Comparison Geometry
MSRI Publications
Volume **30**, 1997

The Comparison Geometry of Ricci Curvature

SHUNHUI ZHU

Abstract. We survey comparison results that assume a bound on the manifold's Ricci curvature.

1. Introduction

This is an extended version of the talk I gave at the Comparison Geometry Workshop at MSRI in the fall of 1993, giving a relatively up-to-date account of the results and techniques in the comparison geometry of Ricci curvature, an area that has experienced tremendous progress in the past five years.

The term "comparison geometry" had its origin in connection with the success of the Rauch comparison theorem and its more powerful global version, the Toponogov comparison theorem. The comparison geometry of *sectional curvature* represents many ingenious applications of these theorems and produced many beautiful results, such as the $\frac{1}{4}$-pinched Sphere Theorem [Berger 1960; Klingenberg 1961], the Soul Theorem [Cheeger and Gromoll 1972], the Generalized Sphere Theorem [Grove and Shiohama 1977], The Compactness Theorem [Cheeger 1967; Gromov 1981c], the Betti Number Theorem [Gromov 1981a], and the Homotopy Finiteness Theorem [Grove and Petersen 1988], just to name a few. The comparison geometry of Ricci curvature started as isolated attempts to generalize results about sectional curvature to the much weaker condition on Ricci curvature. Starting around 1987, many examples were constructed to demonstrate the difference between sectional curvature and Ricci curvature; in particular, Toponogov's theorem was shown not to hold for Ricci curvature. At the same time, many new tools and techniques were developed to generalize results about sectional curvature to Ricci curvature. We will attempt to present the highlights of this progress.

As often is the case, a survey paper becomes outdated before it goes to press. The same can be said about this one. In the past year, many beautiful results

Supported in part by an NSF grant.

were obtained in this area, mainly by T. Colding and J. Cheeger [Colding 1996a; 1996b; 1995; Cheeger and Colding 1995] (see also Colding's article on pages 83–98 of this volume). These results are not included here. To compensate for this, we have tried to give complete proofs of the results we discuss; these results now constitute the "standard" part in the comparison theory of Ricci curvature. We benefited a lot from a course taught by J. Cheeger at Stony Brook in 1988 and from [Cheeger 1991]. Thanks also go to the participants of a topic course I gave at Dartmouth in the winter of 1995.

2. The Main Comparison Theorem through Weitzenböck Formula

The relation between curvature and the geometry is traditionally introduced through the second variation of arclength, as was first used by Myers. We chose to introduce it from the Weitzenböck formula, which gives a uniform starting point for many applications, including the recent results of Colding.

For a smooth function f, we define its gradient, Hessian, and Laplacian by

$$\langle \nabla f, X\rangle = X(f), \quad \operatorname{Hess} f(X,Y) = \langle \nabla_X(\nabla f), Y\rangle, \quad \triangle f = \operatorname{tr}(\operatorname{Hess} f).$$

For a bilinear form A, we write $|A|^2 = \operatorname{tr}(AA^t)$.

Theorem 2.1 (The Weitzenböck formula). *Let (M^n, g) be a complete Riemannian manifold. Then, for any function $f \in C^3(M)$, we have*

$$\tfrac{1}{2}\triangle|\nabla f|^2 = |\operatorname{Hess} f|^2 + \langle \nabla f, \nabla(\triangle f)\rangle + \operatorname{Ric}(\nabla f, \nabla f)$$

pointwise.

Proof. Fix a point $p \in M$. Let $\{X_i\}_1^n$ be a local orthonormal frame field such that

$$\langle X_i, X_j\rangle = \delta_{ij}, \quad \nabla_{X_i} X_j(p) = 0.$$

Computation at p gives

$$\begin{aligned}
\tfrac{1}{2}\triangle|\nabla f|^2 &= \tfrac{1}{2}\sum_i X_i X_i \langle \nabla f, \nabla f\rangle = \sum_i X_i \langle \nabla_{X_i}\nabla f, \nabla f\rangle = \sum_i X_i \operatorname{Hess}(f)(X_i, \nabla f)\\
&= \sum_i X_i \operatorname{Hess}(f)(\nabla f, X_i) \quad \text{(Hessian is symmetric)}\\
&= \sum_i X_i \langle \nabla_{\nabla f}(\nabla f), X_i\rangle = \sum_i \langle \nabla_{X_i}\nabla_{\nabla f}(\nabla f), X_i\rangle + \langle \nabla_{\nabla f}(\nabla f), \nabla_{X_i} X_i\rangle\\
&= \sum_i \langle \nabla_{X_i}\nabla_{\nabla f}(\nabla f), X_i\rangle \quad \text{(the other term vanishes at } p\text{)}\\
&= \sum_i \langle R(X_i, \nabla f)\nabla f, X_i\rangle + \sum_i \langle \nabla_{\nabla f}\nabla_{X_i}\nabla f, X_i\rangle + \sum_i \langle \nabla_{[X_i, \nabla f]}\nabla f, X_i\rangle.
\end{aligned}$$

The first term is by definition $\mathrm{Ric}(\nabla f, \nabla f)$; the second term is

$$\begin{aligned}\sum_i (\nabla f)\langle \nabla_{X_i}\nabla f, X_i\rangle - \langle \nabla_{X_i}\nabla f, \nabla_{\nabla f} X_i\rangle &= (\nabla f)\sum_i \langle \nabla_{X_i}\nabla f, X_i\rangle - 0 \text{ (at } p) \\ &= (\nabla f)\triangle f = \langle \nabla f, \nabla(\triangle f)\rangle,\end{aligned}$$

and the third term is

$$\begin{aligned}\sum_i \mathrm{Hess}(f)([X_i, \nabla f], X_i) &= \sum_i \mathrm{Hess}(f)(\nabla_{X_i}\nabla f - \nabla_{\nabla f} X_i, X_i) \\ &= \sum_i \mathrm{Hess}(f)(\nabla_{X_i}\nabla f, X_i) - \mathrm{Hess}(f)(\nabla_{\nabla f} X_i, X_i) \\ &= \sum_i \mathrm{Hess}(f)(\nabla_{X_i}\nabla f, X_i) - 0 \text{ (at } p) \\ &= \sum_i \mathrm{Hess}(f)(X_i, \nabla_{X_i}\nabla f) \\ &= \sum_i \langle \nabla_{X_i}\nabla f, \nabla_{X_i}\nabla f\rangle = |\mathrm{Hess}(f)|^2.\end{aligned}$$

The theorem follows. □

The power of this formula is that we have the freedom to choose the function f. Most of the results of comparison geometry are obtained by choosing f to be the distance function, the eigenfunction, and the displacement function, among others.

We will consider the distance function. Fix a point p, and let $r(x) = d(p, x)$ be the distance from p to x. This defines a Lipschitz function on the manifold, smooth except on the cut locus of p. It also satisfies $|\nabla r| = 1$ where it is smooth. In geodesic polar coordinates at p, we have $\nabla r = \partial/\partial r$. Let $m(r)$ denote the mean curvature of the geodesic sphere at p with outer normal N, i.e., if $\{e_1, \dots, e_{n-1}\}$ be an orthonormal basis for the geodesic sphere, let

$$m(r) = \sum_{i=1}^{n-1} \langle \nabla_{e_i} N, e_i\rangle.$$

In geodesic polar coordinates, the volume element can be written as

$$d\,\mathrm{vol} = dr \wedge A_\omega(r)\, d\omega$$

where $d\omega$ is the volume form on the standard S^{n-1}. In what follows, we will suppress the dependence of $A_\omega(r)$ on ω for notational convenience. With these notations, we are now ready to state our main result of this section.

THEOREM 2.2 (MAIN COMPARISON THEOREM). *Let (M^n, g) be complete, and assume $\mathrm{Ric}(M) \geq (n-1)H$. Outside the cut locus of p, we have*:
(1) VOLUME ELEMENT COMPARISON:

$$\frac{A(r)}{A^H(r)} \text{ is nonincreasing along radial geodesics.}$$

(2) LAPLACIAN COMPARISON:

$$\triangle r \le \triangle^H r.$$

(3) MEAN CURVATURE COMPARISON:

$$m(r) \le m^H(r).$$

(Here quantities with a superscript H are the counterparts in the simply connected space form of constant curvature H.) Furthermore, equality holds if and only if all radial sectional curvatures are equal to H.

PROOF. We will prove the second part, the Laplacian comparison. The first and third parts follow from a lemma that we will prove momentarily.

Let $f(x) = r(x)$ in Theorem 2.1 and note that $|\nabla r| = 1$. We obtain, outside the cut locus of p,

$$|\operatorname{Hess} r|^2 + \frac{\partial}{\partial r}(\triangle r) + \operatorname{Ric}\left(\frac{\partial}{\partial r}, \frac{\partial}{\partial r}\right) = 0.$$

Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $\operatorname{Hess} r$. Since the exponential function is a radial isometry, one of the eigenvalues, say λ_1, is zero. By the Cauchy–Schwarz inequality, we have

$$|\operatorname{Hess}(r)|^2 = \lambda_2^2 + \cdots + \lambda_n^2 \ge \frac{(\lambda_2 + \cdots + \lambda_n)^2}{n-1} = \frac{\operatorname{tr}^2(\operatorname{Hess}(f))}{n-1} = \frac{(\triangle r)^2}{n-1}.$$

Thus, if $\operatorname{Ric} \ge (n-1)H$, then

$$\frac{(\triangle r)^2}{n-1} + \frac{\partial}{\partial r}(\triangle r) + (n-1)H \le 0.$$

Let $u = (n-1)/\triangle r$. Then

$$\frac{u'}{1 + Hu^2} \ge 1.$$

Note that $\triangle r \to (n-1)/r$ when $r \to 0$; thus $u \to r$. Integrating the above inequality gives

$$\triangle r \le \triangle^H r = \begin{cases} (n-1)\sqrt{H}\cot\sqrt{H}r & \text{for } H > 0, \\ (n-1)/r & \text{for } H = 0, \\ (n-1)\sqrt{-H}\coth\sqrt{-H}r & \text{for } H < 0. \end{cases}$$

We now discuss the equality case. If equality holds at r_0, then for any $r \le r_0$, all the inequalities in the above argument become equalities. In particular, the $n-1$ eigenvalues of $\operatorname{Hess}(r)$ are equal to $\sqrt{N}\cot\sqrt{H}r$ (to simplify the notation, we assume $H > 0$. For $H \le 0$, replace cot by coth.) Let X_i for $i = 2, 3, \dots, n$, be the orthonormal eigenvectors of $\operatorname{Hess}(r)$ at r; thus

$$\nabla_{X_i}\frac{\partial}{\partial r} = \sqrt{H}\cot\sqrt{H}rX_i.$$

Extend X_i in such a way that $[X_i, \partial/\partial r] = 0$ at r, then

$$\begin{aligned}
\sec\left(X_i, \frac{\partial}{\partial r}\right) &= -\left\langle \nabla_{\partial/\partial r}\nabla_{X_i}\frac{\partial}{\partial r}, X_i\right\rangle = -\langle \nabla_{\partial/\partial r}(\sqrt{H}\cot\sqrt{H}r)X_i, X_i\rangle \\
&= H\csc^2\sqrt{H}r - \langle \nabla_{\partial/\partial r}X_i, X_i\rangle \\
&= H\csc^2\sqrt{H}r - \sqrt{H}\cot\sqrt{H}r\left\langle \nabla_{X_i}\frac{\partial}{\partial r}, X_i\right\rangle \\
&= H\csc^2\sqrt{H}r - (\sqrt{H}\cot\sqrt{H}r)^2 = H. \qquad \square
\end{aligned}$$

We now give a more geometric interpretation of $\triangle r$, which proves parts (1) and (3) of Theorem 2.2.

LEMMA 2.3. *Given a complete Riemannian manifold* (M^n, g) *and a point* $p \in M$, *we have* $\triangle r = m(r)$ *and* $m(r) = A'(r)/A(r)$.

PROOF. By definition,

$$\begin{aligned}
\triangle r = \operatorname{tr}(\operatorname{Hess} r) &= \sum_{i=1}^{n-1}\langle \nabla_{e_i}(\nabla r), e_i\rangle + \langle \nabla_N(\nabla r), N\rangle \\
&= \sum_{i=1}^{n-1}\langle \nabla_{e_i}N, e_i\rangle + \langle \nabla_N N, N\rangle = \sum_{i=1}^{n-1}\langle \nabla_{e_i}N, e_i\rangle = m(r).
\end{aligned}$$

This proves the first equality. For the second, consider the map $\phi : T_pM \to M$ defined by $\phi(v) = \exp_p(rv)$. Let $\{v_1, \dots, v_{n-1}\}$ be an orthonormal basis for the unit sphere in T_pM. Then

$$\begin{aligned}
A(r) &= d\operatorname{vol}\left(\frac{\partial}{\partial r}, \phi(v_1), \dots, \phi(v_{n-1})\right) \\
&= d\operatorname{vol}\left(\frac{\partial}{\partial r}, d\exp_p(rv_1), \dots, d\exp_p(rv_{n-1})\right) \\
&= J_1(r) \wedge J_2(r) \wedge \cdots \wedge J_{n-1}(r),
\end{aligned}$$

where $J_i(r) = d\exp_p(rv_i)$. Fix r_0. We have

$$\frac{A'(r_0)}{A(r_0)} = \frac{\sum_{i=1}^{n-1} J_1(r_0) \wedge \cdots \wedge J_i'(r_0) \wedge \cdots \wedge J_{n-1}(r_0)}{J_1(r_0) \wedge J_2(r_0) \wedge \cdots \wedge J_{n-1}(r_0)}.$$

Let $\bar{J}_1(r), \dots, \bar{J}_{n-1}(r)$ be linear combinations (with constant coefficients) of the $J_i(r)$'s such that $\bar{J}_1(r_0), \dots, \bar{J}_{n-1}(r_0)$ form an orthonormal basis. Then

$$\begin{aligned}
\frac{A'(r_0)}{A(r_0)} &= \frac{\sum_{i=1}^{n-1} J_1(r_0) \wedge \cdots \wedge J_i'(r_0) \wedge \cdots \wedge J_{n-1}(r_0)}{J_1(r_0) \wedge J_2(r_0) \wedge \cdots \wedge J_{n-1}(r_0)} \\
&= \frac{\sum_{i=1}^{n-1} \bar{J}_1(r_0) \wedge \cdots \wedge \bar{J}_i'(r_0) \wedge \cdots \wedge \bar{J}_{n-1}(r_0)}{\bar{J}_1(r_0) \wedge \bar{J}_2(r_0) \wedge \cdots \wedge \bar{J}_{n-1}(r_0)} \\
&= \sum_{i=1}^{n-1} \bar{J}_1(r_0) \wedge \cdots \wedge \bar{J}_i'(r_0) \wedge \cdots \wedge \bar{J}_{n-1}(r_0) = \sum_{i=1}^{n-1}\langle \bar{J}_i'(r_0), \bar{J}_i(r_0)\rangle.
\end{aligned}$$

Let $f_i(t,s)=\exp_p(sv_i+t\vec{n})$. Then

$$J_i(r_0)=d\exp_p(r_0v_i)=\frac{\partial}{\partial s}\Big|_{s=0}f_i(t,s)$$

and

$$J_i'(r_0)=\frac{\partial}{\partial t}\Big|_{t=r_0}\frac{\partial}{\partial s}\Big|_{s=0}f_i(t,s)=\frac{\partial}{\partial s}\Big|_{s=0}\frac{\partial}{\partial t}\Big|_{t=r_0}f_i(t,s)=\nabla_{J_i(r_0)}N$$

Therefore, we also have $\bar{J}_i'(r_0)=\nabla_{\bar{J}_i(r_0)}N$. Thus

$$\frac{A'(r_0)}{A(r_0)}=\sum_{i=1}^{n-1}\langle\nabla_{\bar{J}_i(r_0)}N,\bar{J}_i(r_0)\rangle=m(r_0). \qquad \square$$

3. Volume Comparison and Its Applications

The applications of the volume comparison theorem are numerous; we will divide them into several sections. A common feature is that all results in this section are about the fundamental group and the first Betti number.

Bishop–Gromov volume comparison and its direct applications. For most applications of the volume comparison theorem, an integrated form is used. One can integrate the inequality in Theorem 2.2 along radial directions and along a subset of the unit sphere at p. Then:

THEOREM 3.1. *Let $r\le R$, $s\le S$, $r\le s$, $R\le S$, and let Γ be any measurable subset of S_p^{n-1}. Let $A_{r,R}^{\Gamma}(p)$ be the set of $x\in M$ such that $r\le r(x)\le R$ and any minimal geodesic γ from p to x satisfies $\dot\gamma(0)\in\Gamma$. Then*

$$\frac{\operatorname{vol}(A_{s,S}^{\Gamma}(p))}{\operatorname{vol}(A_{r,R}^{\Gamma}(p))}\ge\frac{\operatorname{vol}^H(A_{s,S}^{\Gamma})}{\operatorname{vol}^H(A_{r,R}^{\Gamma})},$$

with equality if and only if the radial curvatures are all equal to H.

REMARK. The strength of this theorem is that now the balls do not have to lie inside the cut locus; hence it is a global result.

We will give a detailed proof of this theorem, since it does not seem to be in the literature.

LEMMA 3.2. *Let f,g be two positive functions defined over $[0,+\infty)$. If f/g is nonincreasing, then for any $R>r>0$, $S>s>0$, $r>s$, $R>S$, we have*

$$\frac{\int_r^R f(t)\,dt}{\int_s^S f(t)\,dt}\le\frac{\int_r^R g(t)\,dt}{\int_s^S g(t)\,dt}.$$

PROOF. It suffices to show that the function

$$F(x,y)=\frac{\int_x^y f(t)\,dt}{\int_x^y g(t)\,dt}$$

satisfies

$$\frac{\partial F}{\partial x} \le 0, \quad \frac{\partial F}{\partial y} \le 0.$$

In fact, if this is true, then

$$\frac{\int_r^R f(t)\,dt}{\int_r^R g(t)\,dt} \le \frac{\int_r^S f(t)\,dt}{\int_r^S g(t)\,dt} \le \frac{\int_s^S f(t)\,dt}{\int_s^S g(t)\,dt},$$

which is what we want.

