ricci curvature. Sha-Yang also exhibited metrics with positive Ricci curvature on the connected sums

$$\underbrace{\left(S^2 \times S^2\right) \sharp \left(S^2 \times S^2\right) \sharp \cdots \sharp \left(S^2 \times S^2\right)}_{k \text{ times}}.$$

For large k, the Betti number bound shows that these connected sums cannot have a metric with nonnegative sectional curvature. Thus, we have simply connected manifolds that admit positive Ricci curvature but not nonnegative sectional curvature. The reader should also consult our discussion of manifolds with nonnegative curvature operator at the end of chapters 7 and 8 to get an appreciation for how rigid manifolds with nonnegative curvature operator are. Let us list the open problems that were posed there and settled for manifolds with nonnegative curvature operator:

(i) (H. Hopf) Does $S^2 \times S^2$ admit a metric with positive sectional curvature?

(ii) (H. Hopf) If M is even-dimensional, does $\sec \geq 0$ (> 0) imply $\chi(M) \geq 0$ (> 0)?

(iii) (Gromov) If $\sec \geq 0$, is $\sum_{i=0}^{n} b_i(M,F) \leq 2^n$?

Recall that these questions were also discussed in chapter 7 under additional assumptions about the isometry group.

First we establish part (1) of Gromov's theorem. The proof resembles that of the critical point estimate lemma from the previous section.

Proof of (1). We shall construct what is called a *short set* of generators for $\pi_1(M)$. We consider $\pi_1(M)$ as acting by deck transformations on the universal covering $\tilde{M}$ and fix $p \in \tilde{M}$. We then inductively select a generating set $\{g_1, g_2, \ldots\}$ such that

(a) $d(p, g_1(p)) \leq d(p, g(p))$ for all $g \in \pi_1(M) - \{e\}$.

(b) $d(p, g_k(p)) \leq d(p, g(p))$ for all $g \in \pi_1(M) - \langle g_1, \ldots, g_{k-1} \rangle$.

Now join p and $g_k(p)$ by segments σ_k (see Figure 11.16). We claim that the angle between any two such segments is $\geq \pi/3$.

Otherwise, the hinge version of Toponogov's theorem would imply

$$\begin{aligned} \left(d\left(g_{k+l}(p), g_k(p)\right)\right)^2 &< \left(d\left(p, g_k(p)\right)\right)^2 + \left(d\left(p, g_{k+l}(p)\right)\right)^2 \\ &\quad - d\left(p, g_k(p)\right) d\left(p, g_{k+l}(p)\right) \\ &\leq \left(d\left(p, g_{k+l}(p)\right)\right)^2. \end{aligned}$$

But then

$$d\left(g_{k+l}^{-1} \circ g_k(p), p\right) < d\left(p, g_{k+l}(p)\right),$$

which contradicts our choice of g_{k+l}.

It now follows that there can be at most

$$\frac{v(n-1,1,\pi)}{v\left(n-1,1,\frac{\pi}{6}\right)}$$

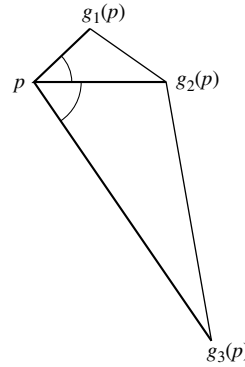

Figure 11.16

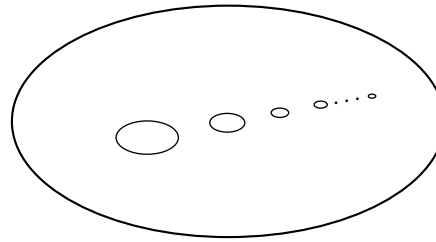

Figure 11.17

elements in the set $\{g_1, g_2, \ldots\}$. We have therefore produced a generating set with a bounded number of elements. □

The proof of the Betti number estimate is established through several lemmas. First, we need to make three definitions for metric balls. Throughout, we fix a Riemannian n-manifold M with $\sec \geq 0$ and a field F of coefficients for our homology theory

$$H_*(\cdot, F) = H_*(\cdot) = H_0(\cdot) \oplus \cdots \oplus H_n(\cdot).$$

The field will be suppressed throughout the proof.

Content: The *content* of a metric ball $B(p, r) \subset M$ is

$$\text{cont}B(p, r) = \text{rank}\left(H_*\left(B\left(p, \frac{1}{5}r\right)\right) \to H_*(B(p, r))\right).$$

The reason for working with content, rather that just the rank of $H_*(B(p, r))$ itself, is that metric balls might not have infinitely generated homology. However, if $O_1 \subset M$ is any bounded subset of a manifold and $\bar{O}_1 \subset O_2 \subset M$, then the image of $H_*(O_1)$ in $H_*(O_2)$ is finitely generated. In Figure 11.17 we have taken a planar domain and extracted infinitely many discs of smaller and smaller size. This yields a compact set with infinite topology. Nevertheless, this set has finitely generated topology when mapped into any neighborhood of itself, as that has the effect of canceling all of the smallest holes.

Corank: The *corank* of a set $A \subset M$ is defined as the largest integer k such that we can find k metric balls $B(p_1, r_1), \ldots, B(p_k, r_k)$ with the properties

(a) There is a critical point x_i for p_i with $d(p_i, x_i) = 10r_i$.

