Toponogov's Theorem and Applications

by

Wolfgang Meyer

These notes have been prepared for a series of lectures given at the College on Differential Geometry at Trieste in the Fall of 1989. The lectures center around Toponogov's triangle comparison theorem, critical point theory and applications. In the short amount of time available not all the aspects can be covered. We focus on those applications which seem to be most important and at the same time most suitable for an exposition. Some basic knowledge in geometry will be assumed. It has been provided by K. Grove in the first series of these lectures. Nevertheless we try to keep the lectures selfcontained and independent as much as possible. For the result about the sum of Betti numbers in section 3.5 a lemma from algebraic topology is needed. A proof for this result has been provided in the appendix.

I am indebted to U. Abresch for many helpful conversations and also for writing and typing the appendix.

Contents

1 Review of notation and some tools

1.1 Covariant derivatives

We consider a complete Riemannian manifold M with tangent bundle TM and Riemannian metric $\langle\ ,\ \rangle$ and corresponding covariant derivative ∇ of Levi Civita, which is the unique torsion free connection for which $\langle\ ,\ \rangle$ is parallel, i.e. for any vector fields X, Y, Z on M we have

$$\nabla_X Y - \nabla_Y X = [X, Y] \tag{1}$$

and

$$X \langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle . \tag{2}$$

The last two equations are equivalent to the Levi Civita equation

$$2 \langle \nabla_X Y, Z\rangle \ = \ X \langle Y, Z\rangle + Y \langle Z, X\rangle - Z \langle X, Y\rangle$$

$$+\langle Z,[X,Y]\rangle + \langle Y,[Z,X]\rangle - \langle X,[Y,Z]\rangle \tag{3}$$

If $\tilde{M}$ is an arbitrary manifold and $f:\tilde{M}\to M$ a differentiable map, $f_*:T\tilde{M}\to TM$ denotes the differential of f. ∇ naturally extends to a covariant derivative for vector fields along f. For any vector field A on $\tilde{M}$ and any vector field Y along f, i.e. $Y:\tilde{M}\to TM$ satisfies $\pi\circ Y=f$ where $\pi:TM\to M$ denotes the projection, the covariant derivative $\nabla_A Y$ is well defined. Due to the fact that $(\nabla_A Y)_p$ depends only on A_p and the values of Y in a neighbourhood of the point p, this extension is uniquely determined by requiring the chain rule $\nabla_v(X_f) = \nabla_{f_* v}X$ for any tangentvector $v\,\epsilon\, T\tilde{M}$ and any vector field X on M.

In a similar way the corresponding covariant derivative for tensor fields carries over to a covariant derivative for tensorfields along a map. As a consequence one obtains for example the Cartan structural equations for the Levi Civita connection:

$$\nabla_A f_* B - \nabla_B f_* A - f_*[A,B] = 0 \tag{4}$$

$$R(f_*A, f_*B)Y = \nabla_A\nabla_B Y - \nabla_B\nabla_A Y - \nabla_{[A,B]}Y, \tag{5}$$

where R is the curvature tensor of ∇, A, B are vector fields on $\tilde{M}$ and Y is a vector field along the map f.

For a curve $c:I\to M$ the parameter vector field on I with respect to the parameter t will be denoted by $\frac{\partial}{\partial t}$ or D_t, $\dot{c}(t) = c_*\frac{\partial}{\partial t}|_t$ is the the tangent vector of c at t. The covariant derivative $\nabla_{D_t}Y$ for a vector field Y along c is abbreviated by Y'. A parallel vector field Y along c is characterized by the linear differential equation $Y'=0$, a geodesic curve by the non-linear second order equation $\dot{c}'=0$. For consistency reasons we avoid the often found notation $\nabla_{\dot{c}}\dot{c}$ resp. $\nabla_{\dot{c}}Y$ for the expressions $\nabla_{D_t}\dot{c}$ resp. $\nabla_{D_t}Y$ when Y is a vector field along c. The inconsistency of such notation becomes apparent when c is a singular curve for example a constant curve and Y a non-constant vector field along c. If X is a vector field on M, $\nabla_{\dot{c}}X = \nabla_{D_t}X_c$ (chain rule) is well defined.

The exponential map $\exp:TM\to M$ is determined by the initial value problem for geodesics. If $v\,\epsilon\, T_pM$, then $\exp(v)=c(1)$ where c is the geodesic with initial condition $c(0)=p$ and $\dot{c}=v$. The restriction of exp to the tangent space T_pM at p is denoted by $\exp_p$. Notice that for complete manifolds the exponential map is defined on all of TM by the Hopf-Rinow theorem.

For a function $f:M\to\mathbb{R}$ and a vector field X on M, Xf denotes the derivative of f in direction X. The gradient of f is defined via the equation

$$\langle \operatorname{grad} f, X\rangle = Xf \tag{6}$$

and the Hessian Hess$f$ of f by

$$\mathrm{Hess}f\,(X) = \nabla_X \mathrm{grad}\, f\,. \tag{7}$$

Hess$f$ is a selfadjoint endomorphism field, i.e.$\langle \nabla_X \mathrm{grad}\, f\,, Y\rangle = \langle \nabla_Y \mathrm{grad}\, f\,, X\rangle$.

Important functions on a Riemannian manifold are distance functions or local distance functions from some point in M or from a submanifold of M. A local distance function is a function in an open subset U of M considered as a Riemannian submanifold. If $p \in U \subset M$ and $r(q) = dist_M(p,q)$, $r_U(q) = dist_U(q,p)$ then $r_U(q) \geq r(q)$. r_U may be differentiable in points where r fails to be differentiable. A typical example arises as follows: Let $c : [\alpha, \beta] \to M$ be an injective geodesic segment with initial point $p = c(\alpha)$ and without conjugate points. Then there is a neighborhood U of $c(]\alpha, \beta])$ where r_U is differentiable. However r is not differentiable in any point of the cut locus of p. For explicit examples look at geodesics on a cylinder.

On the set of points where a (local) distance function is differentiable it satisfies $\|\mathrm{grad}\, f\| = 1$. The gradient lines of any function with this property are geodesics parametrized by arc length, since $\left\langle \nabla_{\mathrm{grad}\, f} \mathrm{grad}\, f\,, X \right\rangle = \langle \mathrm{Hess}f\ \mathrm{grad}\, f\,, X\rangle =$ $\langle \mathrm{Hess}f\, X, \mathrm{grad} f\rangle = \langle \nabla_X \mathrm{grad}\, f\,, \mathrm{grad}\, f\rangle = \frac{1}{2} X \langle \mathrm{grad} f, \mathrm{grad} f\rangle = 0$ for any vector field X on M and hence $\nabla_{\mathrm{grad}\, f} \mathrm{grad} f = 0$. Therefore the level surfaces of such a function are equidistant. They are referred to as a family of parallel surfaces.

1.2 Jacobi fields

Jacobi fields J along a geodesic arise naturally as variational vector fields in one parameter families of geodesic lines and are characterized by the linear second order differential equation

$$J'' + R(J, \dot{c})\dot{c} = 0. \tag{8}$$

If V is a geodesic variation of c, i.e. $V : I \times (-\varepsilon, \varepsilon) \to M$ is differentiable and $V(t,0) = c(t)$ and $t \mapsto V(t,s)$ is a geodesic for all $s \in (-\varepsilon, \varepsilon)$, then $J(t) = V_* \frac{\partial}{\partial s}|_{t,0}$ is a Jacobi field along c:

$$\begin{aligned} J''(t) &= \nabla_{D_t} \nabla_{D_t} V_* D_s\,|_{t,0} = \nabla_{D_t} \nabla_{D_s} V_* D_t\,|_{t,0} + \nabla_{D_t} V_* \underbrace{[D_s, D_t]}_{=0}|_{t,0} \\ &= \nabla_{D_t} \nabla_{D_s} V_* D_t\,|_{t,0} - \nabla_{D_s} \underbrace{\nabla_{D_t} V_* D_t}_{=0}|_{t,0} \\ &= -R(V_* D_s, V_* D_t) V_* D_t\,|_{t,0} = -R(J, \dot{c})\dot{c}\,_t. \end{aligned}$$

Therefore the Jacobi equation is the linearization of the geodesic equation along c. Notice that V can be written in the following way: If p is the curve $p(s) = V(0,s)$ and Y the vector field along p given by $Y(s) = V_* D_t\,|_{0,s}$, then $V(t,s) = \exp tY(s)$. The initial conditions of the Jacobi field in terms of p and Y are $J(0) = \dot{p}(0)$, $J'(0) = Y'(0)$. $Y(0)$ is the initial vector of the geodesic c. Any tangent vector u to TM can be written as the tangent vector $u = \dot{Y}|_0$ of a curve $s \mapsto Y|_s \in TM$. Y is a vector field along the base curve $p(s) = \pi \circ Y|_s$. If Y and V are defined as above, we find $\exp_* u = \widehat{\exp \circ Y}|_0 = V_* D_s|_{1,0} = J(1)$. Therefore the differential of the exponential map is completely determined by Jacobi fields.

For example, the Jacobi field with initial conditions $J(0) = 0$, $J'(0) = w$ along the geodesic $\exp tv$ is obtained from the variation $V(t,s) = \exp t(v + sw)$. Here $p(s)$ is the constant curve, $Y(s) = v + sw$, $J(t) = \exp_* |_{tv} tw$, $J(1) = exp_{p\,*}|_v w$. This shows that the differential of the restriction $exp|_{T_pM}$ is determined by Jacobi fields on M with these initial conditions.

1.3 Interpretation of curvature in terms of the distance function

Consider two geodesics c_0, c_1 emanating from a point p in M, $c_0(\varepsilon) = \exp \varepsilon v$, $c_1(\varepsilon) = \exp \varepsilon w$, $v, w \in T_pM$ and the distance $L(\varepsilon) = dist(c_o(\varepsilon), c_1(\varepsilon))$ in a neighborhood of zero. Then the fourth order Taylor formula for L^2 is given by

$$L^2(\varepsilon) = \varepsilon^2 \|v - w\|^2 - \tfrac{1}{3}\varepsilon^4 \langle R(v,w)w, v\rangle + O(\varepsilon^5)\,. \tag{9}$$

When $v \neq w$ this implies for $\varepsilon \geq 0$:

$$L(\varepsilon) = \varepsilon \|v - w\| - \frac{1}{6}\frac{\langle R(v,w)w, v\rangle}{\|v - w\|}\varepsilon^3 + O(\varepsilon^4)\,. \tag{10}$$

For linearly independent vectors v, w satisfying $\|v\| = \|w\| = 1$ this can be rewritten as

$$L(\varepsilon) = \varepsilon \|v - w\| \left(1 - \tfrac{1}{12} K(v,w)(1 + \langle v, w\rangle)\varepsilon^2\right) + O(\varepsilon^4)\,, \tag{11}$$

where $K(v,w)$ ist the sectional curvature of the plane spanned by v and w. Therefore L grows faster than linear if $K < 0$ and slower than linear if $K > 0$ in a neighborhood of 0.

To prove (9) we consider the variation

$$V(\varepsilon, t) = \exp(t \exp^{-1}_{c_0(\varepsilon)} \circ c_1(\varepsilon))$$

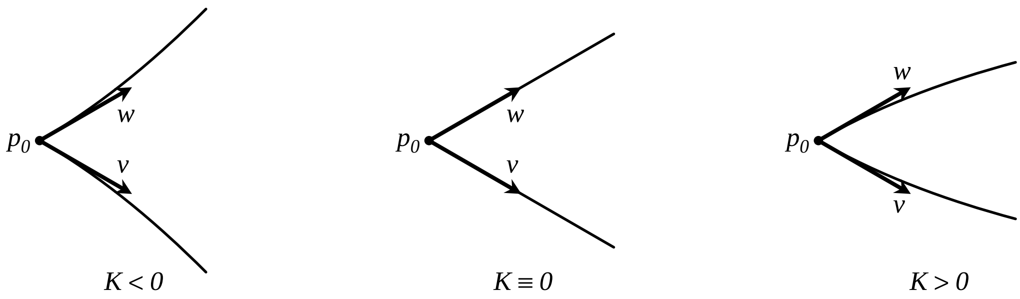

Figure 1: interpretation of sectional curvature

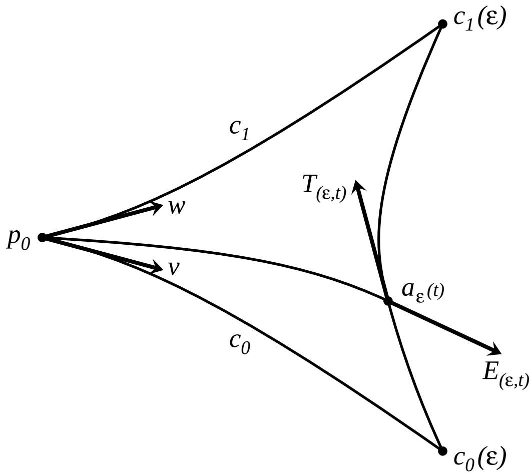

Figure 2: setup for the proof of (9)

for small values of ε and $t \in [0,1]$. The parameter tangent fields along V are $E = V_* D_\varepsilon$ and $T = V_* D_t$. The parameter curves $a_\varepsilon : t \mapsto V(\varepsilon, t)$ are geodesics connecting the points $c_0(\varepsilon)$ and $c_1(\varepsilon)$. T is the tangent field of the geodesics and $t \mapsto E|_{\varepsilon,t}$ is a Jacobi field along a_ε and $E|_{\varepsilon,0} = \dot{c}_0(\varepsilon)$, $E|_{\varepsilon,1} = \dot{c}_1(\varepsilon)$.