We now compute

$$\begin{aligned}\frac{\partial F}{\partial y} &= \frac{1}{(\int_x^y g(t)\,dt)^2}\left(f(y)\int_x^y g(t)\,dt - g(y)\int_x^y f(t)\,dt\right)\\ &= \frac{g(y)\int_x^y g(t)\,dt}{(\int_x^y g(t)\,dt)^2}\left(\frac{f(y)}{g(y)} - \frac{\int_x^y f(t)\,dt}{\int_x^y g(t)\,dt}\right).\end{aligned}$$

But

$$\frac{f(t)}{g(t)} \ge \frac{f(y)}{g(y)} \quad \text{for } x \le t \le y.$$

Thus

$$\int_x^y f(t)\,dt \ge \int_x^y \frac{f(y)}{g(y)} \cdot g(t)\,dt = \frac{f(y)}{g(y)}\int_x^y g(t)\,dt,$$

that is,

$$\frac{f(y)}{g(y)} \le \frac{\int_x^y f(t)\,dt}{\int_x^y g(t)\,dt},$$

which implies $\partial F/\partial y \le 0$. □

Proof of Theorem 3.1. Just as in lemma 3.2, we only need to show that

$$\frac{\text{vol}(A^\Gamma_{x,y})}{\text{vol}(A^H(x,y))}$$

is nonincreasing.

Note that

$$\text{vol}(A^\Gamma_{x,y}) = \int_\Gamma d\omega \int_{\min\{x,\text{cut}(\omega)\}}^{\min\{y,\text{cut}(\omega)\}} A(r,\omega)\,dr,$$

where $\text{cut}(\omega)$ is the distance to the cut locus in the direction $\omega \in S_p^{n-1}$.

Since $A(r,\omega)/A^H(r)$ is nonincreasing for any ω and $r < \text{cut}(\omega)$, Lemma 3.2 implies that, for $z \ge y$,

$$\frac{\int_{\min\{x,\text{cut}(\omega)\}}^{\min\{y,\text{cut}(\omega)\}} A(r,\omega)\,dr}{\int_{\min\{x,\text{cut}(\omega)\}}^{\min\{y,\text{cut}(\omega)\}} A^H(r)\,dr} \ge \frac{\int_{\min\{x,\text{cut}(\omega)\}}^{\min\{z,\text{cut}(\omega)\}} A(r,\omega)\,dr}{\int_{\min\{x,\text{cut}(\omega)\}}^{\min\{z,\text{cut}(\omega)\}} A^H(r)\,dr},$$

that is,

$$\int_{\min\{x,\mathrm{cut}(\omega)\}}^{\min\{y,\mathrm{cut}(\omega)\}} A(r,\omega)\,dr \geq \frac{\int_{\min\{x,\mathrm{cut}(\omega)\}}^{\min\{y,\mathrm{cut}(\omega)\}} A^H(r)\,dr}{\int_{\min\{x,\mathrm{cut}(\omega)\}}^{\min\{z,\mathrm{cut}(\omega)\}} A^H(r)\,dr} \cdot \int_{\min\{x,\mathrm{cut}(\omega)\}}^{\min\{z,\mathrm{cut}(\omega)\}} A(r,\omega)\,dr$$
$$\geq \frac{\int_{x}^{\min\{y,\mathrm{cut}(\omega)\}} A^H(r)\,dr}{\int_{x}^{\min\{z,\mathrm{cut}(\omega)\}} A^H(r)\,dr} \cdot \int_{\min\{x,\mathrm{cut}(\omega)\}}^{\min\{z,\mathrm{cut}(\omega)\}} A(r,\omega)\,dr$$
$$\geq \frac{\int_x^y A^H(r)\,dr}{\int_x^z A^H(r)\,dr} \cdot \int_{\min\{x,\mathrm{cut}(\omega)\}}^{\min\{z,\mathrm{cut}(\omega)\}} A(r,\omega)\,dr,$$

where the last inequality follows by considering the three possibilities $\mathrm{cut}(\omega) \leq y \leq z$, $y \leq \mathrm{cut}(\omega) \leq z$, and $y \leq z \leq \mathrm{cut}(\omega)$. The inequality before that uses the fact that $\int_x^a A^H(r)\,dr / \int_x^b A^H(r)\,dr$ is nonincreasing when $a < b$. Integrate the above over Γ, and we get

$$\mathrm{vol}(A^\Gamma_{x,y}) \geq \frac{\int_x^y A^H(r)\,dr}{\int_x^z A^H(r)\,dr} \cdot \mathrm{vol}(A^\Gamma_{x,z}) = \frac{\mathrm{vol}(A^H(x,y))}{\mathrm{vol}(A^H(x,z))} \cdot \mathrm{vol}(A^\Gamma_{x,z}).$$

The equality part follows from the equality discussion in Theorem 2.2. □

By taking $s = r = 0$ and $\Gamma = S^{n-1}$, one gets the following frequently used corollary.

Corollary 3.3. (1) Gromov's relative volume comparison theorem:

$$\frac{\mathrm{vol}(B_p(r))}{\mathrm{vol}(B_p(R))} \geq \frac{\mathrm{vol}(B^H(r))}{\mathrm{vol}(B^H(R))}.$$

(2) Bishop volume comparison theorem:

$$\mathrm{vol}(B_p(r)) \leq \mathrm{vol}(B^H(r)).$$

In both cases, equality holds if and only if $B_p(r)$ is isometric to $B^H(r)$.

We will give two applications of this result.

Theorem 3.4. *Assume* $\mathrm{Ric} \geq (n-1)H > 0$.
(1) Myers' Theorem [1935]: $\mathrm{diam}(M) \leq \pi/\sqrt{H}$, *and $\pi_1(M)$ is finite.*
(2) Cheng's Maximal Diameter Sphere Theorem [1975]: *If in addition $\mathrm{diam}(M) = \pi/\sqrt{H}$, then M^n is isometric to $S^n(H)$.*

Proof. Without loss of generality, we will assume $H = 1$.

(1) The classical proof of Myers' theorem is through second variation of geodesics, but one can easily see that it also follows from volume comparison. We will now use Theorem 2.2(2) to prove this.

Let p, q be such that $d(p,q) > \pi$, and let γ be a minimal geodesic from p to q. Since γ is minimal, $\gamma(\pi)$ is outside the cut locus of p; thus d_p is smooth at $\gamma(\pi)$, and $\triangle d_p \leq (n-1)\cot d_p$ by Theorem 2.2. Let $d_p \to \pi$ from the left. Then

$\triangle d_p \le -\infty$. This is a contradiction, since the left-hand side is a finite number. Therefore the diameter of M is at most π.

Using the same argument for the universal cover of M, we conclude that the universal cover also has diameter less than π, and thus is compact. It then follows that $\pi_1(M)$ is finite.

(2) Cheng's original proof used an eigenvalue comparison theorem; we will use the volume comparison theorem Corollary 3.3 to give a more geometric argument. The first such proof in print seems to be in [Shiohama 1983].

Let $p, q \in M$ be such that $d(p,q) = \pi$. Consider two balls $B_p(\frac{\pi}{2})$ and $B_p(\frac{\pi}{2})$. If the interiors of the two balls intersect, then there is a point x in the intersection such that $d(x,p) < \frac{\pi}{2}$ and $d(x,q) < \frac{\pi}{2}$; therefore $d(p,q) \le d(x,p) + d(x,q) < \pi$, a contradiction. Thus the two balls do not intersect in the interior. It follows that

$$\begin{aligned} \operatorname{vol}(M) &\ge \operatorname{vol}(B_p(\tfrac{\pi}{2})) + \operatorname{vol}(B_q(\tfrac{\pi}{2})) \\ &\ge \operatorname{vol}(B_p(\pi)) \cdot \frac{\operatorname{vol}(B^1(\frac{\pi}{2}))}{\operatorname{vol}(B^1(\pi))} + \operatorname{vol}(B_q(\pi)) \cdot \frac{\operatorname{vol}(B^1(\frac{\pi}{2}))}{\operatorname{vol}(B^1(\pi))} \\ &= \operatorname{vol}(M) \cdot \tfrac{1}{2} + \operatorname{vol}(M) \cdot \tfrac{1}{2} = \operatorname{vol}(M). \end{aligned}$$

Thus all inequalities are equalities. In particular, equality holds in the volume comparison. Therefore M has constant curvature 1, and $\operatorname{vol}(B_p(\frac{\pi}{2})) = \operatorname{vol}(B^1(\frac{\pi}{2}))$. It then follows M is simply connected and therefore isometric to $S^n(1)$. □

Remark. Myers' theorem is almost the only statement we can make about the fundamental group of manifolds with positive Ricci curvature. One conclusion one can draw from it is that the connected sum of two non-simply connected manifolds does not support any metric with positive Ricci curvature. The question remains open for the connected sum of simply connected manifolds.

We now turn to the second application of the relative volume comparison theorem, which was originally proved by analytic methods by Calabi and Yau.

Theorem 3.5 [Yau 1976]. *If M^n is complete and noncompact with nonnegative Ricci curvature, then* $\operatorname{vol}(B_p(r)) \ge cr$ *for some* $c > 0$.

Remark. This, together with Bishop's theorem, gives the growth of the volume of geodesic balls in noncompact manifolds with nonnegative Ricci curvature as $cr \le \operatorname{vol}(B_p(r)) \le \omega_n r^n$, where ω_n is the volume of the n-dimensional unit disc.

Proof. Since M is noncompact, there is a ray γ with $\gamma(0) = p$. By the relative volume comparison theorem for an annulus, we have

$$\frac{\operatorname{vol}(B_{\gamma(t)}(t-1))}{\operatorname{vol}(A_{\gamma(t)}(t-1, t+1))} \ge \frac{\omega_n (t-1)^n}{\omega_n (t+1)^n - \omega_n (t-1)^n} = c(n)t;$$

therefore

$$\operatorname{vol}(B_{\gamma(t)}(t-1)) \ge c(n) \operatorname{vol}(A_{\gamma(t)}(t-1,\, t+1))t \ge c(n) \operatorname{vol}(B_p(1))t = c(M)t,$$

and then $\mathrm{vol}(B_p(t)) \ge \mathrm{vol}(B_{\gamma(t/2)}(t/2-1)) \ge ct$. □

Packing and Gromov's precompactness theorem. One of the most useful consequences of a lower bound on Ricci curvature is the following.

LEMMA 3.6 (PACKING LEMMA [GROMOV 1981C]). *Let M^n be such that* $\mathrm{Ric} \ge (n-1)H$. *Given $r, \varepsilon > 0$, and $p \in M$, there is a covering of $B_p(r)$ by balls $B_{p_i}(\varepsilon)$, where $p_i \in B_r(p)$, such that the number of balls satisfies $N \le C_1(n, Hr^2, r/\varepsilon)$ and the multiplicity of the covering is bounded by $C_2(n, H\varepsilon^2)$.*

PROOF. Take a maximal set of points $p_i \in B_p(r-\varepsilon/2)$ such that $\mathrm{dist}(p_i, p_j) \ge \varepsilon/2$ for $i \ne j$. It then follows that $B_{p_i}(\varepsilon/4) \cap B_{p_j}(\varepsilon/4) = \varnothing$. Therefore

$$\begin{aligned} N &\le \frac{\mathrm{vol}(B_p(r))}{\min_i \mathrm{vol}(B_{p_i}(\varepsilon/4))} \\ &= \frac{\mathrm{vol}(B_p(r))}{\mathrm{vol}(B_{p_0}(\varepsilon/4))} \quad \text{for some } p_0 \\ &\le \frac{\mathrm{vol}(B_{p_0}(2r))}{\mathrm{vol}(B_{p_0}(\varepsilon/4))} \quad \text{since } B_p(r) \subset B_{p_0}(2r) \\ &\le \frac{\mathrm{vol}(B^H(2r))}{\mathrm{vol}(B^H(\varepsilon/4))} = C_1(n, Hr^2, r/\varepsilon). \end{aligned}$$

Next, if $B_{p_i}(\varepsilon/4) \cap B_{p_0}(\varepsilon/4) \ne \varnothing$, then $\mathrm{dist}(p_i, p_0) \le 2\varepsilon$. Then the disjointness of $B_{p_i}(\varepsilon/4)$ and $B_{p_j}(\varepsilon/4)$ implies

$$\begin{aligned} \text{multiplicity} &\le \frac{\mathrm{vol}(B_{p_0}(2\varepsilon+\varepsilon))}{\min_i \mathrm{vol}(B_{p_i}(\varepsilon/4))} \\ &= \frac{\mathrm{vol}(B_{p_0}(3\varepsilon))}{\mathrm{vol}(B_{p_1}(\varepsilon/4))} \quad \text{for some } p_1 \\ &\le \frac{\mathrm{vol}(B_{p_0}(5\varepsilon))}{\mathrm{vol}(B_{p_1}(\varepsilon/4))} \quad \text{since } B_{p_0}(3\varepsilon) \subset B_{p_1}(5\varepsilon) \\ &\le \frac{\mathrm{vol}(B^H(5\varepsilon))}{\mathrm{vol}(B^H(\varepsilon/4))} = C_2(n, H\varepsilon^2). \end{aligned}$$

□

It is easy to construct examples to see this is not true if one drops the curvature condition—for example, by connecting two spheres with many thin tunnels.

The packing lemma says that under the assumption of Ricci curvature bounded below, there are only finitely many local intersection patterns on a fixed scale. This is made precise by introducing the Hausdorff distance between metric spaces, which induces a very coarse topology on the space of all compact metric spaces.

DEFINITION 3.7. Let X, Y be two compact metric spaces. A map $\phi : X \to Y$ is called an ε-Hausdorff approximation if the ε-neighborhood of $\phi(X)$ is equal to Y, and $|d(x_1, x_2) - d(\phi(x_1), \phi(x_2))| < \varepsilon$, for any $x_1, x_2 \in X$.

We can then define the Hausdorff distance between two compact metric spaces X, Y, denoted by $d_H(X,Y)$, to be the infimum of all ε such that there is a ε-Hausdorff approximation from X to Y and vice versa.

THEOREM 3.8 (GROMOV'S PRECOMPACTNESS THEOREM [1981c]). *The set of n-dimensional Riemannian manifolds satisfying* $\operatorname{Ric} \geq (n-1)H$ *and* $\operatorname{diam} \leq D$ *is precompact with respect to the Hausdorff topology.*

PROOF. By the packing lemma, for any M^n satisfying the conditions and any $j > 0$, there is a subset $x_1, x_2, \dots, x_{N_j}$ with $N_j \leq N(n, D, H, j)$, which is $1/j$ dense and has diameter less than D. Thus, for any manifold under consideration, we have a sequence

$$M^1 \subset M^{1/2} \subset M^{1/3} \subset \cdots$$

such that $M^{1/j}$ is $(1/j)$-dense in M, and $|M^{1/j}| \leq N_1 + N_2 + \cdots + N_j$.

Let M_α be an infinite sequence. Since each $\{M_\alpha^1\}$ has N_1 elements and has bounded diameter, the precompactness of bounded sets in $\mathbb{R}^{N_1}$ implies the existence of a subsequence, still denoted by M_α^1, whose elements have pairwise distance less than 1. Thus the corresponding manifolds $M_{\alpha,1}$ have pairwise Hausdorff distance less than 1. For each $\{M_{\alpha,1}\}$, we can again find a subsequence $M_\alpha^{1/2}$ with pairwise distance less than $\frac{1}{2}$, and therefore a subsequence of manifolds $M_{\alpha,1/2}$, with pairwise Hausdorff distance less than $\frac{1}{2}$. Proceeding in this fashion, and using the diagonal argument, we will get a convergent subsequence. This proves that the set is precompact. □

Growth of fundamental groups. Volume comparison is most often used to get information about the fundamental group. The key idea is that the condition on Ricci curvature is local, and so can be lifted to the universal covering space, and control over the volume will give control over the relative size of fundamental domains. In this section, we will begin this study by examining the size of the fundamental group, as measured by its growth.

Let G be a finitely generated group, $G = \langle g_1, g_2, \dots, g_k \rangle$, and define the r-neighborhood with respect to the set of generators $g = \{g_1, g_2, \dots, g_k\}$ as

$$U_g(r) = \{g \in G \mid g \text{ is a word of length} \leq r\} = \{g \in G \mid g = g_1^{i_1} \cdots g_k^{i_k}, \Sigma |i_j| \leq r\}.$$

DEFINITION 3.9. G has *polynomial growth* if there exists a set of generators g and a positive number s such that $|U_g(r)| \leq r^s$ for r large. G has *exponential growth* if there exists a set of generators g and a positive number c such that $|U_g(r)| \geq e^{cr}$ for r large.

In this definition, we may take for g *any* set of generators. Indeed, if $g = \{g_1, g_2, \dots, g_k\}$ and $h = \{h_1, h_2, \dots, h_l\}$ are two sets of generators, there are constants r_0, s_0 such that $h_i \in U_g(r_0)$ and $g_i \in U_h(s_0)$, so that $U_h(s) \subset U_g(r_0 s)$ and $U_g(s) \subset U_h(s_0 s)$. Therefore, if G is of polynomial growth because g satisfies an inequality as in the definition, h also satisfies such an inequality; and likewise for exponential growth. In fact, this shows that the quantity

$$\operatorname{ord}(G) = \liminf \frac{\ln |U_g(r)|}{r}$$

does not depend on g; we call it the *growth order* of G.