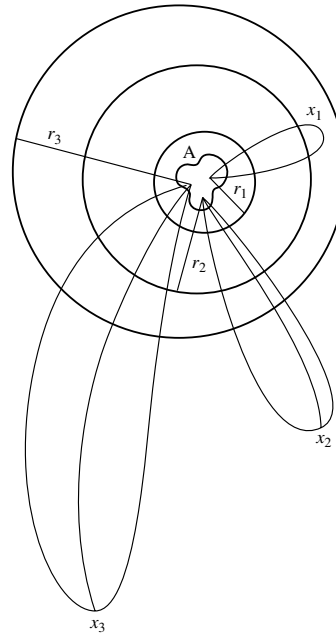

Figure 11.18

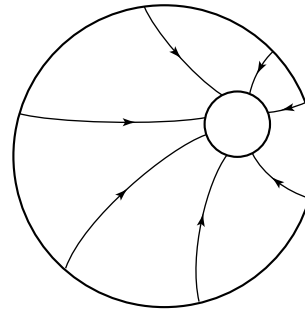

Figure 11.19

(b) $r_i \geq 3r_{i-1}$ for $i = 2, \ldots, k$.
(c) $A \subset \bigcap_{i=1}^{k} B(p_i, r_i)$.

In Figure 11.18 we have a picture of how the set A and the larger circles might be situated relative to each other.

Compressibility: We say that a ball $B(p, r)$ is *compressible* if it contains a ball $B(q, r') \subset B(p, r)$ such that

(a) $r' \leq \frac{r}{2}$.
(b) $\mathrm{cont}B(q, r') \geq \mathrm{cont}B(p, r)$.

If a ball is not compressible we call it *incompressible*. Note that any ball with content > 1, can be successively compressed to an incompressible ball. Figure 11.19 gives a schematic picture of a ball that can be compressed to a smaller ball.

We shall now tie these three concepts together through some lemmas that will ultimately lead us to the proof of the Betti number estimate. Observe that for large r, the ball $B(p,r)$ contains all the topology of M, so

$$\mathrm{cont}B(p,r) = \sum_i b_i(M).$$

Also, the corank of such a ball must be zero, as there can't be any critical points outside this ball. The idea is now to compress this ball until it becomes incompressible and then estimate its content in terms of balls that have corank 1. We shall in this way successively be able to estimate the content of balls of fixed corank in terms of the content of balls with one higher corank. The proof is then finished first, by showing that the corank of a ball is uniformly bounded by 100^n, second, by observing that balls of maximal corank must be contractible and therefore have content 1 (otherwise they would contain critical points for the center, and the center would have larger corank).

LEMMA 63. *The corank of any set $A \subset M$ is bounded by 100^n.*

PROOF. Suppose that A has corank larger than 100^n. Select balls $B(p_1,r_1)$, $\ldots$, $B(p_k,r_k)$ with corresponding critical points $x_1,\ldots,x_k$, where $k > 100^n$. Now choose $z \in A$ and join z to x_i by segments σ_i. As in the critical point estimate lemma from the previous section, we can then find two of these segments σ_i and σ_j that form an angle $< 1/6$ at z.

For simplicity, suppose $i < j$ and define

$$\begin{aligned} a_i &= \ell(\sigma_i) = d(z,x_i), \\ a_j &= \ell(\sigma_j) = d(z,x_j), \\ l &= d(x_i,x_j), \end{aligned}$$

and observe that

$$\begin{aligned} b_i &= d(z,p_i) \leq r_i, \\ b_j &= d(z,p_j) \leq r_j. \end{aligned}$$

Figure 11.20 gives two pictures explaining the notation in the proof.

The triangle inequality implies

$$\begin{aligned} a_i &\leq 10r_i + b_i \leq 11r_i, \\ a_j &\geq 10r_j - r_j \geq 9r_j. \end{aligned}$$

Also, $r_j \geq 3r_i$, so we see that $a_j > a_i$. As in the critical point estimate lemma, we can conclude that

$$l \leq a_j - \frac{3}{4}a_i.$$

Now use the triangle inequality to conclude

$$\begin{aligned} c &= d(p_i,x_j) \geq a_j - b_i \\ &\geq 10r_j - b_j - b_i \\ &\geq 8r_j \\ &\geq 24r_i \\ &\geq 20r_i = 2d(p_i,x_i). \end{aligned}$$

Yet another application of the triangle inequality will then imply

$$l \geq d(x_i,p_i).$$

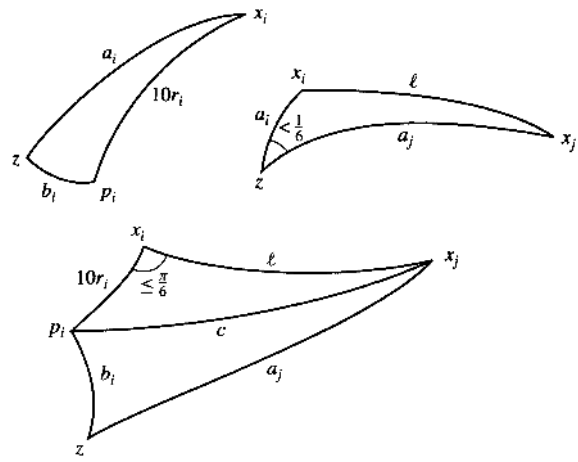

Figure 11.20

Since x_i is critical for p_i, we can now use the hinge version of Toponogov's theorem to conclude

$$\begin{aligned} c^2 &\leq (d(p_i, x_i))^2 + l^2 \\ &\leq \left(l + \frac{1}{2} d(p_i, x_i)\right)^2. \end{aligned}$$

Thus,

$$\begin{aligned} c &\leq l + \frac{1}{2} d(p_i, x_i) \\ &\leq l + 5r_i. \end{aligned}$$

The triangle inequality then implies

$$a_j \leq c + b_i \leq c + r_i \leq l + 6r_i.$$

However, we also have

$$a_i \geq 10r_i - b_i \geq 9r_i,$$

which together with

$$l \leq a_j - \frac{3}{4} a_i$$

implies

$$l \leq a_j - \frac{27}{4} r_i.$$

Thus, we have a contradiction:

$$l + \frac{27}{4} r_i \leq a_j \leq l + 6r_i.$$

□

Having established a bound on the corank, we can now try to check how the topology changes when we pass from balls of lower corank to balls of higher corank. Let $\mathcal{C}(k)$ denote the set of balls in M of corank $\geq k$, and $\mathcal{B}(k)$ the largest content of any ball in $\mathcal{C}(k)$.