Notice that $\|\dot{a}_\varepsilon(t)\|$ is the length of a_ε so that

$$L(\varepsilon) = \|\dot{a}_\varepsilon(t)\| = \|T\|_{\varepsilon,t} \tag{12}$$

which is constant in t for ε fixed. The derivatives of $H = L^2$ up to the fourth order are given by

$$\begin{aligned}
H'(\varepsilon) &= 2\left\langle \nabla_{D_\varepsilon} T, T\right\rangle|_{\varepsilon,t} \\
H''(\varepsilon) &= 2\left(\left\langle \nabla^2_{D_\varepsilon} T, T\right\rangle + \left\langle \nabla_{D_\varepsilon} T, \nabla_{D_\varepsilon} T\right\rangle\right)|_{\varepsilon,t} \\
H'''(\varepsilon) &= 2\left(\left\langle \nabla^3_{D_\varepsilon} T, T\right\rangle + 3\left\langle \nabla^2_{D_\varepsilon} T, \nabla_{D_\varepsilon} T\right\rangle\right)|_{\varepsilon,t} \\
H^{IV}(\varepsilon) &= 2\left(\left\langle \nabla^4_{D_\varepsilon} T, T\right\rangle + 4\left\langle \nabla^3_{D\varepsilon} T, \nabla_{D_\varepsilon} T\right\rangle + 3\left\langle \nabla^2_{D_\varepsilon} T, \nabla^2_{D_\varepsilon} T\right\rangle\right)|_{\varepsilon,t}\ .
\end{aligned}$$

We will now evaluate these derivatives at $(0,t)$ in order to find the coefficients for the Taylor formula. The equation $\nabla_{D_\varepsilon} T = \nabla_{D_t} E$ and $\nabla_{D_t} T = 0$ will be used frequently during this calculation. Also notice that $T|_{0,t} = 0$, since $V(0,t) = p$. We have

$$\nabla_{D_\varepsilon} E\,|_{\varepsilon,0} = 0,\ \ \nabla_{D_\varepsilon} E\,|_{\varepsilon,1} = 0 \tag{13}$$

since $E|_{\varepsilon,0} = \dot{c}_0(\varepsilon)$ and $E|_{\varepsilon,1} = \dot{c}_1(\varepsilon)$. From the Jacobi property of E we obtain

$$\nabla_{D_t} \nabla_{D_t} E\ = -R(E,T)T\,, \tag{14}$$

so that

$$\nabla_{D_t} \nabla_{D_t} E\,|_{0,t} = 0\,. \tag{15}$$

Hence $t \mapsto E|_{0,t}$ is a linear vectorfield along the constant curve a_0. Since $E|_{0,0} = \dot{c}_0(0) = v$, $E|_{0,1} = \dot{c}_1(0) = w$ it follows

$$E|_{(0,t)} = v + t(w-v)\,. \tag{16}$$

With this information we can already evaluate $H'(0)$ and $H''(0)$:

$$\begin{aligned}
H'(0) &= 2\left\langle \nabla_{D_\varepsilon} T, T\right\rangle|_{0,t} = 0 && (17)\\
H''(0) &= 2\left\langle \nabla^2_{D_\varepsilon} T, T\right\rangle|_{0,t} + 2\left\langle \nabla_{D_\varepsilon} T, \nabla_{D_\varepsilon} T\right\rangle|_{0,t} \\
&= 2\left\langle \nabla_{D_t} E, \nabla_{D_t} E\right\rangle|_{0,t} \\
&= 2\|v-w\|^2 && (18)
\end{aligned}$$

from (16). Next we show that

$$\begin{aligned}
\nabla_{D_\varepsilon} E|_{0,t} &= 0 && (19)\\
\nabla_{D_t}\nabla_{D_\varepsilon} E|_{0,t} &= 0 && (20)\\
\nabla_{D_\varepsilon}\nabla_{D_\varepsilon} T|_{0,t} &= 0\,. && (21)
\end{aligned}$$

(20) is a consequence of (19) and (21) follows from (20) since

$$\nabla_{D_\varepsilon} \nabla_{D_\varepsilon} T\ = \nabla_{D_\varepsilon} \nabla_{D_t} E\ = R(E,T)E + \nabla_{D_t} \nabla_{D_\varepsilon} E\ .$$

In view of the equations (13) above it suffices to show $\nabla_{D_t}\nabla_{D_t}\nabla_{D_\varepsilon}E\,|_{0,t} = 0$ for the proof of (19). For this observe

$$\nabla_{D_t}\nabla_{D_t}\nabla_{D_\varepsilon}E \;=\; \nabla_{D_t}(R(T,E)E) + R(T,E)\nabla_{D_\varepsilon}T + \nabla_{D_\varepsilon}(R(T,E)T)\;.$$

The right hand side vanishes at $(0,t)$ since $\nabla_{D_t}T = 0$ and $T|_{0,t} = 0$. This suffices to find $H'''(0)$:

$$\begin{aligned}
H'''(0) &= 6\,\langle \nabla_{D_\varepsilon}\nabla_{D_\varepsilon}T\,,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} \\
&= 6\,\langle \nabla_{D_\varepsilon}\nabla_{D_t}E,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} \\
&= 6\,\langle R(E,T)E,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} + 6\,\langle \nabla_{D_t}\nabla_{D_\varepsilon}E\,,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} \\
&= 0
\end{aligned} \tag{22}$$

from (20). From (21) we get

$$H^{IV}(0) = 8\,\langle \nabla_{D_\varepsilon}\nabla_{D_\varepsilon}\nabla_{D_\varepsilon}T\;\;,\nabla_{D_\varepsilon}T\rangle\,|_{0,t}. \tag{23}$$

Furthermore

$$\begin{aligned}
\nabla_{D_\varepsilon}\nabla_{D_\varepsilon}\nabla_{D_\varepsilon}T|_{0,t} &= \nabla_{D_\varepsilon}\nabla_{D_\varepsilon}\nabla_{D_t}E|_{0,t} \\
&= (\nabla_{D_\varepsilon}R(E,T)E + \nabla_{D_\varepsilon}\nabla_{D_t}\nabla_{D_\varepsilon}E)|_{0,t} \\
&= R(E,\nabla_{D_t}E)E|_{0,t} + \nabla_{D_\varepsilon}\nabla_{D_t}\nabla_{D_\varepsilon}E|_{0,t}\,.
\end{aligned} \tag{24}$$

Using (16),(23),(24) and the symmetries of R we find

$$H^{IV}(0) = 8\,\langle R(v,w)v,w\rangle + \langle \nabla_{D_\varepsilon}\nabla_{D_t}\nabla_{D_\varepsilon}E\;,\nabla_{D_\varepsilon}T\rangle\,|_{0,t}\,.$$

Since this must also be constant in t, the second term on the right hand side is constant in t. Now

$$\begin{aligned}
\langle \nabla_{D_\varepsilon}\nabla_{D_t}\nabla_{D_\varepsilon}E,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} &= D_\varepsilon D_t\,\langle \nabla_{D_\varepsilon}E,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} \\
&= D_t D_\varepsilon\,\langle \nabla_{D_\varepsilon}E,\nabla_{D_\varepsilon}T\rangle\,|_{0,t}
\end{aligned}$$

by using (20) and (21). Therefore

$$D_\varepsilon\,\langle \nabla_{D_\varepsilon}E\,,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} = \langle \nabla_{D_\varepsilon}\nabla_{D_\varepsilon}E\,,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} + \langle \nabla_{D_\varepsilon}E\,,\nabla_{D_\varepsilon}\nabla_{D_\varepsilon}T\,\rangle\,|_{0,t}$$

must be linear in t. But $\nabla_{D_\varepsilon}\nabla_{D_\varepsilon}E\,|_{0,0} = \nabla_{D_\varepsilon}\nabla_{D_\varepsilon}\dot{c}_0(0) = 0$, $\nabla_{D_\varepsilon}\nabla_{D_\varepsilon}E\,|_{0,1} = \nabla_{D_\varepsilon}\nabla_{D_\varepsilon}\dot{c}_1(0) = 0$ and $\nabla_{D_\varepsilon}\nabla_{D_\varepsilon}T\,|_{0,t} = 0$, so that $D_\varepsilon\,\langle \nabla_{D_\varepsilon}E,\nabla_{D_\varepsilon}T\rangle\,|_{0,t} = 0$ since it vanishes at $t=0$ and at $t=1$. This proves

$$H^{IV}(0) = 8\,\langle R(v,w)v,w\rangle\;. \tag{25}$$

Equation (9) now follows from (17), (18), (22) and (25). We leave it to the reader to verify

$$H^V(0) = 10\,\langle(\nabla_{v+w}R)(v,w)v,w\rangle\ . \tag{26}$$

If $H^V(0)=0$ for all choices of v and w, then M must be a locally symmetric space, since (26) can be used to show that the operator $R_{\dot c} = R(..,\dot c)\dot c$ is parallel for any geodesic c.

1.4 The levels of a distance function

In this section we will see that Jacobi fields determine the second fundamental tensor S of the level surfaces of a (local) distance function f. This will be used to establish the Riccati equation for S and a Riccati inequality.

We have a natural unit normal vector field $N = \operatorname{grad} f$ along the level surfaces of f. The second fundamental tensor S of the levels with respect to N is the restriction of the Hessian of f to the tangent spaces of the levels , $Su = \operatorname{Hess} f\, u = \nabla_u N$ for tangent vectors u to the levels. The derivative $S' = \nabla_N S$ in the normal direction is defined by $S'Y = (\nabla_N S)Y = \nabla_N(SY) - S(\nabla_N Y)$ for any vector field Y tangent to the levels of f. Notice that $S'Y$ is again tangent to the levels.
Let M_0 be a fixed level, $M_0 = f^{-1}\{0\}$ after changing f by a constant. The other levels are then given by $M_t = f^{-1}\{t\}$. For small values of t the levels M_0 and M_t are diffeomorphic via the diffeomorphism $E_t(p) = \exp(t\,N(p))$. The differential of $E_t|_{M_o}$ can be desribed in terms of Jacobi fields: Let $s \mapsto p(s)$ be a curve in M_o with tangent vector $v = \dot p(0)$. Then $E_{t*}v = J(t)$ where J is the Jacobi field along the geodesic $t \mapsto E_t(p(0))$ of the geodesic variation $V(t,s) = E_t \circ p(s)$, $J(t) = V_* D_s|_{t,0}$. Its initial conditions are $J(0) = \dot p(0) = v$, $J'(0) = \nabla_{D_s}(N\circ p)|_0 = \nabla_{\dot p(0)}N = Sv$, compare section 1.2.
The geodesic $\gamma(t) = V(t,s)$ is an integral curve of N, so that $V_*D_t|_{t,s} = \dot\gamma(t) = N\circ V(t,s)$. With this information we obtain $J'(t) = \nabla_{D_t}V_*D_s\,|_{t,0} = \nabla_{D_s}V_*D_t\,|_{t,0} = \nabla_{D_s}N\circ V\,|_{t,0} = \nabla_{V_*D_s}N\,|_{t,0} = SJ(t)$. The second fundamental tensor of the levels now is determined by

$$SJ = J'\ . \tag{27}$$

Covariant differentiation of this equation leads to the important Riccati equation for S: Since (27) is an equation along the geodesic c(t)=V(t,0) it reads more precisely $S_cJ = J'$. This is useful to remember for the chain rule in $\nabla_{D_t}S_c = \nabla_{\dot c}S = \nabla_{N_c}S = (S')_c$ for the computation of $J'' = \nabla_{D_t}(S_cJ) = S'J + SJ' = S'J + S^2J$. Using the

Jacobi equation $J'' + R(J,N)N = 0$ we obtain the Riccati equation

$$S' = -R_N - S^2 \tag{28}$$

where R_N denotes the curvature operator $R_N X = R(X,N)N$ in direction N.

If there is a lower bound κ for the sectional curvature K of M, then the Riccati equation leads to a Riccati inequality along the gradient lines c of f. Let Y be a parallel unit vector field along c tangent to the levels, i.e. $\langle Y, \dot{c}\rangle = 0$. Then by (28)

$$\begin{aligned} \langle SY, Y\rangle' &= -\langle R(Y,N)N, Y\rangle - \left\langle S^2 Y, Y\right\rangle \\ &= -K(Y,N) - \|SY\|^2 . \end{aligned}$$

From the assumption $\kappa \le K(Y,N)$ and the Schwarz inequality we obtain the Riccati inequality

$$\langle SY, Y\rangle' \le -\kappa - \langle SY, Y\rangle^2 \tag{29}$$

along c.

1.5 Data in the constant curvature model spaces

Constant curvature model spaces are important in comparison theory because the geometric quantities in these spaces can be calculated explicitly.

M^n_κ denotes the n-dimensional hyperbolic space $\mathbb{H}^n_\kappa$ of curvature κ if $\kappa < 0$, the euclidian space $\mathbb{R}^n$ if $\kappa = 0$ and the standard sphere S^n_κ of radius $\frac{1}{\sqrt{\kappa}}$ if $\kappa > 0$. Since $\langle R(v,u)u, v\rangle = \kappa$ for any pair of orthonormal vectors $u, v \in T_p M^n_\kappa$, we have $R_u := R(\ldots, u)u = \kappa \cdot \mathrm{Id}_p$ on the orthogonal complement of u in $T_p M^n_\kappa$. Therefore the Jacobi equation and the Riccati equation are rather simple.

Jacobi fields along a geodesic $c : \mathbb{R} \to M^n_\kappa$ orthogonal to $\dot{c}$ are given by $f \cdot Y$, where Y is a parallel vector field along c and $f : \mathbb{R} \to \mathbb{R}$ is a solution of the 1-dimensional Jacobi equation

$$f'' + \kappa f = 0 \tag{30}$$

Let sn_κ and cs_κ be the solutions of (30) with initial conditions $\mathrm{sn}_\kappa(0) = 0$, ${\mathrm{sn}_\kappa}'(0) = 1$ and $\mathrm{cs}_\kappa(0) = 1$, ${\mathrm{cs}_\kappa}'(0) = 0$, i.e.

$$\left.\begin{aligned} \mathrm{sn}_\kappa(t) &= \tfrac{1}{\sqrt{\kappa}} \sin \sqrt{\kappa} t \\ \mathrm{cs}_\kappa(t) &= \cos \sqrt{\kappa}\, t \end{aligned}\right\} \text{for} \quad \kappa > 0$$
$$\left.\begin{aligned} \mathrm{sn}_\kappa(t) &= t \\ \mathrm{cs}_\kappa(t) &= 1 \end{aligned}\right\} \text{for} \quad \kappa = 0 \tag{31}$$

$$\left.\begin{array}{rcl} \mathrm{sn}_\kappa(t) & = & \frac{1}{\sqrt{|\kappa|}}\sinh\sqrt{|\kappa|}\,t \\ \mathrm{cs}_\kappa(t) & = & \cosh\sqrt{|\kappa|}\,t \end{array}\right\}\text{for}\quad \kappa<0$$

Furthermore let

$$\mathrm{ct}_\kappa(t)=\mathrm{cs}_\kappa(t)/\mathrm{sn}_\kappa(t)\qquad\text{for}\ \ \mathrm{sn}_\kappa(t)\neq 0 \tag{32}$$

The derivatives of these functions are given by

$$\mathrm{sn}_\kappa{}'=\mathrm{cs}_\kappa\,,\quad \mathrm{cs}_\kappa{}'=-\kappa\,\mathrm{sn}_\kappa\,,\quad \mathrm{ct}_\kappa{}'=-\kappa-\mathrm{ct}_\kappa{}^2\,. \tag{33}$$

Furthermore the following elementary equations hold:

$$1 = \mathrm{cs}_\kappa{}^2+\kappa\,\mathrm{sn}_\kappa{}^2 \tag{34}$$

$$\mathrm{sn}_\kappa(a+b) = \mathrm{sn}_\kappa(a)\mathrm{cs}_\kappa(b)+\mathrm{cs}_\kappa(a)\mathrm{sn}_\kappa(b) \tag{35}$$

$$\mathrm{cs}_\kappa(a+b) = \mathrm{cs}_\kappa(a)\mathrm{cs}_\kappa(b)-\kappa\,\mathrm{sn}_\kappa(a)\mathrm{sn}_\kappa(b)\,. \tag{36}$$

A basis for the Jacobi fields orthogonal to $\dot c$ is given by $\{\mathrm{sn}_\kappa\cdot Y,\mathrm{cs}_\kappa\cdot Y\}$ where Y varies over a basis of parallel vector fields orthogonal to $\dot c$.