It is easy to check that the free abelian groups $\mathbb{Z}^k$ have polynomial growth of order k. In general, the order is related to the degree of commutativity of the group. One of the most striking results is the following characterization of groups with polynomial growth.

THEOREM 3.10 [Gromov 1981b]. *A group has polynomial growth if and only if it is almost nilpotent*, *that is*, *contains a nilpotent subgroup of finite index*.

REMARK. It was shown by Milnor [1968b] that solvable groups have exponential growth unless they are polycyclic.

THEOREM 3.11 [Milnor 1968a]. *If M^n is a complete n-dimensional manifold with* Ric ≥ 0, *then any finitely generated subgroup of $\pi_1(M)$ has polynomial growth of order at most n.*

PROOF. Take the universal cover $(\tilde{M}, \tilde{p}) \to (M, p)$ with the pullback metric. We identify $\pi_1(M)$ with the group of deck transformations of $\tilde{M}$. Let $H = \langle g_1, \dots, g_k \rangle$ be a finitely generated subgroup of G. Then g_i can each be represented by a loop σ_i at p with length l_i. Let $\tilde{\sigma}_i$ be the liftings of σ_i at $\tilde{p}$; then, as deck transformations, we have $g_i(\tilde{p}) = \tilde{\sigma}_i(l_i)$. Let

$$\varepsilon < \min\{l_1, \dots, l_k\}, \quad l = \max\{l_1, \dots, l_k\}.$$

Then, for any distinct $h_1, h_2 \in H$, we have $h_1(B_{\tilde{p}}(\varepsilon)) \cap h_2(B_{\tilde{p}}(\varepsilon)) = \varnothing$, and $\bigcup_{h \in U(r)} h(B_{\tilde{p}}(\varepsilon)) \subset B_{\tilde{p}}(rl + \varepsilon)$. Thus

$$|U(r)| \cdot \operatorname{vol}(B_{\tilde{p}}(\varepsilon)) = \sum_{h \in U(r)} \operatorname{vol}(h(B_{\tilde{p}}(\varepsilon)) \leq \operatorname{vol}(B_{\tilde{p}}(rl + \varepsilon)),$$

and also

$$|U(r)| \leq \frac{\operatorname{vol}(B_{\tilde{p}}(rl + \varepsilon))}{\operatorname{vol}(B_{\tilde{p}}(\varepsilon))} = \frac{\omega_n}{\operatorname{vol}(B_{\tilde{p}}(\varepsilon))}(rl + \varepsilon)^n \leq cr^n. \qquad \square$$

REMARK. If the sectional curvature is nonpositive, one can also show that $\pi_1(M)$ has exponential growth. Note however, that Ric ≤ 0 is not enough.

Theorems 3.11 and 3.12 together show that, for manifolds M with Ric ≥ 0, any finitely generated subgroup of $\pi_1(M)$ is almost nilpotent. We can ask whether the converse is true. It was shown by Wei [1995] that any torsion-free nilpotent group is the fundamental group of some manifold with positive Ricci curvature, although the growth rate of the examples are far from optimal.

We end this section with a conjecture of Milnor [1968a]:

Conjecture 3.12. *If M is complete with* $\operatorname{Ric}(M) \geq 0$, *then* $\pi_1(M)$ *is finitely generated.*

Short loops and the first Betti number. Because of the relation between the fundamental group and the first Betti number given by the Hurewicz Theorem [Whitehead 1978], Ricci curvature can also give control on the first Betti number. This can also be seen through the Bochner technique. In this section, we will prove the following theorem.

Theorem 3.13 [Gromov 1981c; Gallot 1983]. *If M^n is such that* $\operatorname{Ric} \geq (n-1)H$ *and* $\operatorname{diam}(M) \leq d$, *then* $b_1(M,R) \leq c(n, Hd^2)$ *and* $\lim_{Hd^2 \to 0} c(n, Hd^2) = n$.

Remark. If Ricci curvature is replaced by sectional curvature, then the celebrated Betti number theorem of Gromov [1981a] shows that all higher Betti numbers can be bounded by sectional curvature and diameter. It was shown by Sha and Yang [1989a] that such an estimate is not true for Ricci curvature. Also note that in the above theorem, we used homology with real coefficients. The Betti number theorem of [Gromov 1981a] works for coefficients in any field. It is not known whether Theorem 3.13 is true for finite fields.

Remark. The second part of the theorem should be compared with earlier results. Recall that Bochner showed that if $\operatorname{Ric} > 0$ then $b_1(M, \mathbb{R}) = 0$. Using an extension of Bochner's techniques, Gallot proved that if $\operatorname{Ric} \geq 0$ then $b_1(M, \mathbb{R}) \leq n$. For the Bochner Technique, see [Wu 1988; Berard 1988]. We will present the geometric proof due to Gromov [1981c].

Proof. We first show that it is sufficient to prove that there is a finite cover $\hat{M} \to M$ such that $\pi_1(\hat{M})$ has at most $c(n, Hd^2)$ generators. In fact, if $G' = \pi_1(\hat{M}) = \langle \gamma_1, \ldots, \gamma_k \rangle$, and $G = \pi_1(M)$, then $|G/G'| = m < \infty$, i.e., there are $g_1, \ldots, g_m$ such that $g_i^m \in G'$ and G decomposes into left cosets as $G = g_1 G' \cup g_2 G' \cup \cdots \cup g_m G'$. Consider the Hurewicz map

$$G \xrightarrow{f} G/[G,G] \xrightarrow{i} H_1(M, \mathbb{R}).$$

Since $\{\gamma_1, \ldots, \gamma_k, g_1, \ldots, g_m\}$ generates G, $\{f(\gamma_1), \ldots, f(\gamma_k), f(g_1), \ldots, f(g_m)\}$ generates $G/[G,G]$. But

$$m \cdot (i \circ f)(g_i) = (i \circ f)(g_i^m) = (i \circ f)(h)$$

for some $h \in G'$, and $(i \circ f)(g_i)$ can be generated by $\{(i \circ f)(\gamma_i)\}$. Therefore the set $\{(i \circ f)(\gamma_1), \ldots, (i \circ f)(\gamma_k)\}$ generates $H_1(M, \mathbb{R})$. Thus, to bound $b_1(M, \mathbb{R})$, it is sufficient to bound the number of generators for $\pi_1(\hat{M})$.

We will now construct a finite cover $\hat{M}$ and, at the same time, give a set of generators whose size can be bounded by Ricci curvature and diameter. To this end, let $\tilde{\pi} : \tilde{M} \to M$ be the universal cover. Fix $\tilde{x}_0 \in \tilde{M}$ with $\tilde{\pi}(\tilde{x}_0) = x_0$ and $\varepsilon > 0$. Define $\|g\| = d(\tilde{x}_0, g(\tilde{x}_0))$. Take a maximal set of elements $\{g_1, \ldots, g_k\}$ of $\pi_1(M)$ such that $\|g_i\| \leq 2d + \varepsilon$, and $\|g_i g_j^{-1}\| \geq \varepsilon$, for $i \neq j$.

Let Γ be a subgroup of $\pi_1(M)$ generated by $\{g_i\}_1^k$, and let $\hat{\pi} : \hat{M} \to M$ be the covering of M with $\pi_1(\hat{M}) = \Gamma$. We need to show that $\hat{\pi}$ is a finite cover, and to give a bound for the number k.

To show $\hat{\pi}$ is finite, we will show that $\operatorname{diam}(\hat{M}) \le 2d + 2\varepsilon$. Let $\hat{x}_0 \in \hat{M}$ be such that $\hat{\pi}(\hat{x}_0) = x_0$. If $\operatorname{diam}(\hat{M}) > 2d + 2\varepsilon$, then there is a point $\hat{z} \in \hat{M}$ with $d_{\hat{M}}(\hat{x}_0, \hat{z}) = d + \varepsilon$. But $d_M(x_0, \hat{\pi}(\hat{z})) \le \operatorname{diam}(M) = d$; therefore there is a deck transformation $\alpha \in \pi_1(M) \setminus \Gamma$ such that $d_{\hat{M}}(\hat{z}, \alpha\hat{x}_0) \le d$. Then

$$\begin{aligned} d_{\hat{M}}(\hat{x}_0, \alpha\hat{x}_0) &\ge d_{\hat{M}}(\hat{x}_0, \hat{z}) - d_{\hat{M}}(\hat{z}, \alpha\hat{x}_0) \ge \varepsilon, \\ d_{\hat{M}}(\hat{x}_0, \alpha\hat{x}_0) &\le d_{\hat{M}}(\hat{x}_0, \hat{z}) + d_{\hat{M}}(\hat{z}, \alpha\hat{x}_0) \le 2d + \varepsilon. \end{aligned}$$

Note there is a $\beta \in \pi_1(\hat{M}) = \Gamma$ such that

$$d_{\tilde{M}}(\Gamma\tilde{x}_0, \alpha\tilde{x}_0) = d_{\tilde{M}}(\beta\tilde{x}_0, \alpha\tilde{x}_0).$$

Therefore

$$\begin{aligned} \|\beta^{-1}\alpha\| &= d_{\tilde{M}}(\tilde{x}_0, \beta^{-1}\alpha\tilde{x}_0) = d_{\tilde{M}}(\beta\tilde{x}_0, \alpha\tilde{x}_0) \\ &= d_{\tilde{M}}(\Gamma\tilde{x}_0, \alpha\tilde{x}_0) = d_{\hat{M}}(\hat{x}_0, \alpha\hat{x}_0) \le 2d + \varepsilon. \end{aligned}$$

Furthermore, for any $g \in \pi_1(\hat{M}) = \Gamma$, we have

$$\begin{aligned} \|g^{-1}\beta^{-1}\alpha\| &= d_{\tilde{M}}(\tilde{x}_0, g^{-1}\beta^{-1}\alpha\tilde{x}_0) = d_{\tilde{M}}(\beta g\tilde{x}_0, \alpha\tilde{x}_0) \\ &\ge d_{\tilde{M}}(\Gamma\tilde{x}_0, \alpha\tilde{x}_0) = d_{\hat{M}}(\hat{x}_0, \alpha\hat{x}_0) \ge \varepsilon. \end{aligned}$$

Thus $\beta^{-1}\alpha$ should also be in Γ, since Γ is maximal. But α is not in Γ, so this is a contradiction. Therefore $\operatorname{diam}(\hat{M}) \le 2d + 2\varepsilon$.

We now bound k. Since $\varepsilon \le \|g_i^{-1}g_j\| = d_{\tilde{M}}(g_i^{-1}g_j\tilde{x}_0, \tilde{x}_0) = d_{\tilde{M}}(g_j\tilde{x}_0, g_i\tilde{x}_0)$, we have

$$B^{\tilde{M}}_{g_i\tilde{x}_0}(\varepsilon/2) \cap B^{\tilde{M}}_{g_j\tilde{x}_0}(\varepsilon/2) = \varnothing \quad \text{for } i \ne j$$

and

$$\bigcup_{i=1}^k B^{\tilde{M}}_{g_i\tilde{x}_0}(\varepsilon/2) \subset B^{\tilde{M}}_{\tilde{x}_0}(2d + 3\varepsilon/2).$$

Therefore, choosing $d = \varepsilon/2$, we get

$$k = \frac{\operatorname{vol}(\bigcup_1^k B^{\tilde{M}}_{g_i\tilde{x}_0}(d))}{\operatorname{vol}(B^{\tilde{M}}_{\tilde{x}_0}(d))} \le \frac{\operatorname{vol}(B^{\tilde{M}}_{\tilde{x}_0}(2d + 3d))}{\operatorname{vol}(B^{\tilde{M}}_{\tilde{x}_0}(d))} \le \frac{\operatorname{vol}^H(B(2d + 3d))}{\operatorname{vol}^H(B(d))} = c(n, Hd^2).$$

We are now going to refine the preceding argument by choosing longer loops to generate $H_1(M, \mathbb{R})$.

Again, let $\phi : G \to H_1(M, \mathbb{R})$ be the Hurewicz map, and let $\{g_1, \dots, g_k\}$ be chosen as before, with $\varepsilon = 2d$. Then $\{\phi(g_1), \dots, \phi(g_k)\}$ is a basis for $H_1(M, \mathbb{R})$. Let $\Gamma = \langle g_1, \dots, g_k\rangle$. If Γ contains an element γ such that $\|\gamma\| < 2d$ and $\phi(\gamma) \ne 0$, then $\phi(\gamma)$ is of infinite order, and there exists an $m > 0$ such that $2d \le \|\gamma\| \le 4d$. Since

$$\phi(\gamma) = a_1\phi(g_1) + \cdots + a_k\phi(g_k) \ne 0,$$

we can assume, without loss of generality, that $a_1 \neq 0$. Let $\Gamma_1 = \langle \gamma^m, g_2, \dots, g_k \rangle$. Obviously, $\phi(\gamma^m), \phi(g_2), \dots, \phi(g_k)$ form a basis for $H_1(M, \mathbb{R})$. Furthermore, $\gamma \notin \Gamma_1$. In fact, if

$$\gamma = b_1 \cdot (\gamma^m)^{k_1} \cdot b_2 \cdot (\gamma^m)^{k_2} \cdots (\gamma^m)^{k_l} \cdot b_{l+1},$$

where the b_i's are words in $\{g_2, \dots, g_k\}$, then

$$\begin{aligned} \phi(\gamma) &= \phi(b_1) + \phi(\gamma^{mk_1}) + \cdots + \phi(\gamma^{mk_l}) \\ &= \phi(b_1 b_2 \cdots b_{l+1}) + m(k_1 + k_2 + \cdots + k_l)\phi(\gamma), \end{aligned}$$

that is,

$$\phi(b_1 b_2 \cdots b_{l+1}) = (1 - m(k_1 + k_2 + \cdots + k_l))\phi(\gamma).$$

Since $m \geq 2$, the coefficient is not zero; therefore

$$\phi(\gamma) = \frac{1}{1 - m(k_1 + k_2 + \cdots + k_l)} \phi(b_1 b_2 \cdots b_{l+1}) = a_2 \phi(g_2) + \cdots + a_k \phi(g_k),$$

which contradicts $a_1 \neq 0$.

Thus, each time we have an element $\gamma \in \Gamma$ with $\|\gamma\| < 2d$ and $\phi(\gamma) \neq 0$, we can replace it by γ^m such that $\|\gamma^m\| \geq 2d$, and still have a basis. By repeating this process a finite number of times (since $\pi_1(M)$ contains finitely many γ with $\|\gamma\| < 2d$), we get a set, still denoted by $\{g_1, \dots, g_k\}$, such that
(1) $\{\phi(g_1), \dots, \phi(g_k)\}$ is a basis for $H_1(M, \mathbb{R})$;
(2) $2d \leq \|g_i\| \leq 4d$; and
(3) $\|g\| < 2d$ for every element $g \in \langle g_1, \dots, g_k \rangle$ with $\phi(g) \neq 0$.

Let

$$U(N) = \left\{ g \in H_1(M, \mathbb{R}) \,\middle|\, g = \sum r_i \phi(g_i), \sum |r_i| \leq N \right\}.$$

By (3), $\|g^{-1}h\| \geq 2d$, which implies that for all $g \in U(N)$ such that $\phi(g) \neq 0$, the balls $B_{g(\tilde{x}_0)}(d)$ are disjoint. Furthermore,

$$\bigcup_{g \in U(N)} B_{g(\tilde{x}_0)}(d) \subset B_{\tilde{x}_0}(2Nd + 4d).$$

By taking the volume on both sides, we get

$$\begin{aligned} |U(N)| &= \frac{\operatorname{vol}(\bigcup_{g \in U(N)} B_{g(\tilde{x}_0)}(d))}{\operatorname{vol}(B_{\tilde{x}_0}(d))} \\ &\leq \frac{\operatorname{vol}(B_{\tilde{x}_0}(2Nd + 4d))}{\operatorname{vol}(B_{\tilde{x}_0}(d))} \leq \frac{\operatorname{vol}^H(B(2Nd + 4d))}{\operatorname{vol}^H(B(d))} \\ &= \frac{\int_0^{2Nd+4d} \frac{1}{\sqrt{H}} \sinh^{n-1} \sqrt{H} t \, dt}{\int_0^d \frac{1}{\sqrt{H}} \sinh^{n-1} \sqrt{H} t \, dt} = \frac{\int_0^{(2N+4)d\sqrt{H}} \sinh^{n-1} t \, dt}{\int_0^{d\sqrt{H}} \sinh^{n-1} t \, dt}, \end{aligned}$$

which is bounded by cN^n when Hd^2 is small. This implies that $b_1(M, \mathbb{R}) \leq n$. $\square$

Short geodesics and the fundamental group. We now consider a bigger class of manifolds, the set $\mathcal{M}$ of n-dimensional manifolds satisfying

$$\mathrm{Ric} \geq -(n-1)\Lambda^2,$$

diam $\leq D$, and vol $\geq v > 0$. Note that the first condition will not yield any restriction on π_1, and the first two conditions allow infinitely many isomorphism classes of π_1 (as the example of lens spaces shows). In this section, we will prove the following theorem.

THEOREM 3.14 [Anderson 1990a]. *The class* $\mathcal{M}$ *contains only finitely many isomorphism classes of* π_1.

To control the size of the fundamental group is to count the number of fundamental domains, as was shown in the previous sections. For this to work, the fundamental domain should not be too thin, i.e., it should contain a geodesic ball of size bounded from below. This is controlled by the first systol, defined as

$$\mathrm{sys}^1(M,g) = \inf\{\mathrm{length}(\gamma) \mid \gamma \text{ is noncontractible}\}.$$

The following result gives an estimate of most noncontractible curves for the class $\mathcal{M}$.