LEMMA 64. *There is a constant $C(n)$ depending only on dimension such that*

$$\mathcal{B}(k) \leq C(n)\,\mathcal{B}(k+1).$$

PROOF. The number $\mathcal{B}(k)$ is, of course, realized by some incompressible ball $B(p,R)$. Now consider a ball $B(x,r)$ where $x \in B(p,R/4)$ and $r \leq R/20$. We claim that this ball lies in $\mathcal{C}(k+1)$. To see this, consider the ball

$$B(x,R/2) \subset B(p,R) \subset B(x,2R).$$

Since $B(p,R)$ is assumed to be incompressible, there must be a critical point for x in the annulus $B(x,2R) - B(x,R/2)$. For otherwise we could deform $B(p,R)$ to $B(x,R/2)$ inside $B(x,2R)$. This would imply that $\mathrm{cont}B(p,R) \leq \mathrm{cont}B(x,R/2)$ and thus contradict incompressibility of $B(p,R)$. We can now show that $B(x,r) \in \mathcal{C}(k+1)$. Using that $B(p,R) \in \mathcal{C}(k)$, select $B(p_1,r_1), \ldots, B(p_l,r_l)$, $l \geq k$, as in the definition of corank. Then pick a critical point y for x in $B(x,5R) - B(x,R/2)$ and consider the ball $B(x,d(x,y)/10)$. Then the balls $B(p_1,r_1), \ldots, B(p_l,r_l)$, $B(x,d(x,y)/10)$ can be used to show that $B(x,r)$ has corank $\geq l+1 > k$.

Now cover $B(p,R/5)$ by balls $B(p_i,R/100)$, $i = 1,\ldots,m$. If we suppose that the balls $B(p_i,R/200)$ are pairwise disjoint, then we must have:

$$m \leq \frac{v(n,0,2R)}{v\left(n,0,\frac{1}{200}R\right)} = 400^n.$$

Next consider the sets

$$B\left(p_i,\frac{1}{2}R\right) \subset B(p,R).$$

First, we claim that

$$\mathrm{cont}B(p,R) \leq \mathrm{rank}\left(H_*\left(\bigcup_{i=1}^m B\left(p_i,\frac{1}{100}R\right)\right) \to H_*\left(\bigcup_{i=1}^m B\left(p_i,\frac{1}{2}R\right)\right)\right)$$

This follows from the simple observation that if $A \subset B \subset C \subset D$, then

$$\mathrm{rank}\,(H_*(A) \to H_*(D)) \leq \mathrm{rank}\,(H_*(B) \to H_*(C))$$

To estimate the right-hand side of the above inequality, it is natural to suppose that we can use a Mayer-Vietoris argument, together with induction on m, to show

$$\begin{aligned}&\mathrm{rank}\left(H_*\left(\bigcup_{i=1}^m B\left(p_i,\frac{1}{100}R\right)\right) \to H_*\left(\bigcup_{i=1}^m B\left(p_i,\frac{1}{2}R\right)\right)\right)\\ \leq &\sum_{\substack{i_1<\cdots<i_s\\ 1\leq s\leq m}} \mathrm{rank}\left(H_*\left(\bigcap_{t=1}^s B\left(p_{i_t},\frac{1}{100}R\right)\right) \to H_*\left(\bigcap_{t=1}^s B\left(p_{i_t},\frac{1}{2}R\right)\right)\right).\end{aligned}$$

We then observe that if

$$\bigcap_{t=1}^s B\left(p_{i_t},\frac{1}{100}R\right) \neq \emptyset,$$

then the triangle inequality implies (see also below)

$$\bigcap_{t=1}^s B\left(p_{i_t},\frac{1}{100}R\right) \subset B\left(p_{i_1},\frac{1}{100}R\right) \subset B\left(p_{i_1},\frac{1}{20}R\right) \subset \bigcap_{t=1}^s B\left(p_{i_t},\frac{1}{2}R\right).$$

As each of the balls $B(p_i, R/10) \in \mathcal{C}(k+1)$, and there can be at most 2^m nonempty intersections, we then arrive at the estimate

$$\mathrm{cont}B(p,R) = \mathcal{B}(k) \leq 2^{400^n} \cdot \mathcal{B}(k+1).$$

This is the desired inequality. □

We now claim that

$$\mathrm{cont}M \leq 2^{40000^n},$$

which will, of course, prove the theorem. The above lemma clearly yields that

$$\begin{aligned} \mathrm{cont}M &= \mathcal{B}(0) \\ &\leq \mathcal{B}(k) \cdot \left(2^{400^n}\right)^k \\ &= \mathcal{B}(k) \cdot 2^{k \cdot 400^n} \\ &\leq \mathcal{B}(k) \cdot 2^{40000^n}, \end{aligned}$$

where $k \leq 100^n$ is the largest possible corank in M. It then remains to check that $\mathcal{B}(k) = 1$. However, it follows from the above that if $\mathcal{C}(k)$ contains an incompressible ball, then $\mathcal{C}(k+1) \neq \emptyset$. Thus, all balls in $\mathcal{C}(k)$ are compressible, but then they must have minimal content 1.