Notice that the second fundamental tensor of the (local) distance spheres at distance r from a fixed point p in any manifold is determined by equation (27), where J is a Jacobi field with initial value $J(0)=0$ along a normal geodesic emanating from p, i.e. in M^n_κ by

$$J(r)=\mathrm{sn}_\kappa(r)Y(r)$$

with Y parallel along c and $\langle Y,\dot c\rangle=0$. Hence

$$S_{c(r)}Y(r)=\frac{\mathrm{sn}_\kappa{}'(r)}{\mathrm{sn}_\kappa(r)}Y(r)=\mathrm{ct}_\kappa(r)Y(r)\,. \tag{37}$$

Therefore the principal curvatures of distance spheres in M^n_κ are equal to $\mathrm{ct}_\kappa(r)$. The length of the great circles in the distance spheres is $2\pi\,\mathrm{sn}_\kappa(r)$. In any manifold the Hessian of the distance function from a point has a zero eigenvalue in the radial direction. For Karcher's new proof of Toponogov's theorem it is convenient to rescale the distance function f from the point p in M^n_κ so that all the eigenvalues are equal. This is achieved by taking $\mathrm{md}_\kappa\circ f$, where

$$\mathrm{md}_\kappa(r)=\int_0^r\mathrm{sn}_\kappa(t)dt=\begin{cases}\frac{1}{\kappa}(1-\mathrm{cs}_\kappa(r)) & \text{for}\ \ \kappa\neq 0\\ \frac{1}{2}r^2 & \text{for}\ \ \kappa=0\end{cases} \tag{38}$$

Notice the identity

$$\mathrm{cs}_\kappa+\kappa\,\mathrm{md}_\kappa=1\,. \tag{39}$$

From the formula

$$\begin{aligned}\mathrm{Hess}(\mathrm{md}_\kappa\circ f)v &= (\mathrm{md}_\kappa{}'\circ f)\mathrm{Hess}f(v) + (\mathrm{md}_\kappa{}''\circ f)\,\langle \mathrm{grad}f, v\rangle\,\mathrm{grad}f \\ &= (\mathrm{sn}_\kappa\circ f)\mathrm{Hess}f(v) + (\mathrm{cs}_\kappa\circ f)\,\langle \mathrm{grad}f, v\rangle\,\mathrm{grad}f \qquad (40)\end{aligned}$$

it follows that the eigenvalues of $\mathrm{Hess}(\mathrm{md}_\kappa\circ f)$ at a point q with $f(q) = r$ are equal to $\mathrm{cs}_\kappa\circ f(q) = \mathrm{cs}_\kappa(r)$. Using md_κ, the law of cosines in M^n_κ becomes

$$\mathrm{md}_\kappa(c) = \mathrm{md}_\kappa(a-b) + \mathrm{sn}_\kappa(a)\mathrm{sn}_\kappa(b)(1-\cos\gamma) \qquad (41)$$

where a, b, c are the lengths of the edges of a geodesic triangle in M_κ and γ is the angle opposite to the edge corresponding to c. Notice that this is a unified formula for the three classical cases $\kappa = 0, \kappa > 0, \kappa < 0$:

$$\begin{aligned} c^2 &= a^2 + b^2 - 2ab\cos\gamma & (42)\\ \cos(\sqrt{\kappa}\,c) &= \cos(\sqrt{\kappa}\,a)\cos(\sqrt{\kappa}\,b) + \sin(\sqrt{\kappa}\,a)\sin(\sqrt{\kappa}\,b)\cos\gamma & (43)\\ \cosh(\sqrt{|\kappa|}\,c) &= \cosh(\sqrt{|\kappa|}\,a)\cosh(\sqrt{|\kappa|}\,b) - \sinh(\sqrt{|\kappa|}\,a)\sinh(\sqrt{|\kappa|}\,b)\cos\gamma & (44)\end{aligned}$$

1.6 The Riccati comparison argument

A lower curvature bound κ in M leads to an important estimate for the principal curvatures in distance spheres and hence for the tangential eigenvalues of the Hessian of the distance function f from a point. For the modified distance function $\mathrm{md}_\kappa\circ f$ this yields an estimate for all the eigenvalues. This estimate is the key for Karchers proof of Toponogov's theorem and the main reason for introducing md_κ. The basic comparison argument is contained in (i) of the following elementary Lemma and its Corollary, cf. [K].

Lemma 1.1 *Suppose g, G are differentiable functions on some interval satisfying the Riccati inequalities*

$$\begin{aligned} g' &\le -\kappa - g^2 & (45)\\ G' &\ge -\kappa - G^2\,. & (46)\end{aligned}$$

i) If $g(r_0) \ge G(r_0)$, then $g(r) \ge G(r)$ for $r \le r_0$.

ii) If $g(r_0) \le G(r_0)$, then $g(r) \le G(r)$ for $r \ge r_0$.

Proof. From the two Riccati inequalities (45) and (46) we get

$$[(g-G)\cdot \mathrm{e}^{\int (g+G)}]' \leq 0$$

from which i) and ii) follow immediately. □

The statement ii) is useful for estimates involving upper curvature bounds [K]. We are interested mainly in i).

Corollary 1.2 *If* $g:(0,a)\to I\!R$ *(suppose* $a\leq \frac{\pi}{\sqrt{\kappa}}$ *if* $\kappa>0$*) satisfies* $g'\leq -\kappa-g^2$ *and* $\lim_{r\to 0} g(r)=\infty$*, then*

$$g(r)\leq ct_\kappa(r)\,.$$

Proof. If there is a point $r_0\in(0,a)$ for which $g(r_0)>\mathrm{ct}_\kappa(r_0)$, we can choose $\varepsilon>0$ so that $g(r_0)\geq \mathrm{ct}_\kappa(r_0-\varepsilon)$. $G(r)=\mathrm{ct}_\kappa(r-\varepsilon)$ satisfies the Riccati equation $G'=-\kappa-G^2$ on (ε,r_0), so that $g(r)\geq G(r)$ on (ε,r_0). Then $g(\varepsilon)=\lim_{r\searrow\varepsilon} g(r)\geq \lim_{r\searrow\varepsilon} G(r)=+\infty$, contradicting $g(\varepsilon)<\infty$. □

Consider now a normal geodesic segment c with initial point p which does not meet the conjugate locus of p. In a neighborhood U of c we may consider the local distance function $f(q)=dist_U(p,q)$. The principal curvatures of the local distance sphere $f^{-1}(r)$ at the point q are denoted by $\tau_1(q),\ldots,\tau_{n-1}(q)$. From the corollary and (29) we get the estimate

$$\tau_i(q)\leq \mathrm{ct}_\kappa(f(q))\,.$$

$\tau_i(q)$ are the eigenvalues of $\mathrm{Hess}f|_q$ corresponding to eigenvectors tangent to the distance sphere, whereas the radial eigenvalue is zero. The hessean of $\mathrm{md}_\kappa\circ f$ satisfies the corresponding equation (40) and therefore has eigenvalues $\mathrm{sn}_\kappa(f(q))\cdot\tau_i(q)$, $i=1,\ldots,n-1$ in directions tangent to the level r and the eigenvalue $\mathrm{cs}_\kappa(f(q))$ for the radial direction $\mathrm{grad}f|_q$. This proves the operator inequality

$$\mathrm{Hess}(\mathrm{md}_\kappa\circ f)\leq(\mathrm{cs}_\kappa\circ f)\cdot\mathrm{Id}\,. \tag{47}$$

Along c this estimate remains true up to the first conjugate point of c, which in the case $\kappa>0$ appears at a distance not farther away than $\frac{\pi}{\sqrt{\kappa}}$. For $M=M^2_\kappa$ equality holds in (47).

If f is replaced by $g=f+\eta$ where η is a constant, we have $\mathrm{Hess}\,g=\mathrm{Hess}f$ so that the tangential eigenvalues of $\mathrm{Hess}(\mathrm{md}_\kappa\circ g)|q$ according to formula (40) are given by $\mathrm{sn}_\kappa(g(q))\tau_i(q)$ and the radial eigenvalue is $\mathrm{cs}_\kappa(g(q))$. The estimate for $\tau_i(q)$ above leads to $(\mathrm{sn}_\kappa \circ g)\tau_i \leq (\mathrm{sn}_\kappa \circ g)\mathrm{ct}_\kappa(g-\eta) = \mathrm{cs}_\kappa \circ g + \frac{\mathrm{sn}_\kappa(\eta)}{\mathrm{sn}_\kappa(g-\eta)}$. For small values of η and $0 < g - \eta < \frac{\pi}{\sqrt{\kappa}}$ in the case $\kappa > 0$ the Hessian of $\mathrm{md}_\kappa \circ g$ satisfies consequently

$$\mathrm{Hess}(\mathrm{md}_\kappa \circ g) \leq (\mathrm{cs}_\kappa \circ g + \frac{\mathrm{sn}_\kappa(\eta)}{\mathrm{sn}_\kappa(g-\eta)}) \cdot \mathrm{Id}\,. \tag{48}$$

In the case $\kappa > 0$ this estimate along c holds up to the first conjugate point.

2 The Toponogov Theorem

The Toponogov comparison theorem appears to be one of the most powerful tools in Riemannian geometry. It is a global generalization of the first Rauch comparison theorem. The ideas trace back to A.D. Alexandrow who first proved the theorem for convex surfaces. Toponogov's proof of the theorem was technical and contained some difficulties which were resolved in [GKM]. Since then the proof had been simplified considerably by various geometers, compare also [CE]. In this lecture series we shall use an interesting new proof given by Karcher [K]. In contrast to the previous technique the Rauch comparison theorem is not used at all. It uses the estimate for the Hessian given in (47) resp.(48) and fits nicely into our discussion of distance functions. This does not mean, that our approach is necessarily shorter or more geometric than the other viable arguments given before. We certainly encourage the student also to go through some alternate proof of Toponogov's basic result in the literature mentioned above.

Definition 2.1 *A* geodesic hinge c, c_0, α *in* M *consists of two non constant geodesic segments* c, c_o *with the same initial point making the angle* α. *A minimal connection* c_1 *between the endpoints of* c *and* c_o *is called a* closing edge *of the hinge.*

The length of a geodesic segment c will be denoted by $|c|$.

Theorem 2.2 (Toponogov) *Let* M *be a complete Riemannian manifold with sectional curvature* $K \geq \kappa$.

A) Given points p_0, p_1, q *in* M *satisfying* $p_0 \neq q$, $p_1 \neq q$, *a non constant geodesic* c *from* p_0 *to* p_1 *and minimal geodesics* c_i, *from* p_i *to* q, $i = 0, 1$, *all parametrized by arc length. Suppose the triangle inequality* $|c| \leq |c_1| + |c_2|$ *is satisfied and* $|c| \leq \frac{\pi}{\sqrt{\kappa}}$ *in the case* $\kappa > 0$. $\alpha_i \in [0, \pi]$ *denote the angles at* p_i, $\alpha_0 = \measuredangle(\dot{c}_o(0), \dot{c}(0))$, $\alpha_1 = \measuredangle(\dot{c}_1(0), -\dot{c}(|c|)$. *Then there exists a corresponding comparison triangle* $\tilde{p}_0, \tilde{p}_1, \tilde{q}$ *in the model space* M^2_κ *with corresponding geodesics* $\tilde{c}_0, \tilde{c}_1, \tilde{c}$ *which are all minimal of lengths* $|\tilde{c}_i| = |c_i|, |\tilde{c}| = |c|$ *and* *i) the corresponding angles $\tilde{\alpha}_i$ satisfy $\tilde{\alpha}_i \leq \alpha_i$*

ii) $dist(\tilde{q}, \tilde{c}(t)) \leq dist(q, c(t))$ for any $t \in [0, |c|]$.

Except for the case when $\kappa > 0$ and one of the geodesics has length equal to $\frac{\pi}{\sqrt{\kappa}}$ the triangle in M^2_κ is uniquely determined.

B) Let c, c_o, α_o be a hinge in M with c_o minimal and $|c| \leq \frac{\pi}{\sqrt{\kappa}}$ in case $\kappa > 0$ and c_1 a closing edge . Then the closing edge $\tilde{c}_1$ of any hinge $\tilde{c}, \tilde{c}_o, \alpha_o$ in M^2_κ with $|\tilde{c}| = |c|$, $|\tilde{c}_o| = |c_o|$ satisfies

$$|\tilde{c}_1| \geq |c_1| .$$

Remarks

1. Notice that c need not to be minimal and the case $p_0 = p_1$ is not excluded. c_1 and c_0 have to be minimal, otherwise there are counterexamples.

2. With a little effort statement (ii) can be used to show that the length of secants between c and c_i are not shorter than the corresponding secants between $\tilde{c}$ and $\tilde{c}_i$, provided the segment of c in the cut off triangle is minimal:

   iii) $\mathrm{dist}(\tilde{c}_0(t), \tilde{c}(s)) \leq \mathrm{dist}(c_0(t), c(s))$ holds as long as $c|_{[0,s]}$ is minimal,

   iv) $\mathrm{dist}(\tilde{c}_1(t), \tilde{c}(s)) \leq \mathrm{dist}(c_1(t), c(s))$ holds as long as $c|_{[s,|c|]}$ is minimal.

   In the case when c is minimal now any corresponding secants $\sigma, \tilde{\sigma}$ satisfy $|\tilde{\sigma}| \leq |\sigma|$.
   For symmetry reasons only iii) needs to be proved:

   By ii)
   $$\mathrm{dist}(\tilde{q}, \tilde{c}(s)) \leq \mathrm{dist}(q, c(s)) . \tag{49}$$

   Connect $p_s = c(s)$ and q by a minimal geodesic γ_s and consider the triangle p_0, p_s, q with geodesic edges c_0, $c|_{[o,s]}, \gamma_s$ and the corresponding comparison triangle $\tilde{p}_0, \tilde{p}_s, \tilde{q}$ in M^2_κ. Using ii) for this triangle we obtain

   $$\mathrm{dist}(\tilde{p}_s, \tilde{c}_0(t)) \leq \mathrm{dist}(c(s), c_0(t)) . \tag{50}$$

   The monotonicity relation between angle and length of the closing edge of a hinge in M^2_κ and (49) imply

   $$\sphericalangle\, \tilde{c}_0(t)\, \tilde{p}_0\, \tilde{c}(s) = \sphericalangle\, \tilde{q}\, \tilde{p}_0\, \tilde{c}(s) \leq \sphericalangle\, \tilde{q}\, \tilde{p}_0\, \tilde{p}_s = \sphericalangle\, \tilde{c}_0(t)\, \tilde{p}_0\, \tilde{p}_s$$

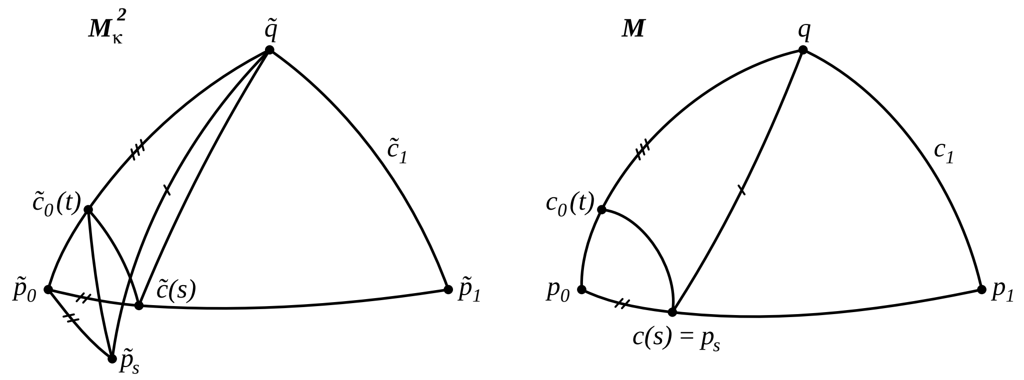

Figure 3: sketch for the proof iii)

and then

$$\mathrm{dist}(\tilde{c}(s), \tilde{c}_0(t)) \leq \mathrm{dist}(\tilde{p}_s, \tilde{c}_0(t)) \,. \tag{51}$$

Inequality iii) now follows from (50) and (51).