LEMMA 3.15. *For any* Λ, v, D, *there exist positive numbers* $N(n,\Lambda,v,D)$ *and* $L(n,\Lambda,v,D)$ *such that if* $M \in \mathcal{M}$ *and* $[\gamma] \in \pi_1(M)$ *have order* $\geq N$, *then*

$$\mathrm{length}(\gamma) \geq L.$$

PROOF. Let $\Gamma \subset \pi_1(M,x_0)$ be the subgroup generated by γ, so that $|\Gamma| \geq N$. Let $\pi : \tilde{M} \to M$ be the universal cover, and let $F \subset \tilde{M}$ be a fundamental domain with $\tilde{x}_0 \in F$. Then $B_{\tilde{x}_0}(r) \cap F$ is mapped isometrically by π onto $B_{x_0}(r)$, modulo a set of measure zero corresponding to the boundary of F. In particular, $\mathrm{vol}(B_{\tilde{x}_0} \cap F) = \mathrm{vol}(B_{x_0}(r))$.

Let $U(r) = \{g \in \Gamma \mid g = \gamma^i,\ |i| < r\}$. Note that $d_{\tilde{M}}(\gamma\tilde{x}_0, \tilde{x}_0) \leq \mathrm{length}(\gamma) = l(\gamma)$; therefore $d_{\tilde{M}}(g\tilde{x}_0, \tilde{x}_0) \leq 2N \cdot l(\gamma)$, for any $g \in U(N)$. Then

$$\bigcup_{g \in U(N)} g(B_{\tilde{x}_0}(D) \cap F) \subset B_{\tilde{x}_0}(2N \cdot l(\gamma) + D).$$

Taking the volume, we obtain

$$(2N+1)\,\mathrm{vol}(M) \leq \mathrm{vol}(B_{\tilde{x}_0}(2N \cdot l(\gamma) + D)) \leq \mathrm{vol}(B^\Lambda(2N \cdot l(\gamma) + D)).$$

If $l(\gamma) < D/2N$, then,

$$N \leq \frac{\mathrm{vol}(B_{\tilde{x}_0}(2N \cdot l(\gamma) + D))}{2\,\mathrm{vol}(M)} < \frac{\mathrm{vol}(B^\Lambda(2D))}{2v}.$$

Thus, if we set $N = N(n,\Lambda,D,v) = \lfloor \mathrm{vol}(B^\Lambda(2D))/2v \rfloor + 1$, we will have

$$l(\gamma) \geq \frac{D}{2N} = \frac{Dv}{\mathrm{vol}(B^\Lambda(2D))} = L(n,\Lambda,D,v).$$

□

Remark. In this proof we only considered a subgroup generated by one loop. The same argument applies to subgroups generated by several loops. We thus can prove that the subgroup of $\pi_1(M)$ generated by loops of length $\leq L$ must have order $\leq N$.

Remark. If the sectional curvature is bounded below by $-\Lambda^2$, one can bound the length of closed geodesics, which gives a lower bound on sys^1. This was proved by Cheeger [1970] in connection with his finiteness theorem.

Lemma 3.16 [Gromov 1981c]. *For any compact manifold M with diameter D, one can choose a set of generators $\{g_1, \dots, g_m\}$ of $\pi_1(M,p)$ and representative loops $\{\gamma_1, \dots, \gamma_m\}$ such that length $(\gamma_i) \leq 2D$ and all relations are of the form $g_i g_j g_k^{-1} = 1$.*

Remark. In this lemma, we do not have a bound on the number of generators. This number is in general very big, and thus such a representation is not efficient for other purposes.

Proof. Fix a constant ε smaller than the injectivity radius. Choose a triangulation K of M such that any n-simplex lies in a ball of radius less than ε. Let $\{x_i\}$ be the vertices and $\{e_{ij}\}$ the edges. Since $\mathrm{cut}(p)$ has measure zero, we can assume that all x_i are not in $\mathrm{cut}(p)$. Let γ_1 be the minimal geodesic from p to x_i, and set $\sigma_{ij} = \gamma_i e_{ij} \gamma_j^{-1}$. Then $\sigma_{ij} \in \pi_1(M,p)$ and the length of $\sigma_{i,j}$ is less than $2D+\varepsilon$. Given any loop σ at p, we can deform σ to lie in the one-skeleton of K. Thus, σ can be written as a product of σ_{ij}s. This shows that the σ_{ij} generate $\pi_1(M,p)$.

If Δ_{ijk} is a two-simplex with vertices x_i, x_j, x_k, we have

$$\sigma_{ij}\sigma_{jk} = \sigma_{ik}.$$

If $\sigma = e$ is a relation with σ a product of σ_{ij}s, the homotopy can be represented by a collection of two simplices (e.g., take simplicial approximations of σ and the homotopy.) Therefore it can be generated by the above set of relations.

Let g_{ij} be a geodesic loop at p in the homotopy class of σ_{ij}. Since the set of lengths of such loops form a discrete set, we can choose δ small enough such that there are no loops with length in $[2D,\, 2D+\delta]$. Hence if we further require $\varepsilon < \delta$, each g_{ij} has length at most $2D$. □

Proof of Theorem 3.14. By Lemma 3.5.2, we only need to bound the number of generators: in fact, if there are p generators as in Lemma 3.16, the relations can be chosen from among a finite number (p^3) of possibilities.

Let $\{g_1, \dots, g_p\}$ be such a set of generators. Fix $\tilde{x}_0 \in \tilde{M}$ in the universal cover. Consider $B_{g_i\tilde{x}_0}(L/2)$ with L as in Lemma 3.14. Since $\mathrm{length}(g_i) \leq 2D$, we have

$$\bigcup_{i=1}^{p} B_{g_i\tilde{x}_0}(L/2) \subset B_{\tilde{x}_0}(2D + L/2).$$

These balls can intersect, but if

$$B_{g_i\tilde{x}_0}(L/2)\cap B_{g_j\tilde{x}_0}(L/2)\neq\varnothing,$$

then $g_ig_j^{-1}$ is represented by a loop of length less than $2\cdot L/2=L$. By Lemma 3.14, there are at most N such loops. Therefore the balls $\{B_{g_i\tilde{x}_0}(L/2)\}$ have multiplicity bounded by N. Thus

$$\begin{aligned}\operatorname{vol}(B_{\tilde{x}_0}(2D+L/2)) &\geq \operatorname{vol}\biggl(\bigcup_{i=1}^{p} B_{g_i\tilde{x}_0}(L/2)\biggr)\geq \operatorname{vol}\biggl(\bigcup B_{g_j\tilde{x}_0}(L/2)\biggr)\\ &\geq \frac{p}{N}\operatorname{vol}(B_{x_0}(L/2)),\end{aligned}$$

where the second union is taken over the balls that do not intersect. It then follows that

$$p\leq N\cdot\frac{\operatorname{vol}(B_{\tilde{x}_0}(2D+L/2))}{\operatorname{vol}(B_{x_0}(L/2))}\leq N\cdot\frac{\operatorname{vol}(B^{\Lambda}(2D+L/2))}{\operatorname{vol}(B^{\Lambda}(L/2))}.\qquad\square$$

4. Laplacian Comparison and Its Applications

Weak maximum principle and regularity. It is clear now that in order to apply any analysis to the distance function, we have to either restrict ourselves to the complement of the cut locus, or to extend the analysis to Lipschitz functions. In this section we will extend the maximum principle to this situation. Much of this section is adapted from [Cheeger 1991].

DEFINITION 4.1. A *lower barrier* (or *support function*) for a continuous function f at the point x_0 is a C^2 function g, defined in a neighborhood of x_0, such that $g(x_0)=f(x_0)$ and $g(x)\leq f(x)$ in the neighborhood.

DEFINITION 4.2. If f is continuous, we say that $\triangle f\geq a$ at x_0 *in the barrier sense* if, for any $\varepsilon>0$, there is a barrier $f_{x_0,\varepsilon}$ of f at x_0 such that

$$\triangle f_{x_0,\varepsilon}\geq a-\varepsilon.$$

THEOREM 4.3 (WEAK MAXIMUM PRINCIPLE, HOPF–CALABI). *Let M be a connected Riemannian manifold and let $f\in C^0(M)$. Suppose that $\triangle f\geq 0$ in the barrier sense. Then f attains no weak local maximum value unless it is a constant function.*

PROOF. Let p be a weak local maximum, so that $f(p)\geq f(x)$ for all x_0 near p. Take a small normal coordinate ball $B_p(\delta)$, and assume that there exists a point $z\in\partial B_p(\delta)$ such that $f(p)>f(z)$. Then, by continuity, $f(p)>f(z')$ for $z'\in\partial B_p(\delta)$ sufficiently close to z. Choose a normal coordinate system $\{x_i\}$ such that $z=(\delta,0,\dots,0)$. Put $\phi(x)=x_1-d(x_2^2+\cdots+x_n^2)$, where d is a number so large that if $y\in\partial B_p(\delta)$ and $f(y)=f(p)$, then $\phi(y)<0$. Note that

$$\nabla\phi=\frac{\partial}{\partial x^1}-\cdots\neq 0.$$

Put $\psi = e^{a\phi} - 1$. Then $\triangle\psi = (a^2|\nabla\phi|^2 + a\triangle\phi)e^{a\phi}$. Thus, for a large enough, $\triangle\psi > 0$. Moreover, $\psi(p) = 0$. Thus, for $\eta > 0$ sufficiently small,

$$(f + \eta\psi)|_{\partial B_p(\delta)} < f(p), \quad (f + \eta\psi)(p) = f(p).$$

Therefore $f + \eta\psi$ has an interior maximum at some point $q \in B_p(\delta)$.

If $f_{q,\varepsilon}$ is a barrier for f at q with $\triangle f_{q,\varepsilon} \geq -\varepsilon$, then $f_{q,\varepsilon} + \eta\psi$ is also a barrier for $f + \eta\psi$ at q. For ε sufficiently small, we have $\triangle(f_{q,\varepsilon} + \eta\psi) > 0$. Since $f + \eta\psi$ has a local maximum at q, and

$$f_{q,\varepsilon} + \eta\psi < f + \eta\psi, \quad (f_{q,\varepsilon} + \eta\psi)(q) = (f + \eta\psi)(q),$$

we find that $f_{q,\varepsilon} + \eta\psi$ has a local maximum at q. This is not possible because $\triangle(f_{q,\varepsilon} + \eta\psi) > 0$.

It follows that for all small δ, we have $f|_{\partial B_p(\delta)} = f(p)$. Since M is connected, this implies that f is constant. □

The following regularity theorem is not necessary for the proof of later results, but it simplifies the proof of the Splitting Theorem considerably.

Theorem 4.4 (Regularity). *If $\triangle f = 0$ in the barrier sense, then f is smooth.*

Proof. Since regularity is a local property, this theorem follows from standard elliptic regularity [Gilbarg and Trudinger 1983, Theorem 6.17]. □

The Splitting Theorem. Recall that a geodesic $\gamma : [0, +\infty) \to M$ is a *ray* if $d(0, \gamma(t)) = t$ for all $t > 0$. A geodesic $\gamma : (-\infty, +\infty) \to M$ is a *line* if $d(\gamma(s), \gamma(t)) = |s - t|$ for all t, s. It is easy to see that if M is noncompact, it contains a ray. If it has at least two ends, it contains a line.

The purpose of this section is to prove the Splitting Theorem of Cheeger and Gromoll [1971]. We will first provide some preliminary properties of the Busemann functions.

Let σ be a ray, and define $b_r^\sigma : M \to \mathbb{R}$ by $b_r^\sigma(x) = r - d(x, \sigma(r))$.

Lemma 4.5. (1) *b_r^σ is increasing in r when x is fixed.*
(2) *b_r^σ is bounded by $d(x, \sigma(0))$.*
(3) *The family b_r^σ is uniformly continuous.*

Proof. (1) By the triangle inequality, for any $r > s$,

$$d(x, \sigma(r)) - d(x, \sigma(s)) \leq d(\sigma(r), \sigma(s)) = r - s,$$

so that $b_r^\sigma(x) \geq b_s^\sigma(x)$. (2) By the triangle inequality,

$$b_r^\sigma(x) = r - d(x, \sigma(r)) = d(\sigma(0), \sigma(r)) - d(x, \sigma(r)) \leq d(\sigma(0), x).$$

(3) We have $|b_r^\sigma(x) - b_r^\sigma(y)| = |d(x, \sigma(r)) - d(y, \sigma(r))| \leq d(x, y)$; uniform continuity follows. □

DEFINITION 4.6. The Busemann function associated to σ is defined as

$$b^\sigma(x) = \lim_{r\to\infty} b^\sigma_r(x) = \lim_{r\to\infty}(r - d(x, \sigma(r))).$$

LEMMA 4.7. *If* $\mathrm{Ric} \geq 0$ *and* σ *is a ray, then* $\triangle b^\sigma \geq 0$ *in the barrier sense.*

PROOF. Fix a point $p \in M$. We will construct a barrier for b^σ at p. For this, we first define the asymptote of σ. Take a sequence of points $t_i \to \infty$, and let δ_i be a minimizing geodesic from p to $\sigma(t_i)$. A subsequence of δ_i converges to a ray δ, which is called an asymptote of σ at p.

We now show that, for any $r > 0$, $b^\delta_r(x) + b^\sigma(p)$ is a barrier for b^σ at p.

In fact, since δ is a ray, $\delta(r)$ is not a cut point of p, hence p is not a cut point of $\delta(r)$, therefore b^δ_r is a smooth function near p. Furthermore, $(b^\delta_r(x)+b^\sigma(p))(p) = b^\sigma(p)$. We thus only need to prove that $b^\delta_r(x) + b^\sigma(p) \leq b^\sigma(x)$ near p.

For any $k > 0$, there exists a geodesic δ_k from p to $\sigma(t_k)$ such that

$$d(\delta(t), \delta_k(t)) \leq 1/k$$

for $t \in [0, t_k]$. Thus,

$$\begin{aligned}
b^\sigma_{t_k}(x) - b^\sigma_{t_k}(p) &= d(p, \sigma(t_k)) - d(x, \sigma(t_k)) \\
&= d(p, \delta_k(r)) + d(\delta_k(r), \sigma(t_k)) - d(x, \sigma(t_k)) \\
&\geq r + (d(\delta(r), \sigma(t_k)) - 1/k) - d(x, \sigma(t_k)) \geq r - 1/k - d(\delta(r), x).
\end{aligned}$$

As $k \to \infty$, we obtain $b^\sigma(x) - b^\sigma(p) \geq r - d(\delta(r), x) = b^\delta_r(x)$, that is, $b^\delta_r(x) + b^\sigma(p) \leq b^\sigma(x)$. This proves the claim.

Now we use the Laplacian comparison theorem to compute

$$\triangle(b^\delta_r(x) + b^\sigma(p)) = \triangle(r - d(\delta(r), x)) = -\triangle d(\delta(r), x) \geq -\frac{n-1}{d(\delta(r), x)} \geq -\varepsilon,$$

for $d(x, \delta(r))$ big enough. Thus $\triangle b^\sigma \geq 0$ in the barrier sense. □

If σ is a line, then we have two rays σ^+ and σ^-, and thus also two Busemann functions b^+ and b^-.

LEMMA 4.8. *If* $\mathrm{Ric} \geq 0$, *and* σ *is a line, then*
(1) $b^+ + b^- = 0$, *and* b^+, b^- *are smooth; and*
(2) *through every point in* M, *there is a unique line perpendicular to the set* $V_0 = \{b^+ = 0\}$.

PROOF. (1) By lemma 4.7, $\triangle(b^+ + b^-) \geq 0$. The triangle inequality implies $b^+ + b^- \leq 0$. Obviously $(b^+ + b^-)(\sigma(0)) = 0$. By Theorem 4.3, $b^+ + b^- = 0$.

This, together with $\triangle b^+ \geq 0$ and $\triangle b^- \geq 0$, implies that $0 \leq \triangle b^+ = -\triangle b^- \leq 0$. Thus, $\triangle b^+ = \triangle b^- = 0$, and is therefore smooth by the regularity theorem.

(2) Fix a point $p \in M$. We have at least two asymptotes δ^+, δ^-. Note that

$$\begin{aligned} b^+(\delta^+(t)) &= \lim_{r\to\infty}(r - d(\delta^+(t), \delta^+(r))) \\ &= \lim_{r\to\infty}(r - d(\delta^+(0), \delta^+(r))) + \lim_{r\to\infty}(d(\delta^+(0), \delta^+(r)) - d(\delta^+(t), \delta^+(r))) \\ &= b^+(\delta^+(0)) + t. \end{aligned}$$

Similarly, $b^-(\delta^-(s)) = b^s(\delta^s(0)) + s$.

Thus,

$$\begin{aligned} d(\delta^+(t), \delta^-(s)) &\geq b^+(\delta^+(t)) - b^+(\delta^-(s)) \\ &\geq b^+(\delta^+(t)) + b^-(\delta^-(s)) = b^+(\delta^+(0)) + t + b^-(\delta^-(0)) + s \\ &= t + s + b^+(p) + b^-(p) = t + s. \end{aligned}$$

Thus, δ^+, δ^- fit together to form a line. Since δ^+, δ^- are arbitrary asymptotes, they are unique.

Moreover, for any $y \in V_0$,

$$d(y, \delta^+(t)) \geq b^+(\delta^+(t)) - b^+(y) = b^+(\delta^+(0)) + t - b^+(y) = t.$$

Thus δ^+ is perpendicular to V_0. □

THEOREM 4.9 (SPLITTING THEOREM [Cheeger and Gromoll 1971]). *If M has nonnegative Ricci curvature, and contains a line, then M is isometric to the product $\mathbb{R} \times V$, for some $(n-1)$-dimensional Riemannian manifold V.*

PROOF. We can now define a map $\phi : V_0 \times \mathbb{R} \to M$ as

$$\phi(v, t) = exp_v(t\dot{\sigma}(0)),$$

where V_0 is as in Lemma 4.8, and σ is the unique line passing through $v \in V_0$. We claim that ϕ is an isometry.