The above estimate on the rank of the inclusion

$$H_*\left(\bigcup_{i=1}^m B\left(p_i, \frac{R}{100}\right)\right) \to H_*\left(\bigcup_{i=1}^m B\left(p_i, \frac{R}{2}\right)\right),$$

in terms of the ranks of all the intersections, is in fact not quite right. One actually needs to consider the doubly indexed family $B\left(p_i, 10^{-j-1}R\right)$, $j = 1, \ldots, n+2$, where we assume that for each fixed j the family covers $B\left(p, \frac{1}{5}R\right)$. The correct estimate is then that the rank of the inclusion

$$H_*\left(\bigcup_{i=1}^m B\left(p_i, \frac{R}{10^{n+2}}\right)\right) \to H_*\left(\bigcup_{i=1}^m B\left(p_i, \frac{R}{2}\right)\right)$$

is bounded by the rank of all of the possible intersections

$$H_*\left(\bigcap_{t=1}^s B\left(p_{i_t}, \frac{R}{10^{j+1}}\right)\right) \to H_*\left(\bigcap_{t=1}^s B\left(p_{i_t}, \frac{R}{2 \cdot 10^{j-1}}\right)\right)$$

Whenever such an intersection

$$\bigcap_{t=1}^s B\left(p_{i_t}, 10^{-j-1}R\right) \neq \emptyset,$$

we still have the inclusions

$$\begin{aligned} \bigcap_{t=1}^s B\left(p_{i_t}, \frac{R}{10^{j+1}}\right) &\subset B\left(p_{i_1}, \frac{R}{10^{j+1}}\right) \\ &\subset B\left(p_{i_1}, \frac{R}{2 \cdot 10^j}\right) \\ &\subset \bigcap_{t=1}^s B\left(p_{i_t}, \frac{R}{2 \cdot 10^{j-1}}\right). \end{aligned}$$

So we can still estimate those ranks by the content of balls in $\mathcal{C}(k+1)$. We have, however, more intersections and also more balls, as this time the smaller balls

$B\left(p_i, 10^{-n-1}R\right)$ have to cover. One can easily compute the correct Betti number estimate with these modifications. The reader should consult the survey by Cheeger in [**24**] for the complete story.

The Betti number theorem can easily be proved in the more general context of manifolds with lower sectional curvature bounds, but one must then also assume an upper diameter bound. Otherwise, the ball covering arguments, and also the estimates using Toponogov's theorem, won't work. Thus, there is a constant $C(n, D, k)$ such that any closed Riemannian n-manifold (M, g) with $\sec \geq k$ and $\operatorname{diam} \leq D$ has the properties that

(1) $\pi_1(M)$ can be generated by $\leq C(n, k, D)$ elements,
(2) $\sum_{i=0}^{n} b_i(M, F) \leq C(n, D, k)$.

6. Homotopy Finiteness

This section is devoted to a result that interpolates between Cheeger's finiteness theorem and Gromov's Betti number estimate. We know that in Gromov's theorem the class under investigation contains infinitely many homotopy types, while if we have a lower volume bound and an upper curvature bound as well, Cheeger's result says that we have finiteness of diffeomorphism types.

THEOREM 87. (Grove-Petersen, 1988) *Given an integer $n > 1$ and numbers $v, D, k \in (0, \infty)$, the class of Riemannian n-manifolds with*

$$\begin{aligned} \operatorname{diam} &\leq D, \\ \operatorname{vol} &\geq v, \\ \sec &\geq -k^2 \end{aligned}$$

contains only finitely many homotopy types.

As with the other proofs in this chapter we need to proceed in stages. First, we present the main technical result.

LEMMA 65. *For a manifold as in the above theorem, we can find $\alpha = \alpha(n, D, v, k) \in \left(0, \frac{\pi}{2}\right)$ and $\delta = \delta(n, D, v, k) > 0$ such that if $p, q \in M$ satisfy $d(p, q) \leq \delta$, then either p is α-regular for q or q is α-regular for p.*

PROOF. The proof is by contradiction and based on a suggestion by Cheeger. Assume there is a pair of points $p, q \in M$ that are not α-regular with respect to each other, and set $l = d(p, q) \leq \delta$. Let $\Gamma(p, q)$ denote the set of unit speed segments from p to q, and define

$$\begin{aligned} \dot{\Gamma}_{pq} &= \{v \in T_pM : v = \dot{\sigma}(0), \sigma \in \Gamma(p, q)\}, \\ \dot{\Gamma}_{qp} &= \{-v \in T_qM : v = \dot{\sigma}(r), \sigma \in \Gamma(p, q)\}. \end{aligned}$$

Then the two sets $\dot{\Gamma}_{pq}$ and $\dot{\Gamma}_{qp}$ of unit vectors are by assumption $(\pi - \alpha)$-dense in the unit sphere. It is a simple exercise to show that if $A \subset S^{n-1}$, then the function

$$t \to \frac{\operatorname{vol} B(A, t)}{v(n-1, 1, t)}$$

is nonincreasing (see also exercises to chapter 9). In particular, for any $(\pi-\alpha)$-dense set $A \subset S^{n-1}$

$$\begin{aligned}\mathrm{vol}\left(S^{n-1}-B\left(A,\alpha\right)\right) &= \mathrm{vol}S^{n-1}-\mathrm{vol}B\left(A,\alpha\right)\\ &\leq \mathrm{vol}S^{n-1}-\mathrm{vol}S^{n-1}\cdot\frac{v\left(n-1,1,\alpha\right)}{v\left(n-1,1,\pi-\alpha\right)}\\ &= \mathrm{vol}S^{n-1}\cdot\frac{v\left(n-1,1,\pi-\alpha\right)-v\left(n-1,1,\alpha\right)}{v\left(n-1,1,\pi-\alpha\right)}.\end{aligned}$$