3. Statement i) is a consequence of ii). To prove for example $\alpha_0 \leq \alpha$, consider the functions $h_0(t) = \mathrm{dist}(c_0(t), c(t))^2$ and $\tilde{h}_0(t) = \mathrm{dist}(\tilde{c}_0(t), \tilde{c}(t))^2$ for small values of t. By iii) we have $\tilde{h}_0 \leq h_0$. According to (9) of section 1 we have the Taylor formulas

$$\begin{aligned} h_0(t) &= t^2 \|\dot{c}_0(0) - \dot{c}(0)\|^2 + O(t^4) \\ \tilde{h}_0(t) &= t^2 \|\dot{\tilde{c}}_0(0) - \dot{\tilde{c}}(0)\|^2 + O(t^4) \end{aligned}$$

so that $\|\dot{\tilde{c}}_0(0) - \dot{\tilde{c}}(0)\| \leq \|\dot{c}_0(0) - \dot{c}(0)\|$ and hence $\tilde{\alpha}_0 \leq \alpha_0$.

The converse implication i) $\Rightarrow$ ii) is also true but more technical to prove.

4. Statement ii) carries over to limits in the sense of Gromov for Riemannian spaces with curvature $K \geq \kappa$, where angles cannot be defined anymore.

5. Part B is equivalent to A)i). This follows immediately from the fact that in M^2_κ the length of a closing edge in a hinge with minimal geodesics and the hinge angle are in a monotone relation. Note that B is trivial in the case when the tiangle inequality is not satisfied in M. For this observe that the triangle inequality in M^2_κ is satisfied since all the corresponding geodesics in M^2_κ are minimal.

6. If c_0 is not minimal in B), the statement is false. For consider in $S^2_{1+\varepsilon}$ a hinge with two geodesics of length π making a positive angle. The end points have a

positive distance for ε small. However, in the corresponding hinge in S_1^2 the end points coincide.

7. An anologue of Toponogov's theorem where the lower curvature bound is replaced by an upper curvature bound is false. For example on the 3-sphere S^3 there are homogeneous metrics (Berger metrics) with positive curvature, upper curvature bound 1 and closed geodesics of length $< 2\pi$. However, if the sectional curvature K of M satisfies $K \le \kappa$ and c_0, c_1, c is a triangle with minimal geodesics and $|c_0| + |c_1| + |c| < \frac{2\pi}{\sqrt{\kappa}}$ which is contained in a ball around p_0 of radius not greater than the injectivity radius at p_0, then there is a triangle $\tilde{c}_0$, $\tilde{c}_1$, $\tilde{c}$ in M_κ^2 with $|c_i| = |\tilde{c}_i|$, $|c| = |\tilde{c}|$ and $\alpha_0 \le \tilde{\alpha}_0$. This is an immediate consequence of Rauch's first comparison theorem.

8. There are generalisations of Toponogov's theorem to a version where the model spaces M_κ^2 are replaced by surfaces of revolution or surfaces with an S^1 - action, c.f. [E], [A]. U. Abresch pointed out to me that these generalisations can be handled with the same technique as used in the proof below.

Proof of Theorem 2.2. By remark 2 above we only have to prove A)ii). Note that in the case $\kappa > 0$ we have $\operatorname{diam}(M) \le \frac{\pi}{\sqrt{\kappa}}$ by Myers' theorem. For the case $\kappa > 0$ the proof is organized in three steps. In step 1 we consider the general case for $\kappa \le 0$, but we assume $\operatorname{diam}(M) < \frac{\pi}{\sqrt{\kappa}}$ and $|c| + |c_0| + |c_1| < \frac{2\pi}{\sqrt{\kappa}}$ for the case $\kappa > 0$. In step 2 the case $\kappa > 0$, $\operatorname{diam}(M) \le \frac{\pi}{\sqrt{\kappa}}$ and $|c| + |c_0| + |c_1| \le \frac{2\pi}{\sqrt{\kappa}}$ is reduced to step 1 by a simple limit argument. Finally, in step 3 we show that in the case $\kappa > 0$ there are no triangles with circumferece $|c| + |c_0| + |c_1| > \frac{2\pi}{\sqrt{\kappa}}$.
Step 1. For the case $\kappa > 0$ we assume $\operatorname{diam}(M) < \frac{\pi}{\sqrt{\kappa}}$ and also that the circumference of the triangle satisfies $|c| + |c_0| + |c_1| < \frac{2\pi}{\sqrt{\kappa}}$, so that the comparison triangles in M_κ^2 exists. From the triangle inequality $|c| \le |c_0| + |c_1|$ we get $|c| < \frac{\pi}{\sqrt{\kappa}}$ for $\kappa > 0$. Therefore we can choose $\varepsilon > 0$ such that $\operatorname{diam} M < \frac{\pi}{\sqrt{\kappa}} - 2\varepsilon$ and $|c| < \frac{\pi}{\sqrt{\kappa}} - 2\varepsilon$. We first look at a simple case: Suppose $q \in c(]0, |c|[)$. Then $|c| \ge |c_0| + |c_1|$ since c_1 and c_0 are minimal. By the triangle inequality we must have $|c| = |c_0| + |c_1|$. Therefore q divides c into two minimal pieces of length $|c_0|$ and $|c_1|$. Consequently equality holds in ii) since the geodesics from q to $c(t)$ are parts of c. If $q \notin c(]0, |c|[)$ we proceed as follows:

We consider the distance functions r from q in M, $\tilde{r}$ from $\tilde{q}$ in M_κ^2 and define

$$\begin{aligned} h(t) &= \operatorname{md}_\kappa \circ r \circ c(t) \\ \tilde{h}(t) &= \operatorname{md}_\kappa \circ \tilde{r} \circ \tilde{c}(t) \end{aligned}$$

$$\lambda(t) \;=\; h(t) - \tilde{h}(t)\,.$$

The idea is, to show that λ cannot have a negative minimum by the use of the Hessian estimate (48) in section 1.6. Unfortunately h is not differentiable in general since r is not differentiable beyond the cutlocus of q. This problem is resolved by a local approximation with a "superdistance function". The argument is slightly different in the cases $\kappa < 0$, $\kappa = 0$ and $\kappa > 0$.

In the case $\kappa = 0$, if λ has a negative minimum -2μ in $]0,|c|[$ also the function $\bar{\lambda}$ defined by

$$\bar{\lambda}(t) = \lambda(t) + \mu\frac{t(|c|-t)}{|c|^2}$$

has a negative minimum $< -\mu$ in $]0,|c|[$.

In the case $\kappa > 0$ we have $|c| \leq \frac{\pi}{\sqrt{\kappa}} - 2\varepsilon$ and define $\sigma_\varepsilon(t) = \mathrm{sn}_\kappa(t+\varepsilon) - \mathrm{sn}_\kappa(\frac{\varepsilon}{2})$ on $[0,|c|]$. If λ has a negative minimum then

$$\hat{\lambda} = \frac{\lambda}{\sigma_\varepsilon}$$

has a negative minimum.

For the point $t_0 \in]0,|c|[$ where λ or $\hat{\lambda}$ or $\bar{\lambda}$ has a negative minimum, we approximate r by local differentiable functions in a neighborhood of $c(t_0)$. Let γ be a normal geodesic from q to $c(t_0)$. For small values $\eta > 0$ we define in some neighborhood U of $\gamma(]\eta,|\gamma|[)$ the local superdistance functions

$$r_\eta(x) = \eta + \mathrm{dist}_U(\gamma(\eta),x) \geq r(x) = \mathrm{dist}(q,x)\,.$$

r_η is differentiable if U is sufficiently small. Therefore the function

$$h_\eta = \mathrm{md}_\kappa \circ r_\eta \circ c \tag{52}$$

is differentiable in some interval around t_0 and

$$h_\eta(t_0) = h(t_0),\;\; h_\eta \geq h\,. \tag{53}$$

Using the estimate (48) for the Hessian we have

$$\begin{aligned} h_\eta'' \;&=\; \langle \mathrm{Hess}(\mathrm{md}_\kappa \circ r_\eta)|_c\, \dot{c}, \dot{c}\rangle \\ &\leq\; \mathrm{cs}_\kappa \circ r_\eta \circ c + \frac{\mathrm{sn}_\kappa(\eta)}{\mathrm{sn}_\kappa(r_\eta \circ c - \eta)} \end{aligned}$$

for η small. The quantity $r_\eta \circ c(t) - \eta$ is bounded away from zero independent of η and $r_\eta \circ c(t) - \eta = \operatorname{dist}(\gamma(\eta), c(t)) \le \frac{\pi}{\sqrt{\kappa}} - 2\varepsilon$ from the diameter assumption. Observing (39) we get

$$h_\eta'' + \kappa h_\eta \le 1 + \text{const} \cdot \operatorname{sn}_\kappa(\eta)$$

with a constant independent of η. Since $\tilde{h}'' + \kappa \tilde{h} = 1$ the difference $\lambda_\eta = h_\eta - \tilde{h}$ satisfies

$$\lambda_\eta'' + \kappa \lambda_\eta \le \text{const} \cdot \operatorname{sn}_\kappa(\eta)\,. \tag{54}$$

Furthermore

$$\lambda_\eta \ge \lambda,\ \ \lambda_\eta(t_0) = \lambda(t_0) \tag{55}$$

by (53).

Case 1. $\kappa < 0$
If λ has a negative minimum $-\mu$ at t_0, then λ_η also has a negative minimum $-\mu$ at t_0, but

$$\lambda_\eta''(t_0) \le -\kappa\lambda(t_0) + \text{const} \cdot \operatorname{sn}_\kappa(\eta) = \underbrace{\kappa\mu}_{<0} + \text{const} \cdot \operatorname{sn}_\kappa(\eta)\,.$$

For η sufficiently small this is a contradiction.

Case 2. $\kappa = 0$
At the point $t_0 \in]0, |c|[$ where $\bar{\lambda}$ has a negative minimum we consider λ_η and also $\bar{\lambda}_\eta$ defined by

$$\bar{\lambda}_\eta = \lambda_\eta + \mu \frac{t(|c| - t)}{|c|^2}\,.$$

Then $\bar{\lambda}_\eta \ge \bar{\lambda}$ and $\bar{\lambda}_\eta(t_0) = \bar{\lambda}(t_0)$ by (55). Therefore $\bar{\lambda}_\eta$ also has a local negative minimum at t_0. But

$$\bar{\lambda}_\eta'' \le -\frac{2\mu}{|c|^2} + \text{const} \cdot \operatorname{sn}_\kappa(\eta)\,,$$

which is a contradiction for small η.

Case 3. $\kappa > 0$
At the point t_0 where $\hat{\lambda} = \frac{\lambda}{\sigma_\varepsilon}$ has a negative minimum $-\mu_o$ we also look at $\hat{\lambda}_\eta = \frac{\lambda_\eta}{\sigma_\varepsilon}$. Again $\hat{\lambda}_\eta \ge \hat{\lambda}$ and $\hat{\lambda}(t_0) = \hat{\lambda}_\eta(t_0)$ so that $\hat{\lambda}_\eta$ has a negative minimum $-\mu_0$ at t_0. Differentiate at t_0 to obtain

$$0 = \hat{\lambda}_\eta'(t_0) = \frac{\lambda_\eta'\sigma_\varepsilon - \lambda_\eta\sigma_\varepsilon'}{\sigma_\varepsilon^2}\Big|_{t_0}$$

and

$$\begin{aligned}
\hat{\lambda}_\eta''(t_0) &= \frac{1}{\sigma_\varepsilon^2}(\sigma_\varepsilon \lambda_\eta'' - \sigma_\varepsilon'' \lambda_\eta)|_{t_0} \\
&= \frac{1}{\sigma_\varepsilon^2}((\lambda_\eta'' + \kappa\lambda_\eta)\sigma_\varepsilon + \kappa\lambda_\eta \mathrm{sn}_\kappa(\frac{\varepsilon}{2}))|_{t_0} \\
&\leq \frac{1}{\sigma_\varepsilon(t_0)}\mathrm{const}\cdot \mathrm{sn}_\kappa(\eta) - \frac{\kappa\mu_0}{\sigma_\varepsilon(t_0)}\mathrm{sn}_\kappa(\frac{\varepsilon}{2}) < 0
\end{aligned}$$

for η sufficiently small, a contradiction.

Step 2. Assume now $\kappa > 0$, $\mathrm{diam}(M) \leq \frac{\pi}{\sqrt{\kappa}}$ and $|c| + |c_0| + |c_1| \leq \frac{2\pi}{\sqrt{\kappa}}$. We choose a sequence κ_i, $0 < \kappa_i < \kappa$ and $\lim_{i\to\infty}\kappa_i = \kappa$. Then $\mathrm{diam}(M) < \frac{\pi}{\sqrt{\kappa_i}}$ and $|c|+|c_0|+|c_1| < \frac{2\pi}{\sqrt{\kappa_i}}$. By step 1 the theorem holds for the sphere $S^2_{\kappa_i} \subset \mathbb{R}^3$ as the comparison space. By compactness, the sequence of comparison triangles $\tilde{\Delta}_i = (\tilde{c}^i, \tilde{c}^i_0, \tilde{c}^i_1)$ has a subsequence converging to a comparison triangle $\tilde{\Delta}$ in S^2_κ. By continuity of the family of distance functions on the family of spheres $S^2_{\tilde{\kappa}} \subset \mathbb{R}^3$, $\tilde{\kappa} > 0$, statement A)ii) now follows for the limit triangle $\tilde{\Delta}$.

Step 3. Suppose $\kappa > 0$ and $|c| + |c_0| + |c_1| > \frac{2\pi}{\sqrt{\kappa}}$. We can choose $\delta > 0$ such that $|c| + |c_0| + |c_1| = \frac{2\pi}{\sqrt{\delta}}$ Then for the comparison triangle in M^2_δ the geodesics $\tilde{c}_0$, $\tilde{c}_1$, $\tilde{c}$ have length $< \frac{\pi}{\sqrt{\delta}}$ and therefore form a great circle. The antipodal point $\bar{q}$ of $\tilde{q}$ has to be a point of $\tilde{c}$, say $\bar{q} = \tilde{c}(t_0)$. By step 1 we have $\frac{\pi}{\sqrt{\delta}} = \mathrm{dist}(\tilde{q}, \tilde{c}(t_0)) \leq \mathrm{dist}(q, c(t_0))$ contradicting $\mathrm{dist}(q, c(t_0)) \leq \frac{\pi}{\sqrt{\kappa}} < \frac{\pi}{\sqrt{\delta}}$. This completes the proof. □

3 Applications of Toponogov's Theorem

3.1 An estimate for the number of generators for the fundametal group

As a first application of Toponogov's theorem we present Gromov's theorem concerning the number of generators for the fundamental group $\pi_1(M)$. Since any element of the fundamental group $\pi_1(M)$ with base point p of a Riemannian manifold M can be represented by a geodesic loop of minimal length at the point p, it is clear that the geometry of M should have strong influence on the structure of $\pi_1(M)$. The earliest result in this direction is Myers' theorem, cf.[CE], [GKM]: the universal cover of a compact Riemannian manifold with strictly positive Ricci curvature is compact and the fundamental group finite. If the sectional curvature K of a compact even dimensional manifold is strictly positive, then by the Synge Lemma, cf. [CE], [GKM], $\pi_1(M) = 1$ or Z_2 depending on the orientability of M. If M is complete non-compact and $K > 0$, then $\pi_1(M) = 1$ since M is diffeomorphic to $\mathbb{R}^n$, cf. [GM]. Finally if M is complete non-compact and $K \geq 0$, then by the soul theorem of Cheeger and Gromoll [CG1], $\pi_1(M)$ contains a lattice group of finite index.

Theorem 3.1 (Gromov)

(i) Suppose the sectional curvature of M^n is nonnegative. Then $\pi_1(M^n)$ can be generated by $N \leq \sqrt{2n\pi}\, 2^{n-2}$ elements.