To see that ϕ is a bijection, note that, for any $x \in M$, there is a line γ such that $\gamma \perp V_0$, $\gamma(0) \in V_0$, and $x = \gamma(t_0)$. Thus $\phi(\gamma(0), t_0) = x$, which shows that ϕ is surjective. If $\phi(v_1, t_1) = \phi(v_2, t_2) = x$, then $t_1 = t_2 = d(x, V_0)$. Now since σ_1 and σ_2 are lines, they cannot intersect unless they are the same, i.e., $v_1 = v_2$. Therefore ϕ is injective. Since σ is a line, $\exp_v$ is a local diffeomorphism. This implies that ϕ is a diffeomorphism.

To see that ϕ is an isometry, we let $S(t) = (b^+)^{-1}(t)$. Then the mean curvature of $S(t)$ is $m(t) = \triangle b^+ = 0$. On the other hand, letting $f = b^+$ in Theorem 2.1 (note that b^+ is smooth by Lemma 4.8) gives

$$m'(t) = \mathrm{Ric} + |\mathrm{Hess}(b^+)|^2,$$

which, together with $\mathrm{Ric} \geq 0$, gives

$$|\mathrm{Hess}(b^+)|^2 \leq 0.$$

Thus, $\mathrm{Hess}(b^+) = 0$, and $S(t)$ is totally geodesic. Therefore, for any X tangent to V_0 and $N = \nabla b^+$,

$$R(N, X)N = \nabla_N \nabla_X N - \nabla_X \nabla_N N - \nabla_{[X,N]} N = \nabla_N \nabla_X N = 0.$$

Let $J(t) = \phi_*(X)$. Then J is a variational vector field of geodesics, and satisfies the Jacobi equation, which, with vanishing curvature, says that $J''(t) = 0$. Thus J is a constant, implying $|\phi_*(X)| = |X|$, and hence ϕ is an isometry. □

From the proof of the Splitting Theorem, it seems natural to make the following conjecture.

Conjecture 4.10 (Local splitting). *Let M be complete, and let γ be a line. If M has nonnegative Ricci curvature in a neighborhood of γ, then a (smaller) neighborhood of γ splits as a product.*

The main difficulty in proving this conjecture is that in proving Lemma 4.7 and in using the Laplacian Comparison Theorem to conclude that $\triangle b \geq 0$, it is necessary that the asymptotes stay in the region where $\mathrm{Ric} \geq 0$. In general, this is very hard to achieve. In [Cai et al. 1994], it was proved that if the Ricci curvature is nonnegative outside of a compact set, then any line in this region will cause a splitting. Furthermore, under the condition $0 \leq \sec \leq c$, the local splitting conjecture is true.

We now prove two corollaries on the fundamental group of manifolds with nonnegative Ricci curvature. They generalize Myers's theorem to manifolds with $\mathrm{Ric} \geq 0$, and strengthen the result of Milnor (Theorems 3.11 and 3.4).

Corollary 4.11. *If M is compact with nonnegative Ricci curvature, there is a finite group F and a Bieberbach group B_k of $\mathbb{R}^n$ such that the sequence*

$$0 \to F \to \pi_1(M) \to B_k \to 0$$

is exact.

Proof. By Theorem 4.9, we may write the universal covering space $\tilde{M}$ of M as $N \times \mathbb{R}^k$, where N contains no lines. Since isometries map lines to lines, the covering transformations $\Gamma = \pi_1(M)$ (actually all isometries) are of the form $(f, g)(x, y) = (f(x), g(y))$ where $f : N \to N$ and $g : \mathbb{R}^k \to \mathbb{R}^k$ are isometries. Let ρ be the projection of $\tilde{M}$ on the first factor in $N \times \mathbb{R}^k$, and let F be a compact fundamental domain for Γ, which exists because M is compact. Then the orbit $\rho(F)$ under $\rho(\Gamma)$ is all of N. We claim that N is compact. Otherwise there exists a ray γ and a sequence $g_i \in \rho(\Gamma)$ such that $g_i^{-1}(\gamma(i)) \in \rho(F)$. By the compactness of $\rho(F)$ we can find a subsequence, still denoted by g_i, such that $(g_i^{-1})_*(\gamma'(i))$ converges to a tangent vector v at some point $p \in \rho(F)$. If σ is a geodesic of N with $\sigma'(0) = v$, then σ is a line. This contradiction shows that N is compact.

Now let $\psi : (f,g) \to g$ be the projection of the isometry group of $\tilde{M}$ into the second factor, and consider the short exact sequence:

$$0 \to \ker\phi \to \pi_1(M) \to \operatorname{im}\phi \to 0.$$

Since $\ker\phi = (f,1)$, and f acts discretely on N which is compact, $\ker\phi$ is thus finite. The image $\operatorname{im}\phi$ acts on $\mathbb{R}^k$ with compact quotients (since M is compact). Therefore $\operatorname{im}\phi$ is a Biebabach group. The corollary follows. ☐

Corollary 4.12. *Let M^n be a complete Riemannian manifold with nonnegative Ricci curvature. Then any finitely generated subgroup of π_1 has polynomial growth of order $n-1$ unless M is a compact flat manifold, in which case the order is n.*

Proof. Let $\pi : \tilde{M} \to M$ be the universal cover of M with the pullback metric, and let $\tilde{M} = N \times \mathbb{R}^k$ be a splitting such that N does not contain any line. Fixing $(p,0) \in N \times \mathbb{R}^k$, we will estimate the volume of $B_p(r) \cap \Gamma(p,0)$ for r large. Note that any deck transformation in $\Gamma = \pi_1(M)$ is of the form (f,g), where f and g act as isometries of the two factors. We claim that there are no sequences (f_i, g_i) such that $f_i(p) = \gamma_i(t_i)$ for a geodesic γ_i, and that $\gamma_i|_{[t_i-i,t_i+i]}$ is minimal. In fact, if such a sequence exists, then $f_i^{-1} \circ \gamma_i|_{[-i,i]}$ is minimal with base point p, and a subsequence will converge to a line in N, contradicting the fact that N contains no lines. Thus there is a number $i_0 > 0$ such that, for any $(f,g) \in \Gamma$ and any geodesic γ from p to $f(p)$ with $f(p) = \gamma(t)$, the restriction $\gamma|_{[t-i_o,t+i_0]}$ is not minimizing. Denote by C the set of points $x \in N$ such that $x \in B_p(i_0)$ or any geodesic γ from p to $x = \gamma(t)$ does not minimize to $t + i_o$. We have shown that

$$\Gamma(p,0) \in C \times \mathbb{R}^k.$$

Since C is the image under $\exp_p$ of a set contained in T_pN, which is the union of sections of annular regions with width i_0 and a ball of radius i_0, it follows from the volume comparison theorem (unless $\dim N = 0$, in which case M is flat) that

$$\operatorname{vol}(B_p(r) \cap C) \le c(i_0^{\dim N} + i_0 r^{\dim N - 1}) \le cr^{n-1}.$$

As in the proof of Theorem 3.11, let $H = \langle g_1, \dots, g_k\rangle$ be a finitely generated subgroup of Γ. Then g_i can each be represented by a loop σ_i at $\pi(p,0)$, with length l_i. Let $\tilde{\sigma}_i$ be the lifting of σ_i at $(p,0)$; then $g_i(p,0) = \tilde{\sigma}_i(l_i)$ as deck transformations. Let

$$\varepsilon < \min\{l_1, \dots, l_k\}, \quad l = \max\{l_1, \dots, l_k\}.$$

Then, for any distinct $h_1, h_2 \in H$, we have $h_1(B_{\tilde{p}}(\varepsilon)) \cap h_2(B_{\tilde{p}}(\varepsilon)) = \varnothing$ and

$$\bigcup_{h \in U(r)} h(B_{\tilde{p}}(\varepsilon)) \subset B_{(p,0)}(rl + \varepsilon) \cap C.$$

Thus

$$|U(r)| \cdot \mathrm{vol}(B_{\tilde{p}}(\varepsilon)) = \sum_{h \in U(r)} \mathrm{vol}(h(B_{\tilde{p}}(\varepsilon))) \leq \mathrm{vol}(B_{\tilde{p}}(rl+\varepsilon) \cap C),$$

and

$$|U(r)| \leq \frac{\mathrm{vol}(B_{\tilde{p}}(rl+\varepsilon) \cap C)}{\mathrm{vol}(B_{\tilde{p}}(\varepsilon))} = \frac{c}{\mathrm{vol}(B_{\tilde{p}}(\varepsilon))}(rl+\varepsilon)^{n-1} \leq cr^{n-1}. \qquad \square$$

Abresch and Gromoll's estimate of the excess function. Given two points $p, q \in M$, the *excess function* is defined as

$$e_{p,q}(x) = d(p,x) + d(x,q) - d(p,q).$$

(When no confusion can occur, we will drop the reference to p, q.) Thus, the excess function measures how much the triangle inequality fails to be an equality. Since the excess function is made up of distance functions, the properties below follow directly from those of the distance function.

LEMMA 4.13. (1) $e(x) \geq 0$;
(2) $\mathrm{dil}(e) \leq 2$, *where* $\mathrm{dil}(f) = \inf |f(x) - f(y)|/d(x,y)$;
(3) $e|_\gamma = 0$ *where* γ *is a minimizing geodesic from* p *to* q; *and*
(4) *if* $\mathrm{Ric} \geq 0$, *then*

$$\triangle e \leq (n-1)\Big(\frac{1}{d(x,p)} + \frac{1}{d(x,q)}\Big)$$

in the barrier sense.

The following analytic lemma uses the Weak Maximum Principle, Theorem 4.3.

LEMMA 4.14. *Suppose* $\mathrm{Ric} \geq 0$, *and let* $u : B_y(R+\eta) \to \mathbb{R}^1$ (*for some* $\eta > 0$) *be a Lipschitz function satisfying*
(1) $u \geq 0$;
(2) $u(y_0) = 0$ *for some* $y_0 \in \overline{B_y(R)}$;
(3) $\mathrm{dil}(u) \leq a$; *and*
(4) $\triangle u \leq b$ *in the barrier sense.*
Then, for all $c \in (0, R)$, u *satisfies* $u(y) \leq a \cdot c + G(c)$, *where* G *is defined as*

$$G(x) = \frac{b}{2n}\Big(x^2 + \frac{2}{n-2}R^n x^{2-n} - \frac{n}{n-2}R^2\Big).$$

PROOF. Take $\varepsilon < \eta$, and define the function G using the value $R + \varepsilon$ in place of R. Since we can eventually let $\varepsilon \to 0$, it will suffice to prove the inequality in this case.

If d denotes the distance function to any point in the n-dimensional space form $S^n(H)$ of curvature H, then $G \circ d$ is the unique function on $S^n(H)$ satisfying
(i) $G \circ d(x) > 0$ for $0 < d(x) < R$,
(ii) G is decreasing for $0 < d(x) < R$,

(iii) $G(R) = 0$, and
(iv) $\triangle^H G \circ d = b$.

Now fix $c \in (0, R)$ and suppose the bound is false. Then $u(y) \geq a \cdot c + G(c)$, and it follows that

$$u|_{\partial B_y(c)} \geq u(y) - c \cdot \mathrm{dil}(u) \geq a \cdot c + G(c) - c \cdot a = G(c) = G|_{\partial B_y(c)},$$

and $u|_{\partial B_y(R)} \geq 0 = G|_{\partial B_y(R)}$. Thus, for ε small, we have $u|_{\partial A} \leq 0$ for the annulus $A = B_y(R) - B_y(\varepsilon)$. But $(G - u)(y_0) = G(y_0) > 0$; hence $G - u$ has a strict interior maximum in A. This violates the maximum principle since $\triangle(G - u) \geq 0$. □

For convenience, we set $s(x) = \min(d(p,x), d(q,x))$ and define the height $h(x) = \mathrm{dist}(x, \gamma)$ for any fixed minimal geodesic γ, from p to q.

THEOREM 4.15 (EXCESS ESTIMATE [Abresch and Gromoll 1990]). *If* $\mathrm{Ric} \geq 0$ *and* $h(x) \leq s(x)/2$, *then*

$$e(x) \leq 8\left(\frac{h(x)^n}{s}\right)^{1/(n-1)}.$$

PROOF. By Lemma 4.13, we can choose $a = 2, b = 4(n-1)/s(x)$, and $R = h(x)$ in Lemma 4.14, and let

$$c = \left(\frac{2h^n}{s}\right)^{1/(n-1)}.$$

Then

$$\begin{aligned} e &\leq 2\left(\frac{2h^n}{s}\right)^{1/(n-1)} + G(c) \\ &= \frac{b}{2n}\left(\left(\frac{2h^n}{s}\right)^{2/(n-1)} - \frac{n}{n-2}h^2 + \frac{2}{n-2}h^n 2^{2-n}\left(\frac{2h^n}{s}\right)^{(2-n)/(n-1)}\right) \\ &\leq 8\left(\frac{h(x)^n}{s}\right)^{1/(n-1)}. \end{aligned}$$

□

REMARK. In the above, we only stated the case where the Ricci curvature is nonnegative. It is easy to see that an estimate holds for general lower bounds on Ricci curvature, and takes the form

$$e(x) \leq E\left(\frac{h}{s}\right) \cdot h,$$

for some function E and $E(0) = 0$.

Critical points of the distance function and Toponogov's theorem. One of the most useful theories of differential topology is Morse theory. The main idea is that, if $f : M \to \mathbb{R}$ is a smooth function, the topology of M is reflected by the critical points of f. In geometry, the classical application of Morse theory is the energy functional of the loop space. As we have seen, the distance function is a natural function closely tied to the geometry and topology of a manifold. In this section, we will try to develop a Morse theory for the distance function.

This theory was originally proposed by Grove–Shiohama [1977], and formally formulated by Gromov [1981a].

DEFINITION 4.16. A point x is a *critical point* of d_p if, for all vectors $v \in M_p$, there is a minimizing geodesic γ from x to p such that the angle $\angle(v, \dot\gamma(0)) \le \pi/2$.

The main result we will use from Morse theory is the following theorem.

THEOREM 4.17 (ISOTOPY LEMMA). *If $r_1 < r_2 \le +\infty$ and $\overline{B_p(r_2)} \setminus B_p(r_1)$ has no critical point, then this region is homeomorphic to $\partial B_p(r_1) \times [r_1, r_2]$.*

PROOF. If x is not critical, there is a vector $w_x \in M_x$ with $\angle(w_x, \dot\gamma(0)) < \pi/2$ for all minimizing geodesics γ from x to p. By continuity, we can extend w_x to a vector field W_x in an open neighborhood U_x of x, such that if $y \in U_x$ and σ is any minimizing geodesic from y to p, then $\angle(\dot\sigma(0), W_x(y)) < \pi/2$. Take a finite subcover of $\overline{B_p(r_2)} \setminus B_p(r_1)$ by sets U_{x_i} (locally finite if $r_2 = +\infty$), and a smooth partition of unity $\sum \phi_i = 1$. Let $W = \sum \phi_i W_{x_i}$. From the restriction on angles, it follows that W is nonvanishing. Let $\psi_x(t)$ be the integral curve of W through x, and let $\sigma_t(s)$ be a minimal geodesic from p to $\psi_x(t)$. The first variation formula gives

$$\begin{aligned} d(\psi_x(t_2)) - d(\psi_x(t_1)) = \int_{t_1}^{t_2} \frac{d}{dt} d(\psi_x(t)) &= \int_{t_1}^{t_2} -\cos\angle(\dot{\sigma_t}(0), W(\psi_x(t))\, dt \\ &\le -\cos(\pi/2 - \varepsilon)(t_2 - t_1), \end{aligned}$$

for some $\varepsilon > 0$. This implies that d is strictly decreasing along $\psi(t)$. It now follows that the flow along ψ gives a homeomorphism: $\phi : \partial B_{r_1} \times [r_1, r_2) \to M$ with $\phi(x,t) = \psi_x(t)$. □

Since the definition of critical points requires a control on the angle, the standard application of this Morse theory requires the following result:

THEOREM 4.18 (TOPONOGOV COMPARISON THEOREM). *Let M^n be a complete Riemannian manifold with sec $\ge H$.*
(1) *Let $\{\gamma_0, \gamma_1, \gamma_2\}$ be a triangle in M, and assume that all three geodesics are minimizing. Then there is a triangle $\bar\gamma_0, \bar\gamma_1, \bar\gamma_2$ in the two-dimensional space form $S^2(H)$ of curvature H with* $\text{length}(\gamma_i) = \text{length}(\bar\gamma_i)$. *Furthermore, if α_i is the angle opposite γ_i, then $\bar\alpha_i \le \alpha_i$.*
(2) *Let $\{\gamma_1, \gamma_2, \alpha\}$ be a hinge in M, and assume that γ_1 and γ_2 are minimizing. Then there is a hinge $\bar\gamma_1, \bar\gamma_2$ in $S^2(H)$ with* $\text{length}(\gamma_i) = \text{length}(\bar\gamma_i) = l_i$ *and same angle α. Furthermore, $d(\gamma_1(l_1), \gamma_2(l_2)) \le d^H(\bar\gamma_1(l_1), \bar\gamma_2(l_2))$.*

We have not stated this theorem in the strongest possible form. For a complete statement and proof, we refer the reader to [Cheeger and Ebin 1975].

To show typical applications for the Morse theory, and provide background for later results, we prove the following two well-known results.

Theorem 4.19 (Grove–Shiohama Diameter Sphere Theorem [1977]). *Let M^n be a Riemannian manifold with sec $\geq H > 0$. If* $\operatorname{diam}(M) > (\pi/2\sqrt{H})$, *then M is homeomorphic to the sphere S^n.*

Proof. Without loss of generality, we assume $H = 1$. Let $p, q \in M$ be such that $d(p,q) = \operatorname{diam}(M)$, a simple first variation argument shows that p, q are mutually critical. The theorem follows from the following claim.