Now choose $\alpha<\frac{\pi}{2}$ such that

$$\mathrm{vol}S^{n-1}\cdot\frac{v\left(n-1,1,\pi-\alpha\right)-v\left(n-1,1,\alpha\right)}{v\left(n-1,1,\pi-\alpha\right)}\cdot\int_0^D\left(\mathrm{sn}_k\left(t\right)\right)^{n-1}dt=\frac{v}{6}.$$

Thus, the two cones (see exercises to chapter 9) satisfy

$$\begin{aligned}\mathrm{vol}B^{S^{n-1}-B\left(\dot{\Gamma}_{pq},\alpha\right)}\left(p,D\right) &\leq \frac{v}{6},\\ \mathrm{vol}B^{S^{n-1}-B\left(\dot{\Gamma}_{qp},\alpha\right)}\left(q,D\right) &\leq \frac{v}{6}.\end{aligned}$$

We now use Toponogov's theorem to choose δ such that any point in M that does not lie in one of these two cones must be close to either p or q (Figure 11.21 shows how a small δ will force the other leg in the triangle to be smaller than r). To this end, pick $r>0$ such that

$$v\left(n,-k^2,r\right)=\frac{v}{6}.$$

We now claim that if δ is sufficiently small, then

$$M=B\left(p,r\right)\cup B\left(q,r\right)\cup B^{S^{n-1}-B\left(\dot{\Gamma}_{pq},\alpha\right)}\left(p,D\right)\cup B^{S^{n-1}-B\left(\dot{\Gamma}_{qp},\alpha\right)}\left(q,D\right).$$

This will, of course, lead to a contradiction, as we would then have

$$\begin{aligned}v &\leq \mathrm{vol}M\\ &\leq \mathrm{vol}\left(B\left(p,r\right)\cup B\left(q,r\right)\cup B^{S^{n-1}-B\left(\dot{\Gamma}_{pq},\alpha\right)}\left(p,D\right)\cup B^{S^{n-1}-B\left(\dot{\Gamma}_{qp},\alpha\right)}\left(q,D\right)\right)\\ &\leq 4\cdot\frac{v}{6}<v.\end{aligned}$$

To see that these sets cover M, observe that if

$$x\notin B^{S^{n-1}-B\left(\dot{\Gamma}_{pq},\alpha\right)}\left(p,D\right),$$

then there is a segment from x to p and a segment from p to q that form an angle $\leq\alpha$. (See Figure 11.22.)

Thus, we have from Toponogov's theorem that

$$\cosh d\left(x,q\right)\leq\cosh l\cosh d\left(x,p\right)-\sinh l\sinh d\left(x,p\right)\cos\left(\alpha\right).$$

If also

$$x\notin B^{S^{n-1}-B\left(\dot{\Gamma}_{qp},\alpha\right)}\left(q,D\right),$$

we have in addition,

$$\cosh d\left(x,p\right)\leq\cosh l\cosh d\left(x,q\right)-\sinh l\sinh d\left(x,q\right)\cos\left(\alpha\right).$$

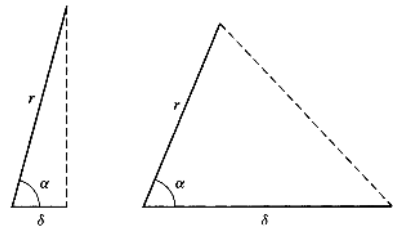

Figure 11.21

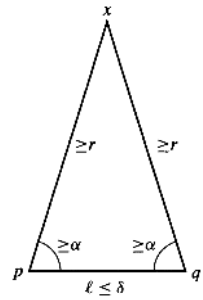

Figure 11.22

If in addition $d(x,p) > r$ and $d(x,q) > r$, we get

$$\begin{aligned} \cosh d(x,q) &\leq \cosh l \cosh d(x,p) - \sinh l \sinh d(x,p) \cos(\alpha) \\ &\leq \cosh d(x,p) \\ &\quad + (\cosh l - 1) \cosh D - \sinh l \sinh r \cos(\alpha) \end{aligned}$$

and

$$\begin{aligned} \cosh d(x,p) &\leq \cosh d(x,q) \\ &\quad + (\cosh l - 1) \cosh D - \sinh l \sinh r \cos(\alpha). \end{aligned}$$

However, as $l \to 0$, we see that the quantity

$$\begin{aligned} f(l) &= (\cosh l - 1) \cosh D - \sinh l \sinh r \cos(\alpha) \\ &= (-\sinh r \cos\alpha)\, l + O\left(l^2\right) \end{aligned}$$

becomes negative. Thus, we can find $\delta(D, r, \alpha) > 0$ such that for $l \leq \delta$ we have

$$(\cosh l - 1) \cosh D - \sinh l \sinh r \cos(\alpha) < 0.$$

We have then arrived at another contradiction, as this would imply

$$\cosh d(x,q) < \cosh d(x,p)$$

and

$$\cosh d(x,p) < \cosh d(x,q)$$

at the same time. Thus, the sets cover as we claimed. As this covering is also impossible, we are lead to the conclusion that under the assumption that $d(p,q) \leq \delta$, we must have that either p is α-regular for q or q is α-regular for p. □