(ii) If the sectional curvature K of M^n is bounded from below, $K \geq -\lambda^2$ and the diameter of M^n is bounded, $diam M^n \leq D$, then $\pi_1(M^n)$ can be generated by $N \leq \frac{1}{2}\sqrt{2n\pi}(2 + 2\cosh(2\lambda D))^{\frac{n-1}{2}}$ elements.

Proof. Let $G = \pi_1(M, p_0)$ be the fundamental group with base point $p_0 \in M$. $\tilde{M}$ denotes the Riemannian universal cover of M. The group of covering transformations G acts on $\tilde{M}$ by isometries. We choose a point $x_0 \in \tilde{M}$ which covers p_0 and define for $\gamma \in G$ the displacement

$$|\gamma| := \mathrm{dist}(x_0, \gamma x_0)$$

A minimal geodesic c from x_0 to γx_0 projects in M to a loop of minimal length $|\gamma|$ in the homotopy class representing γ. There are only finitely many elements of G satisfying $|\gamma| \leq r$. (An infinite sequence $\gamma_i x_0$ of points would have a limit point in the compact ball of radius r around zero contradicting the covering property.) Therefore we can choose an element $\gamma_1 \in G$ with the property $|\gamma_1| = \min\{|\gamma| \mid \gamma \in G\}$. Inductively we can construct generators $\gamma_1, \gamma_2, \ldots$ of G satisfying $|\gamma_1| \le |\gamma_2| \le \ldots$ as follows: Suppose $\gamma_1, \ldots, \gamma_k$ are constructed already and the subgroup $< \gamma_1, \ldots, \gamma_k >$ generated by $\gamma_1, \ldots, \gamma_k$ is not equal to G. Then we can choose $\gamma_{k+1} \in G$ so that $|\gamma_{k+1}| = \min\{|\gamma| \mid \gamma \in G \setminus < \gamma_1, \ldots, \gamma_k >\}$. For $i < j$ we have $|\gamma_i| \le |\gamma_j|$ and

$$\ell_{ij} := \operatorname{dist}(\gamma_i x_0, \gamma_j x_0) \ge |\gamma_j| \, .$$

To prove the last inequality, suppose $\ell_{ij} < |\gamma_j|$. Then $\gamma_j' := \gamma_i^{-1}\gamma_j$ has displacement $|\gamma_j'| = \ell_{ij} < |\gamma_j|$ and $< \gamma_1, \ldots, \gamma_j > = < \gamma_1, \ldots, \gamma_{j-1}, \gamma_j' >$, contradicting the choice of γ_j.

For each γ_i we choose a minimal geodesic c_i from x_0 to $\gamma_i x_0$ of length $\ell_i = |\gamma_i|$. For $i < j$ we choose a minimal geodesic from $\gamma_i x_0$ to $\gamma_j x_0$ of length ℓ_{ij}. By Toponogov's theorem the angle $\alpha_{ij} = \sphericalangle(\dot c_i(0), \dot c_j(0))$ is bounded below by the angle $\tilde\alpha$ of a comparison triangle in M_κ^2 where $\kappa = 0$ for (i) and $\kappa = -\lambda^2$ for (ii). By the law of cosines (42), (44) in M_κ^2

$$\cos\tilde\alpha = \frac{\ell_i^2 + \ell_j^2 - \ell_{ij}^2}{2\ell_i\ell_j} \qquad \text{for } \kappa = 0 \tag{56}$$

$$\cos\tilde\alpha = \frac{\cosh(\lambda\ell_i)\cosh(\lambda\ell_j) - \cosh(\lambda\ell_{ij})}{\sinh(\lambda\ell_i)\sinh(\lambda\ell_j)} \qquad \text{for } \kappa = -\lambda^2 \, . \tag{57}$$

The right hand side of (57) is increasing in the variable ℓ_i (to see this differentiate). The relation $\ell_i \le \ell_j \le \ell_{ij}$ now leads to the estimates

$$\cos\tilde\alpha \le \frac{\ell_i^2 + \ell_j^2 - \ell_j^2}{2\ell_i^2} = \frac{1}{2} \qquad \text{for } \kappa = 0 \tag{58}$$

$$\begin{aligned} \cos\tilde\alpha &\le \frac{\cosh^2(\lambda\ell_j) - \cosh(\lambda\ell_j)}{\sinh^2(\lambda\ell_j)} = \frac{\cosh(\lambda\ell_j)}{\cosh(\lambda\ell_j)+1} \\ &\le \frac{\cosh(2\lambda D)}{\cosh(2\lambda D)+1} \qquad \text{for } \kappa = -\lambda^2 \; . \end{aligned} \tag{59}$$

For the last inequality observe that $\ell_i \le 2D$ by the construction of the generators γ_i of G. To see this, observe that for $\varepsilon > 0$ any loop at p_0 in M is homotopic to a composition of loops with length $\le 2D + \varepsilon$: Subdivide the original loop into segments of length $\le \varepsilon$ and then insert minimal connections from the subdivision points to p_0 and their inverses. Since in the construction $|\gamma_{k+1}|$ is chosen to be minimal in $G \setminus < \gamma_1, \ldots, \gamma_k >$, it follows $|\gamma_{k+1}| \le 2D + \varepsilon$, but ε was arbitrary. Let

$$\alpha_\kappa = \begin{cases} \frac{\pi}{3} & \text{for } \kappa = 0 \\ \arccos\left(\frac{\cosh(2\lambda D)}{1+\cosh(2\lambda D)}\right) & \text{for } \kappa = -\lambda^2 \end{cases} \tag{60}$$

then $\alpha_{ij} \geq \tilde{\alpha} \geq \alpha_\kappa$. To complete the argument consider the initial vectors $v_i = \dot{c}_i(0) \in T_{x_0}\tilde{M}$. We have $\measuredangle(v_i, v_j) \geq \alpha_\kappa > 0$. In $T_{x_0}\tilde{M}$ there can be only a finite number of distinct unit vectors with this property. A rough explicit estimate for the maximal number is obtained as follows: The intrinsic balls of radius $\alpha_\kappa/2$ around the points v_i in the unit sphere S^{n-1} in $T_{x_0}\tilde{M}$ are disjoint. Therefore the maximal number N_κ of points v_i is estimated by the volume of S^{n-1} divided by the volume of a ball of radius $\alpha_\kappa/2$ in S^{n-1}. The volume of this spherical ball is estimated below by the volume of a euclidian (n-1) ball of radius $\sin(\alpha_\kappa/2)$. This estimate, however, can be improved by a factor $\frac{1}{2}$ by the following simple observation: The generators satisfy $|\gamma_i| = |\gamma_i^{-1}|$. Therefore we also have $\measuredangle(v_i, -v_j) \geq \alpha_\kappa$. Hence the volume of the sphere can be replaced by the volume of the real projective space and we obtain

$$N_\kappa \leq \frac{\frac{1}{2}\mathrm{vol}\, S^{n-1}}{\mathrm{vol}\, B^{n-1}(\sin(\alpha_\kappa/2))} = \frac{\sqrt{\pi}\,\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sin^{n-1}(\alpha_\kappa/2)} = \sqrt{\pi}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\left(\frac{2}{1-\cos\alpha_\kappa}\right)^{\frac{n-1}{2}}$$

The logarithmic convexity of Γ can be used to find $\sqrt{\frac{n-1}{2}} \leq \Gamma(\frac{n+1}{2})/\Gamma(\frac{n}{2}) \leq \sqrt{\frac{n}{2}}$. Inserting the appropriate value for α_κ from (60) into the estimate for N_κ finishes the proof. □

The estimates given in the Theorem are never sharp as can be seen by looking at surfaces.

3.2 Critical points of distance functions

Distance functions on a Riemannian manifold M are not differentiable in general. Despite this fact it is possible to develop a critical point theory similar to the Morse theory of a differentiable function. The idea was introduced by Grove and Shiohama [GS] for the proof of their diameter sphere theorem, cf. theorem 3.14. Subsequently it has been refined by Gromov [G2] in connection with his finiteness result for the sum of the Betti numbers, cf. theorem 3.19. These applications deal with distance functions from a point in M. Recently Grove and Petersen have generalized the concept for distance functions from closed subsets of a manifold, in particular from the diagonal Δ in $M \times M$. This leads to an interesting finiteness result concerning the number of homotopy types of Riemannian manifolds, cf. [GP].

Definition 3.2 *Let A be a closed subset of M. Consider the distance function $dist_A$ from A defined by $dist_A(q) = dist(A, q)$. A point $q \in M$ is said to be a* critical point *for*

$dist_A$, if for any vector $v \in T_qM$ there is a distance minimizing geodesic c from q to A satisfying

$$\langle v, \dot{c}(0) \rangle \geq 0 \ . \tag{61}$$

A non-critical point is called a regular *point.*

For points $q \notin A$ this condition is equivalent to $\sphericalangle(v, \dot{c}(0)) \leq \frac{\pi}{2}$. Instead of referring to a critical or regular point for dist_A we also shall say that q is critical or regular for A. Notice that any point $q \in A$ is critical for A.

In the following examples let A={p}.

Examples

i) Consider the flat cylinder $S^1 \times \mathbb{R} \subset \mathbb{R}^3$ and $p = (x_0, y_0, z_0)$, $x_0^2 + y_0^2 = 1$. Then p and $q = (-x_0, -y_0, z_0)$ are the only critical points for p.

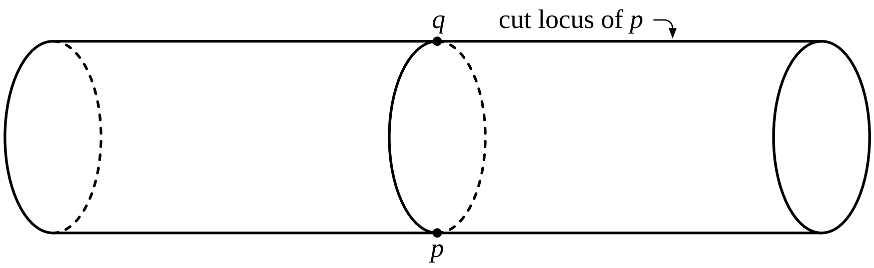

Figure 4: cylinder

ii) Consider the flat torus $T = \mathbb{R}^2 / Z \oplus Z$ and $p = (\frac{1}{2}, \frac{1}{2})$. Then the only critical points for p are p, $q_1 = (1, \frac{1}{2})$, $q_2 = (1, 1)$, $q_3 = (\frac{1}{2}, 1)$.

According to the definition a point q is regular for A if the initial vectors for all minimal geodesic from q to A are contained in an open half space of T_qM, i.e. there is a vector $v \in T_qM$ such that

$$\langle \dot{c}(0), v \rangle < 0$$

for any minimal geodesic c from q to A. Of course equivalently we have a vector $w = -v$ such that

$$\langle \dot{c}(0), w \rangle > 0$$

for any minimal geodesic c from q to A.

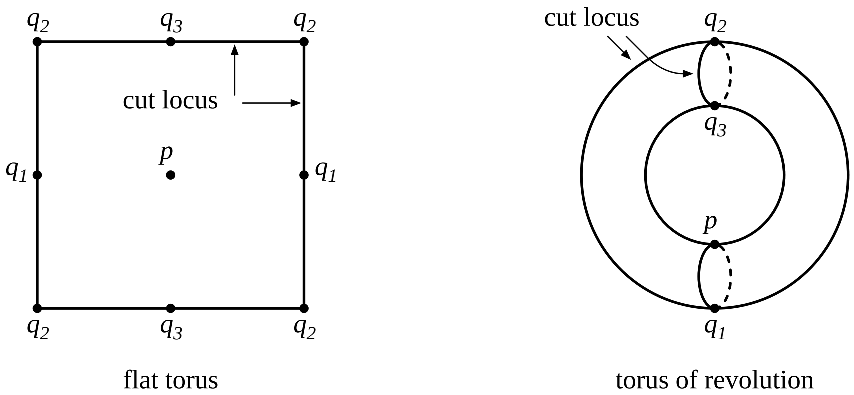

Figure 5: critical points and cut locus for a point p in tori

Lemma 3.3 (local existence of gradient-like vector fields) *Let M be complete and A a closed subset of M. Then for any regular point q of $dist_A$ there is a unit vectorfield X on some open neighborhood U of q such that*

$$\langle X_{\tilde{q}}, \dot{c}(0)\rangle < 0 \tag{62}$$

for any $\tilde{q} \in U$ and any minimal geodesic c from $\tilde{q}$ to A.

Definition 3.4 *A unit vector field X on U satisfying (62) is called a* gradient-like *vector field for $dist_A$.*

Proof. Since q is a regular point we can choose a unit vector $X_q \in T_qM$ with $\langle X_q, \dot{c}(0)\rangle < 0$ for any minimal geodesic c from q to A. Extend X_q to an arbitrary smooth vector field on some open neighborhood of q. Then X satisfies condition (62) on a sufficiently small ball U around q. Otherwise there would be a sequence of points q_i converging to q and minimizing geodesics c_i from q_i to A satisfying $\langle X_{q_i}, \dot{c}_i(0)\rangle \geq 0$. A limiting geodesic c of c_i would be a minimal geodesic from q to A with $\langle X_q, \dot{c}(0)\rangle \geq 0$, contradicting the choice of X_q. □

Corollary 3.5 (existence of global gradient-like vector fields) *As above, let M be complete and A a closed subset of M. Then*

a) The set of regular points of $dist_A$ is open.

b) On the open set U of regular points there exists a gradient-like vector field for $dist_A$.

Proof. a) is obvious from (62). For the proof of b) we point out that local vector fields of the lemma can be glued together by means of a partition of unity to obtain a vector field $\tilde{X}$ on U satisfying (62). This is a consequence of the following observation: If $v_1, \ldots, v_m$ are unit vectors in a euclidian vector space satisfying $\langle v_i, w\rangle < 0$, then any convex linear combination $v = \sum_{i=1}^m \lambda_i v_i$, $\lambda_i \geq 0$, $\sum_{i=1}^m \lambda_i = 1$ satisfies $\langle v, w\rangle < 0$. Now we can take $X = \tilde{X}/\|\tilde{X}\|$. □

The following lemma contains an important monotonicity property for gradient-like vector fields.

Lemma 3.6 *Let M be complete, A a closed subset, U an open subset of M and X a gradient-like vector field for $dist_A$ on U, Φ the flow of $-X$ and Ψ the flow of X. Then*

a) $dist_A$ is strictly decreasing along any integral curve of $-X$.

b) On any compact subset C of U the decreasing rate is controlled by a Lipschitz constant: There is a constant $\Theta > 0$ such that

$$dist_A\Phi(q, t_0 + \tau) \leq dist_A\Phi(q, t_0) - \tau\Theta \tag{63}$$

as long as $\Phi(q, t_0 + \sigma) \in C$ for $0 \leq \sigma \leq \tau$. Equivalently we have

$$dist_A\Psi(q, t_0' + \tau) \geq dist_A\Psi(q, t_0') + \tau\Theta \tag{64}$$

as long as $\Psi(q, t_0' + \sigma) \in C$ for $0 \leq \sigma \leq \tau$.