Any $x \in M \setminus \{p, q\}$ is noncritical for d_p and d_q.

To prove the claim, assume x is a critical point of d_p. Let σ_1 be a minimal geodesic from q to x, with length l_1. Since x is critical, there is a minimal geodesic σ_2 from x to p with length l_2 such that $\angle(\dot\sigma_2(0), \dot\sigma_1(l_1)) \leq \pi/2$. Since p, q are mutual critical points, there are minimal geodesics $\sigma_3, \tilde\sigma_3$ from q to p such that $\angle(\dot\sigma_3(0), \dot\sigma_1(0)) \leq \pi/2$, and $\angle(-\dot{\tilde\sigma}_3(0), -\dot\sigma_2(l_2)) \leq \pi/2$. Let $\{\bar\sigma_1, \bar\sigma_1, \bar\sigma_1\}$ be the comparison triangle in $S^n(1)$. Applying Toponogov's theorem twice—once to the triangle $\{\sigma_1, \sigma_2, \sigma_3\}$, then to the triangle $\{\sigma_1, \sigma_2, \tilde\sigma_3\}$—we obtain $\bar\alpha_i \leq \pi/2$, for $i = 1, 2, 3$. It now follows from elementary spherical geometry that $\operatorname{length}(\bar\sigma_3) \leq \pi/2$. But then Toponogov's theorem implies $d(p,q) = \operatorname{length}(\sigma_3) \leq \operatorname{length}(\bar\sigma_3) \leq \pi/2$, a contradiction. □

Theorem 4.20. *If M is complete and has nonnegative sectional curvature, it has finite topological type, i.e., it is homeomorphic to the interior of a compact manifold with boundary.*

Proof. We first prove that if x is a critical point of d_p, and y is such that $d(y) \geq \mu d(x)$, then for any minimal geodesic γ_1 of length l_1 from p to x and for any minimal geodesic γ_2 of length l_2 from p to y the angle $\theta = \angle(\dot\gamma_1(0), \dot\gamma_2(0))$ is at least $\cos^{-1}(1/\mu)$.

In fact, using part (2) of Toponogov's theorem, we have

$$l_3^2 = d(x,y)^2 \leq l_1^2 + l_2^2 - 2 l_1 l_2 \cos\theta.$$

Choose an arbitrary minimal geodesic γ_3 from x to y. Since x is critical, there is a minimal geodesic $\bar\gamma_1$ from x to p such that $\angle(\dot{\bar\gamma}_1(0), \dot\gamma_3(0)) \leq \pi/2$. Applying Toponogov's theorem again to this triangle we get $l_2^2 \leq l_1^2 + l_3^2$. Combining these two inequalities gives the desired lower bound on θ.

We now claim that d_p does not have critical points outside a compact set. If not, we would have a sequence of critical points x_i such that $d(x_{i+1}) \geq \mu d(x_i)$, for any i. Let γ_i be a minimal geodesic from p to x_i. From what we just proved in the preceding paragraph, we have

$$\angle(\dot\gamma_i, \dot\gamma_j) \geq \cos^{-1}(1/\mu),$$

and setting $\mu = 2$ we get $\angle(\dot\gamma_i, \dot\gamma_j) \geq \pi/3$. This gives a covering of the unit sphere at p by an infinite number of balls of fixed size (corresponding to solid angles of $\pi/6$), no two of which intersect. This is not possible since S_p is compact. Therefore d_p has no critical point outside of a compact set. By the isotopy lemma, M has finite topological type. □

Remark. This result is a weaker version of the Soul Theorem [Cheeger and Gromoll 1972], which says any complete manifold with nonnegative sectional curvature is diffeomorphic to the normal bundle of a compact totally geodesic submanifold, called the *soul*. The argument given above is due to Gromov [1981a].

Diameter growth and topological finiteness. Given the results just seen, it is quite natural to ask whether the conclusion of Theorem 4.20 is still true if one replaces nonnegative sectional curvature by nonnegative Ricci curvature. The examples of Sha and Yang showed this not to be the case [Sha and Yang 1989a; 1989b; Anderson 1990a]. In this section, we will prove a topological finiteness theorem for Ricci curvature under some additional conditions, due to Abresch and Gromoll [1990]. The excess estimate of section 4.3 was originally designed for this purpose. It turned out to be useful for other applications, as we will see in next section.

Definition 4.21. For $r > 0$, the open set $M \setminus \overline{B_p(r)}$ contains only finitely many unbounded components, and each such component has finitely many boundary components Σ_r. Define the diameter growth function $D(r,p)$ as the maximum diameter of the Σ_r, as measured in M.

Theorem 4.22 [Abresch and Gromoll 1990]. *Let M^n be a complete Riemannian manifold of nonnegative Ricci curvature, sectional curvature bounded below by $H > -\infty$, and satisfying $D(r,p) = o(r^{1/n})$. Then d_p has no critical point outside of a compact set. In particular, M has finite topological type.*

Given the lower bound on the Ricci curvature, the excess estimate in the last section gives a upper bound for the excess function. To see the relevance of the lower sectional bound in the above theorem, we state the following lemma, which gives a lower bound for the excess.

Lemma 4.23. *If M^n is complete with $\sec \geq -1$, and x a critical point of d_p, then for any $\varepsilon > 0$, there is a $\delta > 0$ such that if $d(q,x) \geq 1/\delta$, then*

$$e_{p,q}(x) \geq \ln\left(\frac{2}{1+e^{-2d(p,x)}}\right) - \varepsilon.$$

Proof. Take an arbitrary minimal geodesic γ from x to q. Since x is critical for d_p, there is a minimal geodesic σ from x to p such that $\angle(\dot\sigma(0), \dot\gamma)(0) \leq \frac{\pi}{2}$. By Toponogov's theorem, we have

$$\cosh d(p,q) \leq \cosh d(x,p) \cdot \cosh d(x.q).$$

When $d(x,q) \to \infty$, $d(p,q) \to \infty$ with $d(p,x)$ fixed, the above inequality becomes

$$\frac{e^{d(p,q)}}{2} \leq \cosh d(x,p) \frac{e^{d(x,q)}}{2},$$

and the lemma follows from the definition of the excess function. □

Proof of Theorem 4.22. Given a boundary component Σ_r for a noncompact component of $M \setminus B_p(r)$, we can construct a ray γ such that $\gamma(t) \in U(r)$, for all $t > r$. Thus, if $x \in \Sigma_r$ is critical, then $D(r,p) = o(r^{1/n})$, and Theorem 4.15 implies $e(x) \to 0$. But Lemma 4.23 gives a positive lower bound for $e(x)$. This is not possible. Thus, for any $r > r_0$, no point of any set Σ_r is critical for d_p.

Now fix r_0, U_{r_0}, a boundary component Σ_{r_0}, and a ray γ with $\gamma(r_0) \in \Sigma_{r_0}$ and $\gamma(t) \in U_{r_0}$ for $t > r_0$. For each $t \geq r_0$, let Σ_t denote the boundary component of the unbounded component of $M \setminus \overline{B_p(t)}$ with $\gamma(t) \in \Sigma_t$. Using the isotopy lemma we can construct an embedding $\psi : (r_0, \infty) \times \Sigma_0 \to U_{r_0}$ such that $\psi((t, \Sigma_{r_0})) = \Sigma_t$. It follows easily that $\psi((r_0,\infty) \times \Sigma_{r_0}))$ is open and closed in U_{r_0}. Hence $\psi((r_0,\infty) \times \Sigma_{r_0}) = U_{r_0}$. □

Remark. Opinions are divided as to whether the lower bound on sectional curvature is necessary. For that matter, it is interesting to consider the finiteness question for complete manifolds with positive Ricci curvature and bounded diameter. The techniques in [Perelman 1994] may help to settle this question in the negative.

Because of the relation between Ricci curvature and the volume growth of geodesic balls, as given by Theorem 3.5, for example, another approach to obtaining a finiteness theorem is to put conditions on the volume growth of geodesic balls, e.g., to require $\mathrm{vol}(B_p(r))$ to be close to $\omega_n r^n$ or close to cr. A positive result is the following theorem due to Perelman, which we will prove in the next section.

Theorem 4.24 [Perelman 1994]. *For any $n > 0$, there is a positive number $\varepsilon(n)$ such that if M is a complete n-dimensional manifold with* $\mathrm{Ric} \geq 0$ *and*

$$\mathrm{vol}(B_p(r)) \geq (1-\varepsilon)\omega_n r^n,$$

then M is contractible.

This is a topological stability result for the maximal volume growth condition. From [Perelman 1994], it seems possible to construct examples of complete manifolds of positive Ricci curvature with infinite topological type and satisfying the condition that $\mathrm{vol}(B_p(r)) \geq cr^n$, for some positive constant c. But the detailed computation still needs to be completed. No such example is known about the case where the volume grows slowly.

These finiteness considerations are closely related to the attempt to generalize Cheeger's Finiteness Theorem [1970] and Grove–Petersen's Homotopy Finiteness Theorem [1988] to Ricci curvature. The latter theorem says that the class of n-dimensional manifolds of sectional curvature $\geq H$, volume $\geq v > 0$, and diameter $\leq d$ contains only finitely many homotopy types. The crucial step in the proof is to show that geodesic balls of a fixed (small) size have simple topology. When the sectional curvature condition is replaced by Ricci curvature, a small geodesic ball (when rescaled) will resemble a complete manifold with $\mathrm{Ric} \geq 0$ and $\mathrm{vol}(B_p(r)) \geq cr^n$. The above example of Perelman also shows that this finiteness conjecture is false in dimensions four and above. In dimension three, there is a homotopy finiteness theorem due to Zhu [1990].

Perelman's almost maximal volume sphere theorem. Because of the Grove–Shiohama diameter sphere theorem (Theorem 4.19), efforts were made to generalize the result to Ricci curvature. This turned out not to be possible in dimension four and above. In fact, Anderson [1990b] and Otsu [1991] constructed metrics on $CP^n \# CP^n$ and $S^n \times S^m$ that have Ricci curvature at least equal to the dimension minus one, and diameter arbitrarily close to π. In dimension three, it was shown recently by Shen and Zhu [1995] that the diameter sphere theorem for Ricci curvature is still true (with slightly bigger diameter requirement).

It follows from Bishop's Volume Comparison Theorem that if the volume of an n-dimensional manifold with $\operatorname{Ric} \geq n-1$ is close to that of the unit sphere, then the diameter is close to π. One can thus attempt to prove a sphere theorem with conditions on the Ricci curvature and volume. This result was proved by Perelman, using the excess estimate of Abresch and Gromoll (Theorem 4.15).

THEOREM 4.25 [Perelman 1995]. *For any $n > 0$, there is a positive number $\varepsilon(n)$, such that if M^n satisfies $\operatorname{Ric} \geq n-1$ and $\operatorname{vol}(M) \geq (1-\varepsilon)\operatorname{vol}(S^n(1))$, then M is homeomorphic to S^n.*

REMARK. One needs only to prove that $\pi_k(M) = 0$ for all $k < n$, which implies that M is a homotopy sphere. Then the above theorem follows from the solution to Poincaré's conjecture when $n \geq 4$ (see [Smale 1961] and [Freedman 1982]), and from Hamilton's result [1982] when $n = 3$.

Theorem 4.24 can be considered as a noncompact version of Theorem 4.25. Both results are consequences of the next lemma.

MAIN LEMMA 4.26. *For any $C_2 > C_1 > 1$ and any integer $k \geq 0$, there is a constant $\delta = \delta_k(C_1, C_2, n) > 0$ such that, if M^n has Ricci curvature bounded below by $n-1$ and satisfies $\operatorname{vol}(B_q(\rho)) \geq (1-\delta)\operatorname{vol}(B^1(\rho))$ for any $B_q(\rho) \subset B_p(C_2 R)$ and $0 < R < \pi/C_2$, the following two k-parametrized properties hold:*

A(k) *Any continuous map $f : S^k = \partial D^{k+1} \to B_p(R)$ can be extended to a continuous map $g : D^{k+1} \to B_p(C_1 R)$.*

B(k) *Any continuous map $f : S^k \to M \setminus B_p(R)$ can be continuously deformed to a map $h : S^k \to M \setminus B_p(C_1 R)$.*

Theorem 4.25 follows from A(k) and B(k). Theorem 4.24 follows from A(k).

PROOF. We will work by induction on k. The induction step will be stated in Lemma 4.29 below, but first we need two other lemmas.

LEMMA 4.27. *If $\operatorname{Ric} \geq n-1$, there is a continuous function $E : \mathbb{R} \to \mathbb{R}^+$ such that*

$$e_{p,q}(x) \leq E\Big(\frac{h(x)}{d(p,x)}\Big) \cdot h(x),$$

assuming that $d(p,x) \le d(q,x)$ *and* $h(x)$ *is the height function, i.e., the distance from* x *to a minimal geodesic from* p *to* q.

This is a restatement of Theorem 4.15.

LEMMA 4.28. *For any* $C_2 > C_1 > 1$ *and* $\varepsilon > 0$, *there is a positive constant* $\gamma(C_1, C_2, \varepsilon, n)$ *such that, if* $\operatorname{Ric} \ge n-1$ *and* $\operatorname{vol}(B_p(C_2R)) \ge (1-\gamma)\operatorname{vol}(B^1(C_2R))$, *for any* $0 < R < \pi/C_2$, *then for any* $a \in B_p(R)$ *there is a point* $b \in M \setminus B_p(C_1R)$ *such that*

$$d(a, \overline{pb}) \le \varepsilon R,$$

where $\overline{pb}$ *denotes a minimal geodesic from* p *to* b.

Thus, under the curvature and volume conditions, there are lots of thin and long triangles. The excess estimate in Lemma 4.27 is a good estimate only when applied to such thin and long triangles.

PROOF. Set $\Gamma = \{\dot\sigma \mid d(a,\sigma) \le \varepsilon R\} \subset S_p^{n-1}$. Assume that, for all $v \in \Gamma$, the cut point in the direction of v is less than C_1R. We will derive a contradiction. In fact, under this assumption, it follows that

$$\begin{aligned}\operatorname{vol}(B_p(C_2R)) &= \int_\Gamma \int_0^{\operatorname{cut}(v)} A_M(t)\,dt\,dv + \int_{S^{n-1}\setminus\Gamma}\int_0^{\min\{C_2R,\operatorname{cut}(v)\}} A_M(t)\,dt\,dv \\ &\le \operatorname{vol}(\Gamma)\int_0^{C_1R} A^1(t)\,dt + (\operatorname{vol}(S^{n-1}) - \operatorname{vol}(\Gamma))\int_0^{C_2R} A^1(t)\,dt \\ &= -\operatorname{vol}(\Gamma)\int_{C_1R}^{C_2R} A^1(t)\,dt + \operatorname{vol}(S^{n-1})\int_0^{C_2R} A^1(t)\,dt \\ &\le -\operatorname{vol}(\Gamma)\int_{C_1R}^{C_2R} A^1(t)\,dt + \operatorname{vol}(B^1(C_2R)).\end{aligned}$$

(Note that $A^1(t) = \sin^{n-1} t$.) Thus

$$(1-\gamma)\operatorname{vol}(B^1(C_2R)) \le -\operatorname{vol}(\Gamma)\int_{C_1R}^{C_2R} A^1(t)\,dt + \operatorname{vol}(B^1(C_2R)),$$

which implies that

$$\operatorname{vol}(\Gamma) \le \gamma \cdot \frac{\operatorname{vol}(B^1(C_2R))}{\int_{C_1R}^{C_2R} A^1(t)\,dt}.$$

Also note that

$$\begin{aligned}\operatorname{vol}(B_a(\varepsilon R)) \le \operatorname{vol}(A^\Gamma_{0,C_1R}(p)) &\le \operatorname{vol}(\Gamma)\int_0^{C_1R} A^1(t)\,dt \\ &\le \gamma \cdot \frac{\operatorname{vol}(B^1(C_2R))}{\int_{C_1R}^{C_2R} A^1(t)\,dt} \cdot \int_0^{C_1R} A^1(t)\,dt.\end{aligned}$$

(For the definition of $A^{\Gamma}_{r,R}(p)$, see Theorem 3.1.) By the relative volume comparison, we have

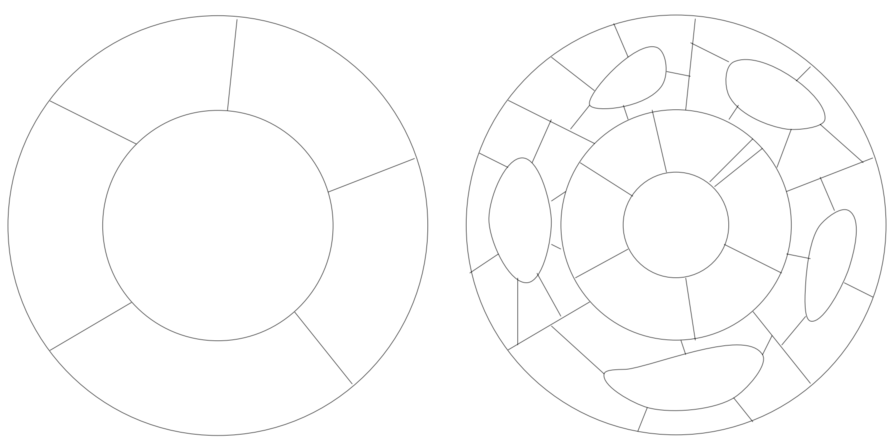
$$\begin{aligned}\mathrm{vol}(B_a(\varepsilon R)) &\geq \mathrm{vol}(B_a(R+C_2R))\cdot\frac{\int_0^{\varepsilon R}A^1(t)\,dt}{\int_0^{R+C_2R}A^1(t)\,dt}\\ &\geq \mathrm{vol}(B_p(C_2R))\cdot\frac{\int_0^{\varepsilon R}A^1(t)\,dt}{\int_0^{R+C_2R}A^1(t)\,dt}\\ &\geq (1-\gamma)\,\mathrm{vol}(B^1(C_2R))\cdot\frac{\int_0^{\varepsilon R}A^1(t)\,dt}{\int_0^{R+C_2R}A^1(t)\,dt}.\end{aligned}$$

These two inequalities together imply that

$$(1-\gamma)\,\mathrm{vol}(B^1(C_2R))\cdot\frac{\int_0^{\varepsilon R}A^1(t)\,dt}{\int_0^{R+C_2R}A^1(t)\,dt}\leq\gamma\cdot\frac{\mathrm{vol}(B^1(C_2R))}{\int_{C_1R}^{C_2R}A^1(t)\,dt}\cdot\int_0^{C_1R}A^1(t)\,dt,$$

which gives a bound $\gamma > C(C_1, C_2, n, \varepsilon)$. This is a contradiction if we choose $\gamma(C_1, C_2, n, \varepsilon) = C$. □

We now give an outline of the proof of A(k), which, as we have mentioned, uses induction on k. As usual, we view D^{k+1} as $S^k \times [1, 0)$ plus a point. In order to extend a function $f : S^k \to B_p(R)$ to D^{k+1}, we need to be able to get a map $\tilde{f} : S^k \to B_p(\alpha R)$, for $\alpha < 1$; we can then add the final point by continuity. The usual requirement is that $\tilde{f}$ be homotopic to f, so the function g is obtained by extending gradually along the radial direction from $S^k \times \{1\}$ to $S^k \times \{0\}$. Perelman's idea is that, in fact, one does not have to require $\tilde{f}$ to be homotopic to f. Instead, one only needs $\tilde{f}$ to be close to f. Then f and $\tilde{f}$ give an extension to the k-skeleton of a cell decomposition K_0 of D^{k+1}, as in the figure.