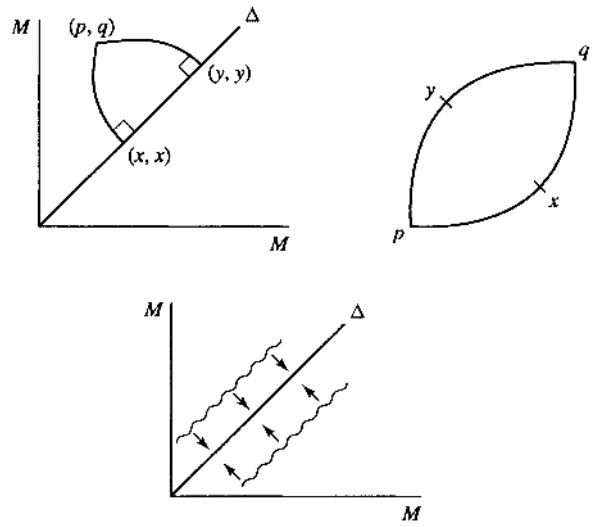

Figure 11.23

As it stands, this lemma seems rather strange and unmotivated. A little analysis will, however, enable us to draw some very useful conclusions from it.

Consider the product $M \times M$ with the product metric. Geodesics in this space are of the form (γ_1, γ_2), where both γ_1, γ_2 are geodesics in M. In $M \times M$ we have the diagonal $\Delta = \{(x, x) : x \in M\}$, which is a compact submanifold. Note that

$$T_{(p,p)}\Delta = \{(v, v) : v \in T_pM\},$$

and consequently, the normal bundle is

$$\nu(\Delta) = \{(v, -v) : v \in T_pM\}.$$

Therefore, if

$$(\sigma_1, \sigma_2) : [a, b] \to M \times M$$

is a segment from (p, q) to Δ, then we must have that $\dot{\sigma}_1(b) = -\dot{\sigma}_2(b)$. Thus these two segments can be joined at the common point $\sigma_1(b) = \sigma_2(b)$ to form a geodesic from p to q in M. This geodesic is, in fact, a segment, for otherwise, we could find a shorter curve from p to q. Dividing this curve in half would then produce a shorter curve from (p, q) to Δ. Thus, we have a bijective correspondence between segments from p to q and segments from (p, q) to Δ. Moreover,

$$\sqrt{2} \cdot d((p, q), \Delta) = d(p, q).$$

The above lemma now implies

COROLLARY 44. *Any point within distance $\delta/\sqrt{2}$ of Δ is α-regular for Δ.*

Figure 11.23 shows how the contraction onto the diagonal works and also how segments to the diagonal are related to segments in M.

Thus, we can find a curve of length $\leq \frac{1}{\cos\alpha} \cdot d((p, q), \Delta)$ from any point in this neighborhood to Δ. Moreover, this curve depends continuously on (p, q). We can translate this back into M. Namely, if $d(p, q) < \delta$, then p and q are joined by a curve $t \to H(p, q, t)$, $0 \leq t \leq 1$, whose length is $\leq \frac{\sqrt{2}}{\cos\alpha} \cdot d(p, q)$. Furthermore,

the map $(p,q,t) \to H(p,q,t)$ is continuous. For simplicity, we let $C = \frac{\sqrt{2}}{\cos\alpha}$ in the constructions below.

We now have the first ingredient in our proof.

COROLLARY 45. *If $f_0, f_1 : X \to M$ are two continuous maps such that*

$$d(f_0(x), f_1(x)) < \delta$$

for all $x \in X$, then f_0 and f_1 are homotopy equivalent.

For the next construction, recall that a k-simplex Δ^k can be thought of as the set of affine linear combinations of all the basis vectors in $\mathbb{R}^{k+1}$, i.e.,

$$\Delta^k = \left\{ (x^0, \ldots, x^k) : x^0 + \cdots + x^k = 1 \text{ and } x^0, \ldots, x^k \in [0,1] \right\}.$$

The basis vectors $e_i = \left(\delta_i^1, \ldots, \delta_i^k\right)$ are called the vertices of the simplex.

LEMMA 66. *Suppose we have $k+1$ points $p_0, \ldots, p_k \in B(p,r) \subset M$. If*

$$2r\frac{C^k - 1}{C - 1} < \delta,$$

then we can find a continuous map

$$f : \Delta^k \to B\left(p, r + 2r \cdot C \cdot \frac{C^k - 1}{C - 1}\right),$$

where $f(e_i) = p_i$.

PROOF. Figure 11.24 gives the essential idea of the proof. The proof goes by induction on k. For $k = 0$ there is nothing to show.

Suppose now that the statement holds for k and that we have $k+2$ points $p_0, \ldots, p_{k+1} \in B(p,r)$. First, we find a map

$$f : \Delta^k \to B\left(p, 2r \cdot C \cdot \frac{C^k - 1}{C - 1} + r\right)$$

with $f(e_i) = p_i$ for $i = p_0, \ldots, p_k$. We then define

$$\begin{aligned} \bar{f} &: \Delta^{k+1} \to B\left(p, r + 2r \cdot C \cdot \frac{C^{k+1} - 1}{C - 1}\right), \\ \bar{f}\left(x^0, \ldots, x^k, x^{k+1}\right) &= H\left(f\left(\frac{x^0}{\sum_{i=1}^k x^i}, \ldots, \frac{x^k}{\sum_{i=1}^k x^i}\right), p_{k+1}, x^{k+1}\right). \end{aligned}$$