Proof. It suffices to prove b). First notice that X satisfies the inequality $\langle -X_q, \dot{c}(0)\rangle \geq \Theta$ for some $\Theta > 0$, any $q \in C$ and any minimal geodesic c from q to A. Otherwise there would be sequences $q_i \in C$ and minimal geodesics c_i from q_i to A with $\lim_{i\to\infty} \langle X_{q_i}, \dot{c}_i(0)\rangle \geq 0$. By compactness there would be a limit point $q \in C$ and a minimal limiting geodesic from q to A with $\langle X_q, \dot{c}(0)\rangle \geq 0$, contradicting (62). Consider now the function $h(t) = \mathrm{dist}_A\Phi(q, t)$. We construct an upper support function $\tilde{h}$ for h as follows: Let $p \in A$ with $\mathrm{dist}(p, \Phi(q, t_0)) = \mathrm{dist}_A\Phi(q, t_0)$ and choose a minimal geodesic $c : [0, 1] \to M$ from $\Phi(q, t_0)$ to p . For a fixed η, $0 < \eta < 1$ let $\tilde{h}(t) = \eta + \mathrm{dist}(c(|c| - \eta), \Phi(q, t))$. $\tilde{h}$ is differentiable in a neighborhood of t_0 and satisfies $\tilde{h}(t) \geq \mathrm{dist}(p, \Phi(q, t)) \geq \mathrm{dist}_A\Phi(q, t) = h(t)$ and $\tilde{h}(t_0) = h(t_0)$. The derivative at t_0 is given by $\tilde{h}'(t_0) = \left\langle \mathrm{grad\ dist}_{c(|c|-\eta)}|_{\Phi(q,t_0)}, \Phi_* \frac{\partial}{\partial t}|_{q,t_0}\right\rangle = \langle -\dot{c}(0), -X \circ \Phi(q, t_0)\rangle$, hence $\tilde{h}'(t_0) \leq -\Theta$. Such a support function exists at any t_0, therefore condition (63) follows easily. $\square$

As an immediate consequence we have:

Corollary 3.7 *Any local maximum point q of $dist_A$ is a critical point for A.*

Corollary 3.8 *Let M^n be complete and $B = B(p,r)$ a ball of radius r around the point $p \in M$. Suppose there are no critical points of $dist_p$ in ∂B.*
Then ∂B is a topological (n-1)-submanifold of M.

Proof. We only have to show that ∂B is locally euclidian. Consider a vector field X on the set of regular points with property (62) and the flow Φ of $-X$. For a given point $q \in \partial B$ let Q be a local (n-1)-dim submanifold through q which is transversal to X, for example the image under the exponential map of a neighborhood V of the origin in the (n-1)-plane orthogonal to X_q in T_qM. By the inverse mapping theorem we can assume that V and $\varepsilon > 0$ are chosen such that $\Phi|_{Q \times [-\varepsilon,\varepsilon]}$ is a local diffeomorphism. Since $t \mapsto \mathrm{dist}_p\Phi(q,t)$ is strictly decreasing, we have $\mathrm{dist}_p\Phi(q,\varepsilon) < \mathrm{dist}_p\Phi(q,0) < \mathrm{dist}_p\Phi(q,-\varepsilon)$. Therefore by continuity we can assume that $\mathrm{dist}_p\Phi(\tilde{q},\varepsilon) < \mathrm{dist}_p\Phi(q,0) < \mathrm{dist}_p\Phi(\tilde{q},-\varepsilon)$ for $\tilde{q} \in Q$, after shrinking Q if necessary. Now any integral curve of X through a point $\tilde{q}$ of Q meets exactly one point of $\partial B = \mathrm{dist}_p^{-1}(r)$ by the monotonicity property. The map from Q to ∂B defined by the projection along the integral curves is a homeomorphism onto its image. $\square$

Corollary 3.9 *Let M be a complete non-compact manifold and suppose that for some point $p \in M$ all the critical points of $dist_p$ are contained in a ball $B = B(p,r)$*
Then M is homeomorhpic to the interior of a compact manifold with boundary.

Proof. Let X be a gradient-like vector field on the set of regular points with flow Ψ. Then $F : \partial B \times [0,\infty[\to M$ defined by $F(q,t) := \Psi(q,t)$ maps $\partial B \times [0,\infty[$ homeomorphic onto $M \setminus B$. For this the properties (64) and $\|X\| = 1 < \infty$ are important. Hence M is homeomorphic to $B \cup F(\partial B \times [0,\infty[) \approx B \cup F(\partial B \times [0,1[)$ with boundary $F(\partial B \times \{1\}) \approx \partial B$. $\square$

Corollary 3.10 *Suppose that there is no critical point of $dist_p$ in $\overline{B(p,r)} \setminus \{p\}$.*
Then $\overline{B(p,r)}$ is contractible.

Proof. An easy exercise. $\square$

Definition 3.11 *An* isotopy *of M (in the topological category) is a homotopy $G : M \times [0,1] \to M$ such that $p \mapsto G(p,\tau)$ is a homeomorphism from M onto a subset of M for any $\tau \in [0,1]$ and $p \mapsto G(p,0)$ is the identity map of M.*
If B_1 and B_2 are subsets of M, we say that the isotopy G moves B_2 into B_1, provided $G(B_2 \times \{1\}) \subset B_1$.

Corollary 3.12 (Isotopy Lemma) *Given a complete manifold M, a point $p \in M$, $0 < r_1 < r_2 \leq \infty$ and an open neighborhood U of the annulus $A = \overline{B(p,r_2) \setminus B(p,r_1)}$. Assume that there are no critical points of $dist_p$ in A.*
Then there is an isotopy of M which is the identity on $M \setminus U$ and which moves $B(p,r_2)$ into $B(p,r_1)$.

Proof. Using a partition of unity one can construct a vector field X on M which is gradient-like on some neighborhood W of A with $\overline{W} \subset U$ and $X|_{M\setminus U} = 0$.
If $r_2 < \infty$ we can choose $\Theta > 0$ such that (62) holds on the compact set A. Then for $t_0 > \frac{1}{\Theta}(r_2 - r_1)$ the isotopy G defined by $G(q,\tau) = \Phi(q, \tau \cdot t_0)$ moves $B(p,r_2)$ into $B(p,r_1)$, where again Φ is the flow of $-X$.
If $r_2 = \infty$, we consider $F : \partial B \times]-\infty, \infty[\to M$, $F(q,t) = \Phi(q,t)$ and use on the domain of this homeomorphism onto a subset of M an isotopy induced from a deformation of $]-\infty,\infty[$ into $]0,\infty[$, for instance $G(t,\tau) = \ln(\tau + \mathrm{e}^t)$. $\square$

The elementary corollaries above demonstrate that gradient-like vector fields can be used for deformations in the same way as gradient vector fields in standard Morse theory. However all these deformation arguments are useless unless one can get additional information on the set of critical points. In standard Morse theory the Morse Lemma is an important tool for this purpose. Unfortunately, there is no analogue of the Morse Lemma available. In fact one cannot say much about the change of topology of $B(p,r) = \mathrm{dist}_p^{-1}(r)$ when r passes a critical level.
In the presence of a lower curvature bound, however, Toponogov's comparison theorem can be used to obtain additional information about critical points leading to rather strong conclusions. In contrast to standard Morse theory the information obtained on the set of critical points is more of a global nature. For the proof of 3.19 one has to consider not just a single distance function but all the distance functions from the various points of M.

3.3 The diameter sphere theorem

One of the famous results in Riemannian geometry is the $\frac{1}{4}$- pinching sphere theorem cf. [GKM], [CE], which can be stated as follows:

Theorem 3.13 (Rauch, Berger, Klingenberg) *Suppose M^n is complete, simply connected and the sectional curvature K satisfies*

$$\frac{1}{4} < K \leq 1 \ .$$

Then M is homeomorphic to the standard sphere.

One of the essential steps in the proof of this theorem is to show that the injectivity radius of the exponential map and hence the diameter of M is $\geq \pi > \frac{\pi}{2\sqrt{\delta}}$, where $\delta > \frac{1}{4}$ is the minimum of the sectional curvatures on M. Grove and Shiohama have generalized the $\frac{1}{4}$-pinching sphere theorem to the diameter sphere theorem below by replacing the upper curvature bound by this lower bound for the diameter. The proof is a nice application of critical point theory and of Toponogov's theorem.

Theorem 3.14 (Grove-Shiohama) *Let M^n be a complete manifold with $K \geq \delta > 0$ and diam$M > \frac{\pi}{2\sqrt{\delta}}$. Then M is homeomorphic to S^n.*

Proof. After rescaling the metric we can assume $K \geq 1$ and $\text{diam}M > \frac{\pi}{2}$. Let p, q be two points of maximal distance in M, $\text{dist}(p,q) = \text{diam}M$. By corollary 3.7 q is critical for p. We show that q is uniquely determined by p. Suppose q_1, q_2 are two points satisfying $\text{dist}(p,q_i) = \text{diam}M$. Choose minimal geodesics c from q_1 to q_2 and c_i from q_i to p. Since q_1 is critical for p, c_1 can be chosen such that $\alpha_1 = \sphericalangle(\dot{c}_1(0), \dot{c}(0)) \leq \frac{\pi}{2}$. Then $\ell_1 := |c_1| = |c_2| = \text{diam}M > \frac{\pi}{2}$ and $\ell := |c| \leq \ell_1$. Consider the corresponding comparison triangle $\tilde{c}$, $\tilde{c}_1$ $\tilde{c}_2$ in the standard sphere with corresponding angle $\tilde{\alpha}_1$ and edge lengths $|\tilde{c}| = \ell$, $|\tilde{c}_1| = |\tilde{c}_2| = \ell_1$. Then by 2.2 $\tilde{\alpha}_1 \leq \alpha_1 \leq \frac{\pi}{2}$. By the law of cosines in S^2_1 we have

$$0 \leq \sin\ell_1 \sin\ell \cos\tilde{\alpha}_1 = \cos\ell_1 - \cos\ell_1 \cos\ell = (1 - \cos\ell)\cos\ell_1 \leq 0$$

and hence $\ell = 0$, i.e. $q_1 = q_2$.

Next we show that p and q are the only critical points for p, more precisely: Let $q_1 := q$, $q_2 \in M$, $q_1 \neq q_2 \neq p$ and $c : [0,1] \to M$ be a minimal geodesic of length ℓ from q_1 to q_2. Then for any minimal geodesic c_2 from q_2 to p we have $\langle \dot{c}(1), \dot{c}_2(0)\rangle > 0$, i.e. the vector $v := -\dot{c}(1) \in T_{q_2}M$ can be used to define the open half space for the regularity of q_2. To show this, choose a minimal geodesic c_1 from q_1 to p and let $\ell_1 = |c_1|$, $\ell_2 = |c_2|$. By the uniqueness of $q = q_1$ we have $0 < \ell < \ell_1$ and $0 < \ell_2 < \ell_1$. For the geodesic triangle c, c_1, c_2 with angle $\alpha_2 = \measuredangle(\dot{c}_2(0), -\dot{c}(1))$ we consider the corresponding comparison triangle in S_1^2 with corresponding angle $\tilde{\alpha}_2$. Then $\langle \dot{c}_2(0), -\dot{c}(1)\rangle = \cos\alpha_2 \leq \cos\tilde{\alpha}_2$ and by the law of cosines

$$\cos\tilde{\alpha}_2 = \frac{cos\ell_1 - \cos\ell\cos\ell_2}{\sin\ell\,\sin\ell_2} < 0$$

since $\ell_1 > \frac{\pi}{2}$.
Now let $\varepsilon > 0$ be sufficiently small such that $\exp|_{B(p,\varepsilon)}$ and also $\exp|_{B(q,\varepsilon)}$ are local diffeomorphisms. The vector field $X_1 = \operatorname{grad}(\operatorname{dist}_p|_{B(p,\varepsilon)\setminus\{p\}})$ satisfies condition (62) in section 3.2. The regularity argument above shows that $X_2 = -\operatorname{grad}(\operatorname{dist}_q|_{B(q,\varepsilon)\setminus\{q\}})$ satisfies this condition as well. Therefore one can construct a gradient-like vector field X on $M\setminus\{p,q\}$ which coincides with X_1 on $B(p,\frac{\varepsilon}{2})\setminus\{p\}$ and with X_2 on $B(q,\frac{\varepsilon}{2})\setminus\{q\}$ and $\|X\| = 1$. The flow Ψ of X satisfies (64) on all of $M\setminus(B(p,\frac{\varepsilon}{2})\cup B(q,\frac{\varepsilon}{2}))$. Hence all the integral curves of X have finite lengths and extend continuously to the end points p and q. For a unit vector $v\in T_pM$ the integral curve $\varphi_v(t) := \Psi(\exp(\frac{\varepsilon}{2}v), t-\frac{\varepsilon}{2})$ is defined on an interval $]0,\ell_v[$, where ℓ_v is the length of φ_v. Since X is differentiable, the function $v\mapsto \ell_v$ is differentiable. Let $F(t,v) = \varphi_v(t\cdot\ell_v)$, $F(0,v) = p$, $F(1,v) = q$ for $t\in]0,1[$ and $\|v\| = 1$. Then the map $G : tv \mapsto F(t,v)$ maps the closed unit ball $\overline{B}$ of T_pM onto M inducing a homeomorphism from the quotientspace $B/\partial B \approx S^n$ to M. □

With a slight modification of the reparametrisation of φ_v in the proof above the map G can be made smooth in the interior B of $\overline{B}$. However there is no information about the "twist" of the map near q.

The diameter sphere theorem may also be viewed as a diameter pinching theorem for manifolds of positive curvature: The quantity

$$\partial_M = \min(K)\frac{(\operatorname{diam}M)^2}{\pi^2}$$

is invariant under scalings of the metric. By Myers' theorem we have $\partial_M \leq 1$. According to the diameter sphere theorem M^n is homeomorphic to S^n if $\partial_M > \frac{1}{4}$. If $\partial_M = 1$, M^n is isometric to the sphere S^n. This rigidity result was originally obtained by Toponogov as an application of the triangle comparison theorem, cf. [CE], but it also follows from the more general theorem of Cheng [Cg], which has been discussed in the first part of this lecture series.

If one relaxes the assumption in the $\frac{1}{4}$-pinching theorem to $\frac{1}{4} \leq K \leq 1$, then there is the rigidity theorem of Berger, cf. [CE]. In view of this result one also should expect a rigidity theorem if one assumes $K \geq 1$ and diam$M = \frac{\pi}{2}$. In fact Gromoll and Grove cf. [GG] have obtained a corresponding result:
Under the given hypothesis either

a) M is homeomorphic to a sphere, or

b) M has the cohomology ring of the Cayley plane, or

c) M is isometric to one of the following spaces with their standard metrics: $\mathbb{C}\mathrm{P}^m$, $\mathbb{H}\mathrm{P}^\ell$, $\mathbb{C}\mathrm{P}^{2d-1}/\{[z_1, \ldots . z_{2d}] \sim [\overline{z}_{d+1}, \ldots, \overline{z}_{2d}, -\overline{z}_1, \ldots, -\overline{z}_d]\}$, S_1^n/Γ, where the orthogonal representation of $\Gamma = \pi_1(M)$ on $\mathbb{R}^{n+1}$ is reducible.

The proof is somewhat technical for our exposition.

3.4 A critical point lemma and a finiteness result

The critical point lemma below was one of the basic observations which lead Gromov to the finiteness theorem in the next section. Its proof is a simple application of Toponogov's theorem (used twice). The given estimate is somewhat stronger than in Gromov's original lemma. It was also used by Abresch [A].