K_0 K_1

The requirement that $\tilde{f}$ and f be close guarantees that each k-cell of K_0 is mapped into a smaller ball. Now regard the boundary of each k-cell as S^k, and proceed in each such cell to extend the map to a finer cell decomposition K_1 of D^{k+1}, as in the second figure. Continuing in this fashion, one obtains the desired extension by adding infinitely many points (corresponding to the center of the k-cells of K_j when $j \to \infty$.) Thus, the main difficulty is to construct the map $\tilde{f}$ that maps into a smaller ball and is close to f. We will prove the existence of $\tilde{f}$ in the following lemma, which uses Lemmas 4.27 and 4.28, and the induction hypothesis A$(k-1)$. In the lemma, d_0 is a (small) positive number to be chosen later.

Lemma 4.29 (Induction Lemma P(k)). *Assume $A(k-1)$ is true. Then, given any map $\phi : S^k \to B_q(\rho)$, and a fine triangulation T of S^k with* $\operatorname{diam}(\phi(\Delta)) \leq d_0\rho$ *for any $\Delta \in T$, there exists a continuous map $\tilde{\phi} : S^k \to B_q((1-d_0)\rho)$ such that*

$$\operatorname{diam}(\phi(\Delta) \cup \tilde{\phi}(\Delta)) \leq \beta\rho,$$

for any $\Delta \in T$, where $\beta = (4 + 2d_0/k)^{-k}(1 - 1/C_1)$.

Proof. We will construct the map $\tilde{\phi}$ on the skeletons $\operatorname{skel}_i(T)$ of T, for $i = 0, 1, \ldots, k$. We proceed by induction on i.

When $i = 0$, for any $x \in \operatorname{skel}_0(T)$, let γ_x be a minimal geodesic from q to $\phi(x)$. Define $\tilde{\phi}(x) = \gamma_x((1-2d_0)\rho)$. Then $\tilde{\phi}(\Delta) \subset B_q((1-2d_0)\rho)$ and $\operatorname{diam}(\phi(\Delta) \cup \tilde{\phi}(\Delta)) \leq 10d_0\rho$ (we could use $2d_0\rho$ on the right-hand side for the latter inequality).

Assume that $\tilde{\phi}$ has already been defined on the i-skeleton skel_i, with $\tilde{\phi}(\Delta) \subset B_q((1-d_0(2-\frac{i}{k}))\rho)$ and $\operatorname{diam}(\phi(\Delta) \cup \tilde{\phi}(\Delta)) \leq 10d_i\rho$, where d_i will be determined later. We now construct $\tilde{\phi}$ on $\operatorname{skel}_{i+1}$.

For any $\Delta \in \operatorname{skel}_{i+1}$, we can assume that $\phi(\Delta) \not\subset B_q((1-2d_0)\rho)$; otherwise we are done. By Lemma 4.28, there is a point $y_\Delta \in M \setminus B_q(C_1\rho)$ such that

$$d(\phi(\Delta), \overline{qy_\Delta}) \leq 2d_0\rho.$$

Let σ be a minimal geodesic from q to y_Δ, and let $q_\Delta = \sigma((1-d_{i+1})\rho)$. Then, for any $x \in \partial\Delta$, we check that the triangle $\tilde{\phi}(x)q_\triangle q$ is thin and long:

$$d(\tilde{\phi}(x), q_\Delta \bar{y}_\Delta) \leq d(\phi(x), q_\Delta \bar{y}_\Delta) + \operatorname{diam}(\phi(\partial(\Delta)) \cup \tilde{\phi}(\partial\Delta)) \leq 2d_0\rho + 10d_i\rho \leq 20d_i\rho,$$

$$d(\tilde{\phi}(x), q_\Delta) \geq d(q_\Delta, \phi(\Delta)) - \operatorname{diam}(\phi(\partial\Delta) \cup \tilde{\phi}(\partial\Delta)) \geq d_{i+1}\rho - 10d_i\rho \geq \frac{d_{i+1}}{2}\rho,$$

$$d(\tilde{\phi}(x), y_\Delta) \geq d(y_\Delta, q) - d(\tilde{\phi}(x), q) \geq C_1\rho - (1 - d_0(2 - i/k))\rho \geq \tfrac{1}{2}d_{i+1}\rho.$$

Here we have assumed that $d_{i+1} \geq 100d_i$ and $d_k \leq C_1 - 1$.

By Lemma 4.27, we have

$$d(\tilde{\phi}(x), y_\Delta) + d(\tilde{\phi}(x), q_\Delta) - d(q_\Delta, y_\Delta) \leq 20d_i\rho E\Big(\frac{100d_i}{d_{i+1}}\Big).$$

The triangle inequality gives

$$d(q_\Delta, y_\Delta) + \rho(1 - d_{i+1}) = d(q, y_\Delta) \leq d(\tilde{\phi}(x), y_\Delta) + d(\tilde{\phi}(x), q).$$

Adding these two inequalities, we obtain

$$\begin{aligned} d(\tilde{\phi}(x), q_\Delta) &\leq d(\tilde{\phi}(x), q) + 20 d_i \rho E\Big(\frac{100 d_i}{d_{i+1}}\Big) - \rho(1 - d_{i+1}) \\ &\leq (1 - d_0(2 - \frac{i}{k}))\rho + 20 d_i \rho E\Big(\frac{100 d_i}{d_{i+1}}\Big) - \rho(1 - d_{i+1}) \\ &\leq \Big(d_{i+1} - d_0\Big(2 - \frac{2i+1}{2k}\Big)\Big)\rho, \end{aligned}$$

where we need

$$20 d_i E\Big(\frac{100 d_i}{d_{i+1}}\Big) \leq d_0/2k.$$

Since A(i) is true (note that $i \leq k-1$), we can use A(i) to extend the map (using $1 + d_0/2k$ as the factor C_1; we thus require $\delta_{i+1} < \delta_i(1 + d_0/2k, C_2)$). We get

$$\begin{aligned} \tilde{\phi}\Big(\Delta, q_\Delta\Big) &\leq \Big(1 + \frac{d_0}{2k}\Big)\Big(d_{i+1} - d_0\Big(2 - \frac{2i+1}{2k}\Big)\Big)\rho \\ &\leq \Big(d_{i+1} - d_0\Big(2 - \frac{i+1}{k}\Big)\Big)\rho. \end{aligned}$$

Therefore

$$\begin{aligned} \operatorname{diam}(\phi(\Delta) \cup \tilde{\phi}(\Delta)) &\leq \operatorname{diam}(\phi(\partial\Delta) \cup \tilde{\phi}(\partial\Delta)) + \operatorname{diam}(\phi(\Delta)) + \operatorname{diam}(\tilde{\phi}(\Delta)) \\ &\leq 10 d_i \rho + d_0 \rho + \Big(d_{i+1} - d_0\Big(2 - \frac{i+1}{k}\Big)\Big)\rho \leq 10 d_{i+1}\rho \leq \beta. \end{aligned}$$

Here we require $10 d_k \leq \beta$. Also,

$$\begin{aligned} d(\tilde{\phi}(\Delta), q) &\leq d(q, q_\Delta) + d(\tilde{\phi}(x), q_\Delta) \\ &\leq (1 - d_{i+1})\rho + \Big(d_{i+1} - d_0\Big(2 - \frac{i+1}{k}\Big)\Big)\rho \leq (1 - d_0)\rho. \end{aligned}$$

Thus we complete the proof of the Induction Lemma if we choose

$$\delta_k \leq \min\{\gamma(C_1, C_2, d_0),\ \delta_i(1 + d_0/(2k), C_2) \text{ for } i = 0, 1, \ldots, k-1\}. \qquad \square$$

We now turn to the proof of Lemma 4.26 proper, by induction on k. The case $k = 0$ is obvious. Assume A($k-1$) is true, i.e., any map $f : S^i \to B_p(R)$, for $i \leq k-1$, can be extended to $g : D^{i+1} \to B_p(\alpha R)$, with $\alpha = 1 + d_0/2k$. Now consider a map $f : S^k \to B_p(R)$ and view $S^k = \partial D^{k+1}$. Give a fine triangulation T of S^k, view $D^{k+1} = T \times (0,1] \cup \{0\}$. We define a sequence of cell decompositions $K_j (j = 0, 1, \ldots)$ of D^{k+1} as discussed before Lemma 4.29. We only give the k-skeleton; the lower skeletons are naturally induced from T:

$$\begin{aligned} &\operatorname{skel}_k(K_0) = \partial D^{k+1} = S^k \\ &\operatorname{skel}_k(K_1) = (S^k \times \{\tfrac{1}{2}\}) \cup (S^k \times \{1\}) \cup (\operatorname{skel}_{k-1}(T) \times [\tfrac{1}{2}, 1]) \end{aligned}$$

Each k-simplex in K_1 can be considered as a map $\sigma : S^k \to B_p(R)$, and then can be further subdivided to define K_j inductively using the above formula for each k-cell of K_{j-1}. We then define a function f_j on K_j so that $f_0 = f,\ f_{j+1}|_{S^k\times\{1\}} = f_j,\ f_{j+1}|_{S^k\times\{\frac{1}{2}\}} = \tilde{f}_j$, by Lemma 4.29 $P(j)$, and $f_{j+1}|_{\mathrm{skel}_{k-1}(T)\times[\frac{1}{2},1]}$ (extension by induction hypothesis on $\mathrm{skel}_i(T)\times[\frac{1}{2},1]$, for $i = 0, 1, \ldots, k-1$).

We will let the desired extension be $g = \lim f_j$. To see that this gives a continuous function with the desired properties, we estimate the size of its image as follows.

If $j = 0$, we have $f_0 : \mathrm{skel}_k(K_0) \to B_p(R)$.

If $j = 1$, the map $f_1 : \mathrm{skel}_k(K_1) \to M$ satisfies, for any $\Delta \in \mathrm{skel}_k(K_1)$,

$$f_1|_{\Delta^k\times\{1\}} = f_0 \subset B_p(R),$$
$$f_1|_{\Delta^k\times\{\frac{1}{2}\}} = \tilde{f}_0 \subset B_p((1-d_0)R);$$

therefore

$$\begin{aligned}\mathrm{diam}(f_1|_{\mathrm{skel}_k(K_1)}(\Delta)) &\le \mathrm{diam}(f_1(\Delta^k\times\{1\})\cup f_1(\Delta^k\times\{\tfrac{1}{2}\})) + \mathrm{diam}(f_1|_{\partial\Delta\times[\frac{1}{2},1]})\\ &\le \beta R + \alpha^k\beta R = (1+\alpha^k)\beta R\end{aligned}$$

and

$$d(f_1|_{\mathrm{skel}_k(K_1)}, p) \le d(f_1|_{\Delta^k\times\{\frac{1}{2}\}}) + \mathrm{diam}(f_1|_{\mathrm{skel}_k(K_1)}) \le (1-d_0)R + (1+\alpha^k)\beta R,$$

where we denote by p_Δ the center of the ball containing $f_1(\partial\Delta)$.

We have spelled out these two cases to make it easier to see the general formula. For general j, we have

$$\begin{aligned}\mathrm{diam}(f_j|_{\mathrm{skel}_k(K_j)}(\Delta)) &\le \big((1+\alpha^k)\beta\big)^j R,\\ d(f_j|_{\mathrm{skel}_k(K_j)}(\Delta), p_\Delta)) &\le (1-d_0)R + R\sum_{i=1}^{j}\big((1+\alpha^k)\beta\big)^i + R(1-d_0)\sum_{i=1}^{j-1}\big((1+\alpha^k)\beta\big)^i\\ &\le (1-d_0)R + 2R\sum_{i=1}^{j}\big((1+\alpha^k)\beta\big)^i \le \frac{1-C_1^{-1}}{2},\end{aligned}$$

where we used the definition of α.

Therefore

$$\begin{aligned}\mathrm{diam}(f_j|_{\mathrm{skel}_k(K_j)}(\Delta)) &\le \left(\frac{1-C_1^{-1}}{2}\right)^j R \to 0\\ d(f_j|_{\mathrm{skel}_k(K_j)}(\Delta), p_\Delta)) &\le \left((1-d_0) + \frac{1-C_1^{-1}}{1-\frac{1}{2}(1-C_1^{-1})}\right)R\\ &\le C_1 R.\end{aligned}$$

Thus, each cell is mapped into a ball that is smaller by a factor bounded away from 1. Therefore when $j \to \infty$, the limit $g = \lim f_j$ is continuous. This concludes the proof of Lemma 4.26. ☐

5. Mean Curvature Comparison and its Applications

Direct applications of the Mean Curvature Comparison Theorem as stated in Theorem 2.2(3) have not been numerous. On the other hand, we should point out that Ricci curvature naturally enters into the second variation for the area of hypersurfaces. This aspect of the relation between Ricci curvature and mean curvature was explored extensively in dimension three by Meeks, Schoen, and Yau, among others, as in the proof of the Positive Mass Conjecture by Schoen and Yau. In this section, we will only give one application of the mean curvature comparison, to the Diameter Sphere Theorem of Perelman. The result was also obtained by Colding by a different method.

Recall that the direct generalization of the Grove–Shiohama Diameter Sphere Theorem to Ricci curvature is not correct in dimensions four and above, as shown by the examples of Anderson and Otsu. In dimension three, it holds by a recent result of Shen and Zhu [1995]. In the preceding section we strengthened the condition to volume. Here we will keep the diameter condition, but add a condition on the sectional curvature.

THEOREM 5.1 [Perelman 1997]. *For any positive integer n and any number H, there exists a positive number $\varepsilon(n,H)$ such that if a n-dimensional manifold M^n satisfies* $\operatorname{Ric} \geq n-1$, $\sec \geq H$, *and* $\operatorname{diam} \geq \pi - \varepsilon$, *then M is a twisted sphere.*

LEMMA 5.2. *For any $\delta > 0$ and any positive integer n, there is a positive number $\varepsilon(\delta,n)$ such that if $p,q \in M^n$ satisfy* $\operatorname{Ric} \geq n-1$ *and $d(p,q) \geq \pi - \varepsilon$, then $e_{p,q}(x) \leq \delta$.*

PROOF. Let $e = e_{p,q}(x)$. The triangle inequality implies that the three geodesic balls $B_p(d(x,p)-e/2)$, $B_q(d(x,q)-e/2)$, and $B_x(e/2)$ have disjoint interiors. Therefore

$$\begin{aligned}\operatorname{vol}(M) &\geq \operatorname{vol}(B_x(e/2)) + \operatorname{vol}(B_p(d(x,p)-e/2)) + \operatorname{vol}(B_q(d(x,q)-e/2))\\ &\geq \operatorname{vol}(M)\left(\frac{\operatorname{vol}(B^1(e/2))}{\operatorname{vol}(B^1(d(p,q)))} + \frac{\operatorname{vol}(B^1(d(p,x)-e/2))}{\operatorname{vol}(B^1(d(p,q)))} + \frac{\operatorname{vol}(B^1(d(q,x)-e/2))}{\operatorname{vol}(B^1(d(p,q)))}\right).\end{aligned}$$

Thus

$$\begin{aligned}\operatorname{vol}(d(p,q)) &\geq \operatorname{vol}(B^1(e/2)) + \operatorname{vol}(B^1(d(p,x)-e/2)) + \operatorname{vol}(B^1(d(q,x)-e/2))\\ &\geq \operatorname{vol}(B^1(e/2)) + 2\operatorname{vol}\Big(B^1\Big(\frac{(d(p,x)-e/2)+(d(q,x)-e/2)}{2}\Big)\Big)\\ &= \operatorname{vol}(B^1(e/2)) + \operatorname{vol}\Big(B^1\Big(\frac{d(p,q)}{2}\Big)\Big),\end{aligned}$$

where, for the second inequality, we used the fact that the volume of balls in $S^n(1)$ is a convex function of the radius. Therefore

$$\operatorname{vol}(B^1(e/2)) \leq \operatorname{vol}(B^1(d(p,q))) - 2\operatorname{vol}(B^1(\frac{d(p,q)}{2})).$$

The right-hand side approaches 0 when $\varepsilon \to 0$. This gives the desired bounds. □

Corollary 5.3. *For any* $\rho > 0$, *any* H, *and any positive integer* n, *there is a positive number* $\varepsilon(n, \rho, H)$ *such that if* $p, q \in M^n$ *satisfy* $\mathrm{Ric} \geq n-1$, $\sec \geq H$, *and* $d(p,q) \geq \pi - \varepsilon$, *then no* x *with* $\min\{d(x,p), d(x,q)\} > \rho$ *is not a critical point of* d_p, d_q.