This clearly gives a well-defined continuous map as long as

$$\begin{aligned} & d\left(f\left(\frac{x^0}{\sum_{i=1}^k x^i}, \ldots, \frac{x^k}{\sum_{i=1}^k x^i}\right), p_{k+1}\right) \\ \leq\ & d\left(f\left(\frac{x^0}{\sum_{i=1}^k x^i}, \ldots, \frac{x^k}{\sum_{i=1}^k x^i}\right), p\right) + d(p, p_{k+1}) \\ \leq\ & \left(2r \cdot C \cdot \frac{C^k - 1}{C - 1} + r\right) + r \\ =\ & 2r \cdot \frac{C^{k+1} - 1}{C - 1} \\ <\ & \delta. \end{aligned}$$

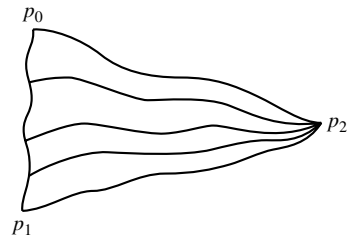

Figure 11.24

Moreover, it has the property that

$$\begin{aligned} d\left(p, \bar{f}\left(\cdot\right)\right) &\leq d\left(p, p_{k+1}\right) + d\left(p_{k+1}, \bar{f}\left(\cdot\right)\right) \\ &\leq r + 2r \cdot C \cdot \frac{C^{k+1} - 1}{C - 1}. \end{aligned}$$

This concludes the induction step. □

Note that if we select a face spanned by, say, $(e_1, \ldots, e_k)$ of the simplex Δ^k, then we could, of course, construct a map in the above way by mapping e_i to p_i. The resulting map will, however, be the same as if we constructed the map on the entire simplex and restricted it to the selected face.

We can now prove finiteness of homotopy types. Observe that the class we work with is precompact in the Gromov-Hausdorff distance as we have an upper diameter bound and a lower bound for the Ricci curvature. Thus it suffices to prove

LEMMA 67. *There is an $\varepsilon = \varepsilon(n, k, v, D) > 0$ such that if two Riemannian n-manifolds (M, g_1) and (N, g_2) satisfy*

$$\begin{aligned} \text{diam} &\leq D, \\ \text{vol} &\geq v, \\ \text{sec} &\geq -k^2, \end{aligned}$$

and

$$d_{G-H}(M, N) < \varepsilon,$$

then they are homotopy equivalent.

PROOF. Suppose M and N are given as in the lemma, together with a metric d on $M \amalg N$, inside which the two spaces are ε Hausdorff close. The size of ε will be found through the construction.

First, triangulate both manifolds in such a way that any simplex of the triangulation lies in a ball of radius ε. Using the triangulation on M, we can now construct a continuous map $f : M \to N$ as follows. First we use the Hausdorff approximation to map all the vertices $\{p_\alpha\} \subset M$ of the triangulation to points $\{q_\alpha\} \subset N$ such that $d(p_\alpha, q_\alpha) < \varepsilon$. If now $(p_{\alpha_0}, \ldots, p_{\alpha_n})$ forms a simplex in the triangulation of M, then we constructed the triangulation such that

$$(p_{\alpha_0}, \ldots, p_{\alpha_n}) \subset B(x, \varepsilon)$$

for some $x \in M$. Thus

$$(q_{\alpha_0}, \ldots, q_{\alpha_n}) \subset B(q_{\alpha_0}, 4\varepsilon).$$

Therefore, if

$$8\varepsilon\frac{C^n-1}{C-1}<\delta,$$

then we can use the above lemma to define f on the simplex spanned by $(p_{\alpha_0},\ldots,p_{\alpha_n})$. In this way we get a map $f:M\to N$ by constructing it on each simplex as just described. To see that it is continuous, we must check that the construction agrees on common faces of simplices. But this follows, as the construction is natural with respect to restriction to faces of simplices. We now need to estimate how good a Hausdorff approximation f is. To this end, select $x\in M$ and suppose that it lies in the face spanned by the vertices $(p_{\alpha_0},\ldots,p_{\alpha_n})$. Then we have

$$\begin{aligned} d(x,f(x)) &\leq d(x,p_{\alpha_0})+d(p_{\alpha_0},f(x))\\ &\leq 2\varepsilon+\varepsilon+d(q_{\alpha_0},f(x))\\ &\leq 3\varepsilon+4\varepsilon+8\varepsilon\cdot C\cdot\frac{C^n-1}{C-1}\\ &= 7\varepsilon+8\varepsilon\cdot C\cdot\frac{C^n-1}{C-1}. \end{aligned}$$

We can now construct $g:N\to M$ in the same manner. This map will, of course, also satisfy

$$d(y,g(y))\leq 7\varepsilon+8\varepsilon\cdot C\cdot\frac{C^n-1}{C-1}.$$

It is now possible to estimate how close the compositions $f\circ g$ and $g\circ f$ are to the identity maps on N and M, respectively, as follows:

$$\begin{aligned} d(y,f\circ g(y)) &\leq d(y,g(y))+d(g(y),f\circ g(y))\\ &\leq 14\varepsilon+16\varepsilon\cdot C\cdot\frac{C^n-1}{C-1};\\ d(x,g\circ f(x)) &\leq 14\varepsilon+16\varepsilon\cdot C\cdot\frac{C^n-1}{C-1}. \end{aligned}$$

As long as

$$14\varepsilon+16\varepsilon\cdot C\cdot\frac{C^n-1}{C-1}<\delta,$$

we can then conclude that these compositions are homotopy equivalent to the respective identity maps. In particular, the two spaces are homotopy equivalent. □

Note that as long as

$$16\varepsilon\cdot\frac{C^{n+1}-1}{C-1}<\delta,$$

the two spaces are homotopy equivalent. Thus, ε depends in an explicit way on $C=\frac{\sqrt{2}}{\cos\alpha}$ and δ. It is possible, in turn, to estimate α and δ from n,k,v, and D. We can therefore get an explicit estimate for how close spaces must be to ensure that they are homotopy equivalent. Given this explicit ε, it is then possible, using our work from the section on Gromov-Hausdorff distance, to find an explicit estimate for the number of homotopy types.