Lemma 3.15 (critical point lemma) *Let M be complete and p, q_1, q_2 $\in M$, $q_i \neq p$ and assume q_1 is critical for p. Furthermore let c_i be minimal geodesics from p to q_i of length $\ell_i, \ell_1 \leq \ell_2$ and $\alpha = \sphericalangle(\dot{c}_1(0), \dot{c}_2(0))$.*

a) If the sectional curvature satisfies $K \geq 0$, then

$$\cos\alpha \leq \frac{\ell_1}{\ell_2} \ .$$

b) If $K \geq -\lambda^2$ $(\lambda > 0)$ and diam$M < D$, then

$$\cos\alpha \leq \frac{\ell_1}{\ell_2}\lambda D \coth(\lambda D) \ .$$

Proof. Let c be a minimal geodesic from q_1 to q_2 of length ℓ. Since q_1 is critical for p there is a minimal geodesic $\overline{c}_1$ from q_1 to p of length ℓ_1 such that $\alpha_1 = \sphericalangle(\dot{\overline{c}}_1(0), \dot{c}(0)) \leq \frac{\pi}{2}$. Using Toponovs's theorem 2.2 , part B for this hinge and the law of cosines we get

$$\begin{array}{rcll} \ell_2^2 & \leq & \ell_1^2 + \ell^2 & \text{for } K \geq 0 \\ \cosh\lambda\ell_2 & \leq & \cosh\lambda\ell_1 \, \cosh\lambda\ell & \text{for } K \geq -\lambda^2 \end{array}$$

Consider now the geodesic triangle c_1, c_2, c and the corresponding triangle $\tilde{c}_1$, $\tilde{c}_2$, $\tilde{c}$ with the same edge length in the comparison space $\mathbb{R}^2$ respectively $M^2_{-\lambda^2}$. Then the angle comparison theorem 2.2 A (i) leads to $\tilde{\alpha} = \sphericalangle(\dot{\tilde{c}}_1, \dot{\tilde{c}}_2) \le \alpha$ or eqivalently $\cos\alpha \le \cos\tilde{\alpha}$. Applying again the law of cosines we can finish the argument:

$$\cos\tilde{\alpha} = \frac{\ell_1^2 + \ell_2^2 - \ell^2}{2\ell_1\ell_2} \le \frac{\ell_1}{\ell_2} \qquad \text{for} \quad K \ge 0$$

and

$$\begin{aligned}\cos\tilde{\alpha} &= \frac{\cosh\lambda\ell_1\,\cosh\lambda\ell_2 - \cosh\lambda\ell}{\sinh\lambda\ell_1\,\sinh\lambda\ell_2} \le \frac{\cosh\lambda\ell_1\,\cosh\lambda\ell_2 - \frac{\cosh\lambda\ell_2}{\cosh\lambda\ell_1}}{\sinh\lambda\ell_1\,\sinh\lambda\ell_2} \\ &= \tanh\lambda\ell_1\coth\lambda\ell_2 \le \frac{\ell_1}{\ell_2}\lambda\ell_2\coth\lambda\ell_2 \le \frac{\ell_1}{\ell_2}\lambda D\coth\lambda D\end{aligned}$$

for $K \ge -\lambda^2$. $\square$

Corollary 3.16

a) *Given a complete manifold M^n with $K \ge 0$ and a constant $L > 1$. Then there are only finitely many critical points $q_1, \dots, q_k$ for the distance function $dist_p$ satisfying*

$$dist_p(q_{i+1}) \ge L \cdot dist_p(q_i)\,.$$

If $L \ge 3(1+\sqrt{2})^{n-1}$, then $k \le 2n$.

b) *For manifolds with $K \ge -\lambda^2$ and $diamM < D$ the same statement holds for $L \ge 3(1+\sqrt{2})^{n-1}\lambda D\coth\lambda D$.*

Remark

By reversing the indexing of the points q_i we also have at most $2n$ critical points satisfying

$$\mathrm{dist}_p(q_{i+1}) \le \frac{1}{L}\mathrm{dist}_p(q_i)$$

if L is chosen as specified in the corollary.

Proof of corollary 3.16. We consider the case $K \ge 0$ and leave the simple modification for b) to the reader. Connect p and q_i by minimal geodesics c_i of lengths ℓ_i. Then $\ell_i \ge L\ell_j$ for $i > j$. By the critical point lemma the angles $\alpha_{ij} = \sphericalangle(\dot{c}_i(0), \dot{c}_j(0))$ satisfy $\cos\alpha_{ij} \le \frac{\ell_j}{\ell_i} \le \frac{1}{L}$ or equivalently $\alpha_{ij} \ge \arccos\frac{1}{L} > 0$. There are only finitely many vectors in T_pM with this condition, compare also the proof of theorem 3.1. If $L = 2$ we have $\alpha_{ij} \ge \frac{\pi}{3}$ and $k \le \sqrt{2\pi n}\,2^{n-2}$. If $L \ge 3(1+\sqrt{2})^{n-1}$, then $\alpha_{ij} \ge \frac{\pi}{2} - \alpha_n$, where $\alpha_n = \arcsin \frac{1}{3}\left(\frac{1}{1+\sqrt{2}}\right)^{n-1}$. By the ball packing argument due to Abresch, cf. [A] part II, there are at most $2n$ vectors in $\mathbb{R}^n$ making a pairwise angle $\geq \frac{\pi}{2} - \alpha_n$. $\square$

Corollary 3.17 *Let M^n be a complete non-compact manifold with $K \geq 0$, $p \in M$. Then all critical points of $dist_p$ are contained in some ball of finite radius around p.*

As a consequence we obtain the following

Theorem 3.18 (Gromov) *Let M^n be a complete non-compact manifold with $K \geq 0$. Then M is homeomorphic to the interior of a compact manifold with boundary, hence M is of "finite" topological type.*

Proof. Since the critical points of dist_p are contained in some ball of finite radius, corollary 3.9 applies. $\square$

Recent examples of Sha and Yang show that a similar result does not hold for manifolds with positive Ricci curvature, cf. [SY1]. However if $Ric > 0$ and in addition $K > -\infty$ and the "diameter growth" of $\partial B(p,r)$ is of the order $o(r^{\frac{1}{n}})$, then the same conclusion as in the theorem holds, cf. [AG].

The above theorem 3.18 may be viewed as a weak version of the much more subtle soul theorem of Cheeger and Gromoll, by which M contains a compact totally geodesic submanifold S such that M is diffeomorphic to the normal bundle of S in M, cf. [CG1], [CE] and section 3.6.

3.5 An estimate for the sum of Betti numbers

In this section $\mathrm{H}_*(M)$ denotes the singular homology of M with coefficients in some arbitrary field $\mathcal{F}$. The k^{th} Betti number of M with respect to $\mathcal{F}$ is given by $b_k(M) = \dim_{\mathcal{F}} \mathrm{H}_k(M)$. For a compact n-manifold and by theorem 3.18 also for complete n-manifolds M of nonnegative curvature we have $\sum_{k=0}^{n} b_k(M) = \dim_{\mathcal{F}} \mathrm{H}_*(M) < \infty$.

By an ingeniously designed Morse theory for distance functions Gromov [G2] obtained the following result:

Theorem 3.19 (Gromov)

a) There is a constant $C(n)$ such that any complete n-manifold M of nonnegative curvature satisfies

$$\dim_{\mathcal{F}} H_*(M) \leq C(n).$$

b) *Given $D > 0$ and $\kappa < 0$ and n, then there is a constant $C_*(D, \kappa, n)$ such that any complete n-manifold with sectional curvature $K \geq \kappa$ and $diam M \leq D$ satisfies*

$$\dim_{\mathcal{F}} H_*(M) \leq C_*(D, \kappa, n)\,.$$

In his paper [G2] Gromov indicated that a similar theorem as (a) holds for manifolds with assymptotically nonnegative curvature. U. Abresch [A] gave the precise definition of "assymptotically nonnegative curvature" for which such a theorem can be proved. He also refined Gromov's method and developed the necessary tools to obtain the following result:

c) Let $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ be a decreasing function satisfying $\int_0^\infty r\lambda(r)\,dr \leq \infty$. then there is a constant $C_\#(n, \lambda)$ such that

$$\dim_{\mathcal{F}} H_*(M) \leq C_\#(n, \lambda)$$

for any complete Riemannian manifold M^n with sectional curvatures $K_r \geq -\lambda(r)$ at distance r from a given point $p \in M$.

Remarks

1. The lower bound for the sectional curvature cannot be replaced by a lower bound for the Ricci curvature: Sha and Yang recently have constructed metrics of positive Ricci curvature on the connected sum of an arbitrary number of copies of $S^n \times S^m$, cf. [SY2]. In [SY1] they also gave complete noncompact examples with positive Ricci curvature of infinite homology type.

2. Under the hypothesis in the theorem one cannot expect finiteness for the number of homotopy types. Here the lens spaces and also the simply connected Wallach examples [AW] should be observed.
   However Grove and Petersen have shown that there are only finitely many homotopy types of compact manifolds if in addition to the lower curvature bound and the upper diameter bound one assumes a lower bound for the volume, cf. [GP].

3. The methods for the proof of a) and b) are essentially the same. For the proof of c) Abresch had to develop a more general version of Toponogov's triangle comparison theorem, compare remark 8 in section 2. Though the proof of c) is somewhat more technical, the refined method of Abresch leads to a simplified proof of a) and b). It also gives a better estimate for the constants $C(n)$ and $C_*(n, \kappa, D)$ than in [G2].

For reasons of exposition we concentrate on the proof of a) using the refined version due to Abresch. So we assume $K \geq 0$ for the remainder of this section unless stated otherwise. We also will fix the constant

$$L = 3(1+\sqrt{2})^{n-1}$$

as determined in corollary 3.16.

It is convenient to use the the following notation in connection with metric balls: If B is a ball of radius r around p then ρB denotes the concentric ball of radius ρr around p.

In contrast to standard Morse theory one cannot estimate the dimension of the homology of the sublevels of distance functions (i.e. of metric balls) directly since the intersection of a ball with the cutlocus can be rather complicated. As a replacement for this part of the Morse theory Gromov introduces the concept of content:

Definition 3.20 *Let $Y \subset X$ be open subsets of M. The* content *of Y in X is defined as the rank of the inclusion map on the homology level*

$$cont(Y, X) := rk(H_*(Y) \to H_*(X)) .$$

The content of a metric ball B in M is defined as

$$cont(B) := cont(B, 5B) .$$

The content of B is a measure for how much of its homology survives after the inclusion map into $5B$. Clearly $\mathrm{cont}(B) = 1$ for any contractible ball B. By corollary 3.17 and the Isotopy Lemma 3.12 for sufficiently large balls B there is an isotopy of M which moves M into B. Therfore there is a map $f : M \to M$ such that the induced map f_* is the identity on $\mathrm{H}_*(M)$ and $f(M) \subset B \subset 5B \subset M$. Hence $\mathrm{cont}(B) = \mathrm{cont}(M, M) = \dim_{\mathcal{F}} \mathrm{H}_*(M)$.

The strategy for the proof now consists in showing that the content of any metric ball and hence of M is bounded by a constant $C(n)$. For this purpose Gromov introduces the concepts of *corank* and *compressibility* for metric balls with the following properties:

(i) Either a ball of content > 1 is incompressible or it can be deformed into an incompressible smaller ball of at least the same content and of at least the same corank.

(ii) The corank is bounded by a constant $k_0 \leq 2n$.

(iii) If a ball B of radius r and of corank k is incompressible, then any ball of radius $\leq \frac{r}{5L}$ with center in $\frac{3}{2}B$ has corank at least $k+1$.

(iv) A ball with maximal corank has content 1.

Now the proof is based on a reverse induction over the corank: By (i) only incompressible balls need to be considered. Suppose that the content of any ball of corank $> k$ is bounded by $a_k(n)$. Let B be an incompressible ball of radius r and corank k. Then B is covered by balls B_i of radius $\rho = \frac{r}{5L \cdot 10^{n+1}}$ such that the concentric balls $\frac{1}{2}B_i$ are disjoint. The maximal number N of these balls can be estimated from above by the Bishop-Gromov volume comparison argument. It depends only on n. Using property (iii) and the induction assumption, a topological argument stemming from a generalized Mayer-Vietoris sequence for nested coverings then is used to show that the content of B is bounded by $a_k(n) \cdot N^{n+1}$, completing the induction argument.

We start introducing the concept of compressibility which essentially corresponds to "*ρ − compressibility*" used by Abresch with the fixed value $\rho = 5$.

Definition 3.21 *A ball B of radius r in M is called* compressible *if there is a ball $\tilde{B}$ of radius $\tilde{r} \leq \frac{3}{5}r$ around some point in $2B$ such that there is an isotopy of M which is fixed outside $5B$ and which moves B into $\tilde{B}$. Briefly we say that B is compressible into $\tilde{B}$ when these conditions hold.*

If B is compressible into $\tilde{B}$, then $\tilde{B} \subset 5\tilde{B} \subset 5B$ and the pairs $(5B, B)$ and $(5B, \tilde{B})$ are homotopically equivalent. Therefore it is clear that

$$\mathrm{cont}(B) \leq \mathrm{cont}(\tilde{B})\,.$$

Consequently for each ball B of content > 1 there is an incompressible ball $B_0 \subset 5B_0 \subset 5B$ such that $\mathrm{cont}(B_0) \geq \mathrm{cont}(B)$. For this observe that the injectivity radius on the compact ball $5\overline{B}$ is bounded below by some constant $\delta > 0$ and if a ball can be compressed successively into a final ball of radius $\leq \delta$ then it must have content 1 since the δ-balls are contractible.

Lemma 3.22 *Suppose B is an incompressible ball of radius r. Then for any point $\tilde{p} \in 2B$ there must be a critical point $\tilde{q}$ for the distance function $dist_{\tilde{p}}$ in the compact annulus $A_{\tilde{p}} = \overline{B(\tilde{p}, 3r) \setminus B(\tilde{p}, \frac{3}{5}r)}$.*

Proof. Suppose for some point $\tilde{p} \in 2B$ there is no critical point in $A_{\tilde{p}}$. Let $\tilde{B} = B(\tilde{p}, \frac{3}{5}r)$. Then we have inclusions $B \subset B(\tilde{p}, 3r)$ and $\overline{B(\tilde{p}, 3r)} \subset 5B$. By the isotopy lemma 3.12 there is an isotopy of M which is fixed outside $5B$ moving $B(\tilde{p}, 3r)$ and hence B into $\tilde{B}$ contradicting the incompressibility of B. $\square$

Definition 3.23 *Given $p \in M$, $r > 0$. Let $k_r(p)$ be the maximal number of critical points q_j, $j = 1, \ldots, k_r(p)$, for $dist_p$ satisfying*

$$dist_p(q_j) \geq 3Lr \quad and \quad dist_p(q_{j+1}) \leq \frac{1}{L} dist_p(q_j)\,.$$

The corank *of the ball $B = B(p, r)$ is defined as*

$$corank(B) = \inf\{k_r(\tilde{p}) \mid \tilde{p} \in 5B\}\,.$$

Note that $k_r(p) \leq 2n$ and therefore $\mathrm{corank}(B) \leq 2n$ by the choice of L. If B is compressible into $\tilde{B}$, then $\mathrm{corank}(B) \leq \mathrm{corank}(\tilde{B})$.
As an immediate consequence of the previous Lemma 3.22 we have

Corollary 3.24 *Suppose $B = B(p, r)$ is incompressible and $\hat{r} \leq \frac{r}{5L}$.*

a) If $\tilde{p} \in 2B$, then $k_{\hat{r}}(\tilde{p}) \geq 1 + corank(B)$.

b) If $\hat{p} \in \frac{3}{2}B$ and $\hat{B} = B(\hat{p}, \hat{r})$, then $corank(\hat{B}) \geq 1 + corank(B)$.