Proof. Without loss of generality, we will assume $H = -K^2$ is negative. Denoting by α the angle at x, and applying Toponogov's theorem to the triangle pqx, we obtain

$$\cosh Kd(p,q) \leq \cosh Kd(p,x)\cosh Kd(x,q) - \sinh Kd(p,x)\sinh Kd(x,q)\cos\alpha.$$

Using the excess estimate of the previous lemma, we immediately conclude that $\cos\alpha \to -1$ as $\varepsilon \to 0$. $\square$

Proof of Theorem 5.1. By the preceding discussion, we only need to consider points that are very close to p or q. Let m be a critical point of d_p such that $d_p(m) \leq \rho$. Define a function $g : M \to R$ by

$$g(x) = \min_{y \in [pm]} \{d(x,y) + (d(p,y) - d(m,y))^2 - d(p,m)^2\},$$

where $[pm]$ denotes the union of all minimal geodesics from p to m.

Lemma 5.4. (1) $g(x) < d(x,m)$.
(2) *If* $M(R) = \{x | g(x) \leq R\}$ *and* $R_1 = \inf\{r > 0 | M(R) \cap \overline{B_q(r)} \neq \varnothing\}$, *then for any* $x_0 \in M(R) \cap \overline{B_q(R_1)}$, *we have* $g(x_0) < d(x_0, p)$.
(3) $|R + R_1 - \pi| < 2\rho$.

Remark. This lemma implies that $M(R) \cap \overline{B_q(R_1)} \neq \{m, p\}$. This is the reason for considering the function g; we could just use the distance function to $[pm]$ if $[pm]$ forms a close geodesic.

Proof. (1) Note that if we let $y = m$, then $d(x,y) + (d(p,y) - d(m,y))^2 - d(p,m)^2 = d(x,m)$. Thus $g(x) \leq d(x,m)$. Too see the strict inequality, we note that since m is a critical point, the angle at m is at most $\pi/2$. By Toponogov's theorem,

$$d(x,y) \leq \sqrt{d^2(x,m) + d^2(m,y)} = d(x,m) + O(d^2(m,y)).$$

Thus

$$\begin{aligned} d(x,y) + (d(p,y) - d(m,y))^2 - d(p,m)^2 &= d(x,y) - 4d(p,y)d(y,m) \\ &\leq d(x,m) + O(d^2(m,y)) - 4d(p,y)d(y,m) \\ &< d(x,m). \end{aligned}$$

(2) Since $d(q,m) \leq d(q,p) \leq d(q,x_0) + d(x_0,p) = R_1 + d(x_0,p)$, we have

$$d(m, B_q(R_1)) \leq d(x_0, p).$$

Thus, there exists a point $x_1 \in \partial B_q(R_1)$, such that $d(m, x_1) \leq d(x_0, p)$. This, together with the definition of R_1, implies that

$$g(x_0) \leq g(x_1) < d(x_1, m) \leq d(x_0, p).$$

(3) Take x_0 as in (2). Then

$$R + R_1 = g(x_0) + d(x_0, q) < d(x_0, q) + d(x_0, q) \leq d(p, q) \leq \pi.$$

Similarly, if $g(x_0)$ is realized at y_0, an interior point of some shortest line from p to m, then

$$\begin{aligned} R + R_1 &= g(x_0) + d(x_0, q) = d(x_0, y_0) - 4d(p, y_0)d(y_0, m) + d(x_0, q) \\ &\geq d(y_0, q) - \rho \geq \pi - 2\rho, \end{aligned}$$

if we choose ε in Corollary 5.3 smaller than ρ. □

LEMMA 5.5. *For any point $x_0 \in \partial M(R)$, let $\vec{n}$ be the unit normal vector pointing away from p. Then the mean curvature of $\partial M(R)$ can be estimated as*

$$m_{\vec{n}}(x_0) \leq (n-2)\coth\sqrt{K}R + \tanh\sqrt{K}R + 10.$$

PROOF. By Lemma 5.4, at x_0, the value of $g(x_0)$ is achieved at an interior point y_0 of some minimal geodesic γ from p to m. Then the function $g(x)$ is a perturbation of the distance function to the geodesic γ, whose mean curvature does not exceed $(n-2)\coth\sqrt{K}R + \tanh\sqrt{K}R$.

For convenience of notation, we parametrize γ so that $\gamma(0) = y_0$. Let σ be a minimal geodesic from y_0 to x_0 with $\sigma(t_0) = x_0$. Let $V(t)$ be the vector field along σ obtained by parallel translation of $\dot{\gamma}(0)$ along σ. Let v_1 be the projection of $V(t_0)$ in the tangent plane of $\partial M(R)$ (v_1 is not a unit vector). Let $\{v_2, \ldots, v_{n-1}\}$ be orthonormal tangent vectors of $\partial M(R)$ that are all perpendicular to v_1. By the first variation formula, the v_i are all perpendicular to $\dot{\sigma}(t_0)$, and the values of g along the directions v_i are all achieved at y_0. It now follows from the Hessian Comparison Theorem that $\langle \nabla_{v_i}\dot{\sigma}(t_0), v_i\rangle \leq \coth\sqrt{K}t_0$ for $i = 2, \ldots, n-1$.

We now consider the direction v_1. Let $J(t)$ be a Jacobian field along σ such that $J(0) = \dot{\gamma}(0)$ and $J(t_0) = v_1$. We decompose J into directions along σ and orthogonal to σ, then call the orthogonal component $W(t)$, i.e.,

$$J(t) = (at + b)\dot{\sigma}(t) + W(t).$$

In general, $W(t_0)$ is not a unit vector. Let $\bar{J}(t) = cJ(t)$ be such that the orthogonal component $\bar{W}$ of $\bar{J}$ at t_0 has unit length. Then the Hessian comparison theorem applied to this case implies

$$\langle \nabla_{\bar{W}(t_0)}\dot{\sigma}(t_0), \bar{W}(t_0)\rangle \leq \tanh\sqrt{K}t_0.$$

To get the desired estimate on the mean curvature of $\partial M(R)$, we need to estimate the numbers a, b, c in this decomposition, which in return depend on the angle $\angle(\dot{\sigma}(0), \dot{\gamma}(0))$.

Note that $g(x_0) = d(x_0, y_0) - 4d(y_0, p)d(y_0, m)$. The first variation formula gives

$$0 = -\cos\angle(\dot\sigma(0), \dot\gamma(0)) - 4(\rho - 2d(y_0, m));$$

thus

$$|\cos\angle(\dot\sigma(0), \dot\gamma(0))| \le 4\rho.$$

This will give a uniform bound for a, b, c. Note also that

$$|R - t_0| = 4d(y_0, p)d(y_0, m) \le 4\rho^2.$$

The lemma follows if we choose ρ small enough. □

We continue with the proof of Theorem 5.1. Choose $R > 0$ such that

$$(n-1)\cot R - (n-2)\sqrt{K}\coth\sqrt{K}R - \sqrt{K}\tanh\sqrt{K}R - 10 > 0.$$

Thus, R depends on n, K. Then choose ρ small enough so that Lemma 5.5 holds; this in return determines the number ε. For such an ε, by Corollary 5.3, there are no critical points at distance ρ from p and q. Assume there is a critical point m such that $(p, m) \le \rho$; we will derive a contradiction using the mean value comparison.

On the one hand, using the condition $\mathrm{Ric} \ge n-1$ and applying Theorem 2.2 to d_q, we conclude, letting $m^1(R_1)$ be the mean curvature of the sphere of radius R_1 in $S^n(1)$:

$$\begin{aligned} m_{\partial B_q(R_1)}(x_0) &\le m^1(R_1) \le (n-1)\cot R_1 \\ &\le (n-1)\cot(\pi - R - 2\rho) = -(n-1)\cot(R + 2\rho) \\ &\le -(n-1)\cot R + \kappa, \end{aligned}$$

where $\kappa = \kappa(n, \rho)$ and $\lim_{\rho\to 0}\kappa = 0$. On the other hand, Lemma 5.5 implies

$$m_{\partial M(R),N}(x_0) \ge -(n-2)\sqrt{K}\coth\sqrt{K}R - \sqrt{K}\tanh\sqrt{K}R - 10.$$

Denote by N the unit normal vector pointing away from q. Since $\partial M(R)$ and $\partial B_q(R_1)$ are tangent to each other at x_0, and $\partial M(R)$ lies outside (with respect to N) of $\partial B_q(R_1)$, we can write

$$m_{\partial M(R),N}(x_0) \le m_{\partial B_q(R_1)}(x_0),$$

which implies

$$(n-1)\cot R - (n-2)\sqrt{K}\coth\sqrt{K}R - \sqrt{K}\tanh\sqrt{K}R - 10 \le \kappa.$$

We get a contradiction if we choose ε so small that κ violates this inequality. □

References

[Abresch and Gromoll 1990] U. Abresch and D. Gromoll, "On complete manifolds with nonnegative Ricci curvature", *J. Amer. Math. Soc.* **3** (1990), 355–374.

[Anderson 1990a] M. T. Anderson, "Short geodesics and gravitational instantons", *J. Diff. Geom.* **31** (1990), 265–275.

[Anderson 1990b] M. T. Anderson, "Metrics of positive Ricci curvature with large diameter", *Manuscripta Math.* **68** (1990), 405–415.

[Berard 1988] P. Berard, "From vanishing theorems to estimating theorems: the Bochner technique revisited", *Bull. Amer. Math. Soc.* **19** (1988), 371–406.

[Berger 1960] M. Berger, "Les variétés riemanniennes $\frac{1}{4}$-pincées", *Ann. Scuola Norm. Sup. Pisa* **14** (1960), 161–170.

[Bochner 1946] S. Bochner, "Vector fields and Ricci curvature", *Bull. Amer. Math. Soc.* **52** (1946), 776–797.

[Cai et al. 1994] M. Cai, G. Galloway and Z. Liu, "Local splitting theorems for Riemannian manifolds", *Proc. Amer. Math. Soc.* **120** (1994), 1231–1239.

[Cheeger 1967] J. Cheeger, "Comparison and finiteness theorems for Riemannian manifolds", Ph.D. Thesis, Princeton University, 1967.

[Cheeger 1970] J. Cheeger, "Finiteness theorems for Riemannian manifolds", *Am. J. Math.* **92** (1970), 61–74.

[Cheeger 1991] J. Cheeger, "Critical points of distance functions and applications to geometry", pp. 1–38 in *Geometric topology: recent developments* (Montecatini Terme, 1990), edited by J. Cheeger et al., Lecture Notes in Math. **1504**, Springer, Berlin, 1991.

[Cheeger and Colding 1995] J. Cheeger and T. H. Colding, "Almost rigidity of warped products and the structure of spaces with Ricci curvature bounded below", *C. R. Acad. Sci. Paris* Sér. I **320** (1995), 353–357.

[Cheeger and Ebin 1975] J. Cheeger and D. Ebin, *Comparison theorems in Riemannian geometry*, North-Holland, Amsterdam, and Elsevier, New York, 1975.

[Cheeger and Gromoll 1971] J. Cheeger and D. Gromoll, "The splitting theorem for manifolds of nonnegative Ricci curvature", *J. Differential Geom.* **6** (1971), 119–129.

[Cheeger and Gromoll 1972] J. Cheeger and D. Gromoll, "On the structure of complete manifolds with nonnegative curvature", *Ann. of Math.* **96** (1972), 413–443.

[Cheng 1975] S. Y. Cheng, "Eigenvalue comparison theorems and its geometric applications", *Math. Z.* **143** (1975), 289–297.

[Colding 1995] T. H. Colding, "Stability and Ricci curvature", *C. R. Acad. Sci. Paris* Sér. I **320** (1995), 1343–1347.

[Colding 1996a] T. H. Colding, "Shape of manifolds with positive Ricci curvature", *Invent. Math.* **124** (1996), 175–191.

[Colding 1996b] T. H. Colding, "Large manifolds with positive Ricci curvature", *Invent. Math.* **124** (1996), 193–214.

[Freedman 1982] M. H. Freedman, "Topology of four-dimensional manifolds", *J. Diff. Geom.* **17** (1982), 357–453.

[Gallot 1983] S. Gallot, "A Sobolev inequality and some geometric applications", pp. 45–55 in *Spectra of Riemannian manifolds*, edited by M. Berger et al., Kaigai, Tokyo, 1983.

[Gilbarg and Trudinger 1983] D. Gilbarg and N. Trudinger, *Elliptic partial differential equations of second order*, 2nd ed., Grundlehren der math. Wiss. **224**, Springer, Berlin, 1983.

[Greene and Wu 1988] R. E. Greene and H. Wu, "Lipschitz convergence of Riemannian manifolds", *Pacific J. Math.* **131** (1988), 119–141.

[Gromov 1981a] M. Gromov, "Curvature, diameter, and Betti numbers", *Comment. Math. Helv.* **56** (1981), 53–78.

[Gromov 1981b] M. Gromov, "Groups of polynomial growth and expanding maps", *Publ. Math. IHES* **53** (1981), 183–215.

[Gromov 1981c] M. Gromov, *Structures métriques pour les variétés riemanniennes*, edited by J. Lafontaine and P. Pansu, CEDIC, Paris, 1981.

[Grove and Petersen 1988] K. Grove and P. Petersen, "Bounding homotopy types by geometry", *Ann. of Math.* (2) **128** (1988), 195–206.

[Grove and Shiohama 1977] K. Grove and K. Shiohama, "A generalized sphere theorem", *Ann. of Math.* (2) **106** (1977), 201–211.

[Hamilton 1982] R. S. Hamilton, "Three-manifolds with positive Ricci curvature", *J. Diff. Geom.* **17** (1982), 255–306.

[Klingenberg 1961] W. Klingenberg, "Über Mannigfaltigkeiten mit positiver Krümmung", *Comm. Math. Helv.* **35** (1961), 47–54.

[Milnor 1968a] J. Milnor, "A note on curvature and fundamental group", *J. Diff. Geom.* **2** (1968), 1–7.

[Milnor 1968b] J. Milnor, "Growth of finitely generated solvable groups", *J. Diff. Geom.* **2** (1968), 447–449.

[Myers 1935] S. Myers, "Riemannian manifolds in the large", *Duke Math. J.* **1** (1935), 39–49.

[Otsu 1991] Y. Otsu, "On manifolds of positive Ricci curvature with large diameter", *Math. Z.* **206** (1991), 255–264.

[Perelman 1994] G. Perelman, "Manifolds of positive Ricci curvature with almost maximal volume", *J. Amer. Math. Soc.* **7** (1994), 299–305.

[Perelman 1995] G. Perelman, "A diameter sphere theorem for manifolds of positive Ricci curvature", *Math. Z.* **218** (1995), 595–596.

[Perelman 1997] G. Perelman, "Construction of manifolds of positive Ricci curvature with big volume and large Betti numbers", pp. 157–163 in this volume.

[Peters 1987] S. Peters, "Convergence of Riemannian manifolds", *Compositio Math.* **62** (1987), 3–16.

[Schoen and Yau 1979] R. Schoen and S.-T. Yau, "On the proof of the positive mass conjecture in general relativity", *Comm. Math. Phys.* **65** (1979), 45–76.

[Schoen and Yau 1982] R. Schoen and S.-T. Yau, "Complete three-dimensional manifolds with positive Ricci curvature and scalar curvature", pp. 209–228 in

Seminar on Differential Geometry, edited by S.-T. Yau, Ann. of Math. Stud. **102**, Princeton Univ. Press, Princeton, NJ, 1982.

[Sha and Yang 1989a] J. Sha and D. Yang, "Examples of manifolds of positive Ricci curvature", *J. Diff. Geom.* **29** (1989), 95–103.

[Sha and Yang 1989b] J. Sha and D. Yang, "Positive Ricci curvature on the connected sums of $S^n \times S^m$", *J. Diff. Geom.* **33** (1991), 127–137.

[Shen and Zhu 1995] Y. Shen and S. Zhu, "Ricci curvature, minimal surfaces and sphere theorems", preprint, 1995. To appear in *J. Geom. Anal.*

[Shiohama 1983] K. Shiohama, "A sphere theorem for manifolds of positive Ricci curvature", *Trans. Amer. Math. Soc.* **275** (1983), 811–819.

[Smale 1961] S. Smale, "Generalized Poincaré's conjecture in dimensions greater than four", *Ann. of Math.* **74** (1961), 391–466.

[Wei 1995] G. Wei, "Aspects of positive curved spaces", Ph.D. Thesis, SUNY Stony Brook, 1989.

[Whitehead 1978] G. Whitehead, *Elements of homotopy theory*, Graduate Texts in Math. **61**, Springer, New York, 1978.

[Wu 1988] H.-H. Wu, *The Bochner technique in differential geometry*, Mathematical reports **3**, part 2, Harwood, Chur (Switzerland) and New York, 1988.

[Yau 1976] S.-T. Yau, "Some function-theoretic properties of complete Riemannian manifolds and their applications to geometry", *Indiana Math. J.* **25** (1976), 659–670.

[Zhu 1990] S. Zhu, "A finiteness theorem for Ricci curvature in dimension three", *Bull. Amer. Math. Soc.* **23** (1990), 423–426.

Shunhui Zhu
Department of Mathematics
Dartmouth College
Hanover, NH 03755