To conclude, let us compare the three finiteness theorems by Cheeger, Gromov, and Grove-Petersen. We have inclusions of classes of closed Riemannian n-manifolds

$$\left\{\begin{array}{lll}\text{diam}&\leq&D\\ \text{sec}&\geq&-k^2\end{array}\right\}\supset\left\{\begin{array}{lll}\text{diam}&\leq&D\\ \text{vol}&\geq&v\\ \text{sec}&\geq&-k^2\end{array}\right\}\supset\left\{\begin{array}{lll}\text{diam}&\leq&D\\ \text{vol}&\geq&v\\ |\text{sec}|&\leq&k^2\end{array}\right\}$$

with strengthening of conclusions from bounded Betti numbers to finitely many homotopy types to compactness in the $C^{1,\alpha}$ topology. In the special case of non-negative curvature Gromov's estimate actually doesn't depend on the diameter, thus yielding obstructions to the existence of such metrics on manifolds with complicated topology. For the other two results the diameter bound is still necessary. Consider for instance the family of lens spaces $\left\{S^3/\mathbb{Z}_p\right\}$ with curvature $= 1$. Now rescale these metrics so that they all have the same volume. Then we get a class which contains infinitely many homotopy types and also satisfies

$$\begin{gathered} \text{vol} = v, \\ 1 \geq \text{sec} > 0. \end{gathered}$$

The family of lens spaces $\left\{S^3/\mathbb{Z}_p\right\}$ with curvature $= 1$ also shows that the lower volume bound is necessary in both of these theorems.

Some further improvements are possible in the conclusion of the homotopy finiteness result. Namely, one can strengthen the conclusion to state that the class contains finitely many homeomorphism types. This was proved for $n \neq 3$ in [**51**] and in a more general case in [**76**]. One can also prove many of the above results for manifolds with certain types of integral curvature bounds, see for instance [**79**] and [**80**]. The volume [**50**] also contains complete discussions of generalizations to the case where one has merely Ricci curvature bounds.

7. Further Study

There are many texts that partially cover or expand the material in this chapter. We wish to attract attention to the surveys by Grove in [**45**], by Abresch-Meyer, Colding, Greene, and Zhu in [**50**], by Cheeger in [**24**], and by Karcher in [**28**]. The most glaring omission from this chapter is probably that of the Abresch-Gromoll theorem and other uses of the excess function. The above-mentioned articles by Zhu and Cheeger cover this material quite well.

8. Exercises

(1) Let (M, g) be a closed simply connected positively curved manifold. Show that if M contains a totally geodesic closed hypersurface (i.e., the shape operator is zero), then M is homeomorphic to a sphere. (Hint: first show that the hypersurface is orientable, and then show that the signed distance function to this hypersurface has only two critical points - a maximum and a minimum. This also shows that it suffices to assume that $H^1(M, \mathbb{Z}_2) = 0$.)

(2) Show that the converse of Toponogov's theorem is also true. In other words, if for some k the conclusion to Toponogov's theorem holds when hinges (or triangles) are compared to the same objects in S^2_k, then sec $\geq k$.

(3) (Heintze-Karcher) Let $\gamma \subset (M, g)$ be a geodesic in a Riemannian n-manifold with sec $\geq -k^2$. Let $T(\gamma, R)$ be the normal tube around γ of radius R, i.e., the set of points in M that can be joined to γ by a segment of length $\leq R$ that is perpendicular to γ. The last condition is superfluous when γ is a closed geodesic, but if it is a loop or a segment, then not all points in M within distance R of γ will belong to this tube. On this tube introduce coordinates (r, s, θ), where r denotes the distance to γ, s is the arc-length parameter on γ, and $\theta = \left(\theta^1, \ldots, \theta^{n-2}\right)$ are spherical

coordinates normal to γ. These give adapted coordinates for the distance r to γ. Show that as $r \to 0$ the metric looks like

$$g(r) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \cdot r^2 + O\left(r^3\right)$$

Using the lower sectional curvature bound, find an upper bound for the volume density on this tube. Conclude that

$$\mathrm{vol}T(\gamma, R) \leq f(n, k, R, \ell(\gamma)),$$

for some continuous function f depending on dimension, lower curvature bound, radius, and length of γ. Moreover, as $\ell(\gamma) \to 0$, $f \to 0$. Use this estimate to prove Cheeger's lemma from Chapter 10 and the main lemma on mutually critical points from the homotopy finiteness theorem. This shows that Toponogov's theorem is not needed for the latter result.

(4) Show that any vector bundle over a 2-sphere admits a complete metric of nonnegative sectional curvature. Hint: You need to know something about the classification of vector bundles over spheres. In this case k-dimensional vector bundles are classified by homotopy classes of maps from S^1, the equator of the 2-sphere, into $SO(k)$. This is the same as $\pi_1(SO(k))$, so there is only one 1-dimensional bundle, the 2-dimensional bundles are parametrized by $\mathbb{Z}$, and 2 higher dimensional bundles.

(5) Use Toponogov's theorem to show that b_γ is convex when $\sec \geq 0$.