Proof. For a) let q_j, $1 \leq j \leq k_r(\tilde{p}) =: k$ be critical points with $\mathrm{dist}_{\tilde{p}}(q_j) \geq 3Lr$ and $dist_{\tilde{p}}(q_{j+1}) \leq \frac{1}{L}\mathrm{dist}_{\tilde{p}}(q_j)$. Since B is incompressible, there is a critical point q_{k+1} for $\tilde{p}$ in the annulus $A_{\tilde{p}}$ as in lemma 3.22. Thus $3L\hat{r} \leq \frac{3}{5}r \leq \mathrm{dist}_{\tilde{p}}(q_{k+1}) \leq 3r \leq \frac{1}{L}\mathrm{dist}_{\tilde{p}}(q_j)$. Now the $k_r(\tilde{p}) + 1$ points q_j satisfy the condition for the definition of $k_{\hat{r}}(\tilde{p})$ hence a) follows.
For b) observe the inclusions $5\hat{B} \subset (\frac{3}{2} + \frac{1}{L})B \subset 2B$. Now $k_{\hat{r}}(\tilde{p}) \geq 1 + \mathrm{corank}(B)$ for any $\tilde{p} \in 5\hat{B}$. $\square$

As a consequence a ball of maximal corank must have content 1: If B has maximal corank, then because of b) in the lemma, B must be compressible into a ball of radius $\frac{3}{5}r$ with the same maximal corank and at least the same content. This procedure can be repeated k times until one reaches a ball $\tilde{B}$ of radius $(\frac{3}{5})^k r$ which is smaller than the injectivity radius on the compact ball $5\overline{B}$. Then $\mathrm{cont}(\tilde{B}) = 1$ and hence $\mathrm{cont}(B) = 1$.

These are the basic ingredients from critical point theory. We now turn to the covering arguments.

Lemma 3.25 *Given an n-dim Riemannian manifold M of nonnegative Ricci curvature, a ball B of radius r and a covering of B by balls $B_1, \ldots, B_N$ of radius $\varepsilon \leq r$ with center in B such that the corresponding balls $\frac{1}{2}B_1, \ldots, \frac{1}{2}B_N$ are disjoint. Then*

$$N \leq (6\frac{r}{\varepsilon})^n\,.$$

Proof. Choose i_0 such that the ball $\frac{1}{2}B_{i_0}$ with center p_0 has the smallest volume among all the given balls. The ball $\hat{B}$ around p_0 of radius $3r$ contains all the B_i. Therefore

$$N \leq \frac{\mathrm{vol}\hat{B}}{\mathrm{vol}\frac{1}{2}B_{i_0}} \leq \left(\frac{3r}{\frac{1}{2}\varepsilon}\right)^n$$

where the last inequality is the Bishop-Gromov estimate for the volume of concentric balls, which has been discussed in the first series of these lectures given by K. Grove, compare also [K] for a proof. □

Since the proof of theorem 3.19 will be based on reverse induction over the corank, we introduce the following notation:
Let $k_0 \leq 2n$ be the maximal corank of metric balls. For $0 \leq k \leq k_0$ we denote by $\mathcal{B}_k$ the set of balls having corank $\geq k$.

The topological information for the induction step is contained in the next lemma:

Lemma 3.26 *Suppose cont*$(\tilde{B})$ *is bounded by a constant* a_k *for any* $\tilde{B} \in \mathcal{B}_k$. *Furthermore let* $B \in \mathcal{B}_{k-1}$ *be incompressible. Then*

$$\mathit{cont}(B) \leq a_k \cdot N^{n+1}$$

where $N \leq (3L \cdot 10^{n+2})^n$.

Proof. Choose a covering of B by balls $B_1, \ldots, B_N$ of radius $\varepsilon(r) = \frac{r}{5L \cdot 10^{n+1}}$ such that $\frac{1}{2}B_1, \ldots, \frac{1}{2}B_N$ are disjoint. Then by lemma 3.25 $N \leq (3L \cdot 10^{n+2})^n$. For $0 \leq j \leq n+1$ we also consider the coverings $B_1^j, \ldots, B_N^j$ where $B_i^j = 10^j \cdot B_i$. The radii of all these balls are $\leq \frac{r}{5L}$. By corollary 3.24 we have $\mathrm{corank}(B_i^j) \geq 1 + \mathrm{corank}(B) \geq k$, hence $\mathrm{cont}(B_i^j) \leq a_k$. Using the result on the nested coverings in corollary 4.2 of the appendix, we obtain

$$\mathrm{cont}(\bigcup_i B_i^0, \bigcup_i B_i^{n+1}) \leq \sum_{\ell=0}^{n} \sum_{i_0 < \ldots < i_\ell} \mathrm{cont}(B_{i_0}^{n-\ell} \cap \ldots \cap B_{i_\ell}^{n-\ell}, B_{i_0}^{n+1-\ell} \cap \ldots \cap B_{i_\ell}^{n+1-\ell}).$$

By the choice of the radii and the triangle inequality we have inclusions $5B_i^j \subset 5B$, $5B_{i_1}^j \subset B_{i_2}^{j+1}$ and therefore

$$B \subset \bigcup_i B_i^0 \subset \bigcup_i B_i^{n+1} \subset 5B$$

and

$$B_{i_0}^{n-\ell} \cap \ldots \cap B_{i_\ell}^{n-\ell} \subset B_{i_0}^{n-\ell} \subset 5B_{i_0}^{n-\ell} \subset B_{i_0}^{n+1-\ell} \cap \ldots \cap B_{i_\ell}^{n+1-\ell}.$$

The first chain of inclusions implies $\mathrm{cont}(B) \leq \mathrm{cont}(\bigcup_i B_i^0, \bigcup_i B_i^{n+1})$, and from the second we conclude that the content of any of the intersections is bounded by $\mathrm{cont}(B_{i_0}^{n-\ell}) \leq a_k$. Since the number of terms in the sum on the right hand side is bounded by N^{n+1}, compare (81) in the appendix, the proof is complete. □

Proof of Theorem 3.19 a): Reverse induction over the corank: For $B \in \mathcal{B}_{k_0}$ we have $\mathrm{cont}(B) = 1$. Assume now that $\mathrm{cont}(B) \leq a_k(n)$ for any $B \in \mathcal{B}_k$. Let $B \in \mathcal{B}_{k-1}$ If B is compressible and $\mathrm{cont}(B) > 1$, then B can be compressed into a ball of at least the same content and of at least the same corank. Therefore we can assume that B is incompressible. Now lemma 3.26 applies and we get $\mathrm{cont}(B) \leq a_k(n) \cdot N^{n+1}$. Since $k_0 \leq 2n$ we get recursively

$$\dim_{\mathcal{F}} \mathrm{H}_*(M) = \mathrm{cont}(M) \leq N^{2n^2+2n}$$

where $N = (3L \cdot 10^{n+2})^n$, $L = 3(1+\sqrt{2})^{n-1}$. Using $L < 3^{n+1}$, an explicit rough estimate for $C(n)$ is given by

$$C(n) \leq 10^{3n^4+9n^3+6n^2}\,.$$

□

Remarks

1. Note that the exponent in the estimate for $C(n)$ is a polynomial of order 4 in n. Gromov's original constant depended double exponentially on n. The reason for this improvement due to Abresch is the choice of L, the modification of corank and compressibility to eliminate one of Gromov's critical point lemmas which all together gave a better estimate for the corank, and finally the improvement of the estimate in the inductive lemma 3.26 where Gromov uses the estimate $\mathrm{cont}(B) \leq a_k \cdot 2^N$.

2. The estimate for the constant $C(n)$ still seems to be far away from reality. Known examples of n-manifolds with nonnegative curvature all have a sum of Betti numbers $\leq 2^n$.

3.6 The soul theorem

This final section is devoted to the soul theorem, cf. [GM1], [CG1].

Theorem 3.27 *(Cheeger, Gromoll) Let M^n be a complete noncompact manifold of nonnegative curvature K. Then there is a compact totally geodesic submanifold S in*

M such that M is diffeomorphic to the normal bundle $\nu(S)$ of S. If $K > 0$, then M is diffeomorphic to $I\!R^n$.

We first introduce a few basic concepts which are needed for the proof.

Definition 3.28 *A nonempty subset C of M is called* totally convex *if for arbitrary points $p, q \in C$ any geodesic with endpoints p and q is contained in C.*

Definition 3.29 *A* ray *in M is a normal geodesic $c : [0, \infty[\to M$ for which any finite segment is minimal. For a ray $c : [0, \infty[\to M$ we define the* halfspaces *B_c respectively H_c by*

$$\begin{aligned} B_c &= \bigcup_{t>0} B(c(t), t) \\ H_c &= M \setminus B_c \end{aligned}$$

where $B(c(t), t)$ is the open metric ball of radius t around $c(t)$.

Note that in a complete noncompact manifold M for any $p \in M$ there exists a ray $c : [0, \infty[\to M$ with initial point $c(0) = p$. For a sequence $q_i \in M$ with $\lim_{i\to\infty}(p, q_i) = \infty$ and normal minimal geodesics c_i from p to q_i any limiting geodesic c obtained from a convergent subsequence of c_i will be a ray emanating from p. ($\dot{c}_i(0)$ has an accumulation point in the compact unit sphere in T_pM).

The basic observation about the halfspaces H_c is the following.

Lemma 3.30 *If M is complete, noncompact of nonnegative sectional curvature, then H_c is totally convex for any ray in M.*

Proof. Suppose H_c is not totally convex, i.e. there is a geodesic $c_0 : [0, 1] \to M$ with endpoints $c_0(0), c_0(1) \in H_c$ but $c_0(s) \in B_c$ for some $s \in]0, 1[$. Then $q := c_0(s) \in B(c(t_0), t_0)$ for some $t_0 > 0$ and hence $q \in B(c(t), t)$ for any $t \geq t_0$ by the triangle inequality: In fact setting

$$t_0 - \varepsilon = \mathrm{dist}(q, c(t_0)), \quad \varepsilon > 0$$

we have

$$\begin{aligned} \mathrm{dist}(q, c(t)) &\leq \mathrm{dist}(q, c(t_0)) + \mathrm{dist}(c(t_o), c(t)) \\ &= (t_0 - \varepsilon) + (t - t_0) = t - \varepsilon \end{aligned}$$

for $t \geq t_0$.

Let $c_0(s_t)$ be a point on c_0 which is closest to $c(t)$. Further consider the restriction $c_0^t := (c_0|_{[0,s_t]})^{-1}$ and a minimal geodesic c_1^t from $c_0(s_t)$ to $c(t)$. Since $c_0^t(0) = c_0(s_t)$ is the closest point to $c(t)$ on c_0 we have $\sphericalangle(\dot{c}_0^t(0), \dot{c}_1^t(0)) = \frac{\pi}{2}$. Furthermore $|c_1^t| = \mathrm{dist}(c_0^t(0), c(t)) = \mathrm{dist}(c_0(s_t), c(t)) \leq \mathrm{dist}(q, c(t)) \leq (t-\varepsilon)$ and $|c_0^t| \leq |c_0|$. Consider now the hinge $c_0^t, c_1^t, \frac{\pi}{2}$. Using Toponogov's theorem 2.2 part B with comparison space $\mathbb{R}^2$ and the law of cosines we obtain

$$\mathrm{dist}^2(c_0^t(s_t), c(t)) = \mathrm{dist}^2(c_0(0), c(t)) \leq |c_0^t|^2 + |c_1^t|^2 \leq |c_0|^2 + (t-\varepsilon)^2$$

Furthermore $\mathrm{dist}(c_0(0), c(t)) \geq t$ since $c_0(0) \in H_c = M \setminus B_c$. Therefore $t^2 \leq |c_0|^2 + (t-\varepsilon)^2$, which for large values of t is a contradiction. $\square$

We now fix a point $p \in M$. For a ray $c : [0,\infty[\to M$ we also consider the restriction $c_t := c|_{[t,\infty[}$. Let

$$C_t := \bigcap_c H_{c_t}$$

where the intersection is taken over all the rays c emanating from p.

Lemma 3.31 *C_t is a compact totally convex set for all $t \geq 0$, moreover*

a) $C_{t_2} \supset C_{t_1}$ *for* $t_2 \geq t_1$ *and*
$C_{t_1} = \{q \in C_{t_2} \mid \mathit{dist}(q, \partial C_{t_2} \geq t_2 - t_1\}$,
in particular
$\partial C_{t_1} = \{q \in C_{t_2} \mid \mathit{dist}(q, \partial C_{t_2}) = t_2 - t_1\}$

b) $\bigcup_{t \geq 0} C_t = M$

c) $p \in \partial C_0$

Proof. Clearly C_t is totally convex and closed and $p \in C_t$. If some C_t were not compact it would contain a ray $c : [0,\infty[\to C_t$ starting from p (use the same argument as for the existence of rays in a noncompact manifold). Now $c(t') \notin C_t$ for $t' > t$, contradicting the definition of C_t. Statement c) is obvious from the construction of C_t. The proof of a) and b) now is an exercise using only the definiton of C_t and the triangle inequality, cf. [CE]. $\square$

Note that the interior of C_t is nonempty for $t > 0$. This is not true for C_0 in general as can be seen on the paraboloid of revolution in $\mathbb{R}^3$: If p is the umbilic point of the paraboloid then $C_0 = \{p\}$.

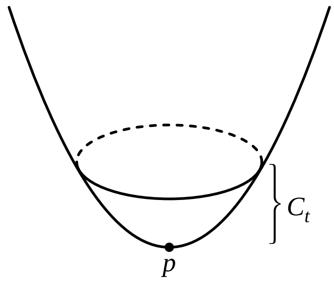

Figure 6: paraboloid

The C_t provide an expanding filtration of M by compact totally convex sets. Our next goal is to construct minimal totally convex sets by a contraction procedure which will be used to find a soul S. For this important part of the proof we also need the local concept of convexity:

Definition 3.32 *A subset A of M is called* strongly convex *if for any $q, q' \in A$ there is a unique minimal geodesic from q to q' which is contained in A.*

Recall that there is a continuous function $r : M \to]0, \infty]$, the *convexity radius* such that for any $p \in M$, any open metric ball B which is contained in $B(p, r(p))$ is strongly convex, cf [GKM].

Definition 3.33 *We say that a subset C of M is* convex *if for any $p \in \overline{C}$ there is a number $0 < \varepsilon(p) < r(p)$ such that $C \cap B(p, \varepsilon(p))$ is strongly convex.*

Note that a totally convex set is convex and connected. Also the closure of a convex set is again convex.

Let C be a connected nonempty convex subset of M. For $0 \leq l \leq n$ we may consider the collection $\{N_\alpha^l\}$ of smooth $l-$dim submanifolds of M such that $N_\alpha^l \subset C$. Let k denote the largest integer such that $\{N_\alpha^k\}$ is nonempty and $N := \bigcup_\alpha N_\alpha^k \subset C$.

Lemma 3.34 *N is a smooth connected totally geodesic submanifold of M and $C \subset \overline{N}$. Moreover $\overline{N} = \overline{C}$ is a topological manifold with possibly empty bounary $\partial\overline{N} = \overline{N} \setminus N$.*

Proof (outline). The full details are technical, therefore we only give the main idea, [CG1], [CE]. Let $p \in N$ and $\varepsilon(p)$ as in the definition above. Then $p \in N_\alpha^k$ for some α. Therefore we can choose a neighborhood $U \subset N_\alpha \cap B(p, \frac{1}{2}\varepsilon(p))$ of p in

N and $0 < \delta < \frac{1}{2}\varepsilon(p)$ such that $\exp|_{\nu_\delta(U)}$ is a diffeomorphism onto a neighborhood T_δ of p in M, where $\nu_\delta(U) = \{v \in (TU)^\perp \mid \|v\| < \delta\} \subset TM$ is the δ-tube in the normal bundle $\nu(U)$ of U. To prove that N is a submanifold it suffices to show that $N \cap T_\delta = U$. Suppose $q \in (N \cap T_\delta) \setminus U \subset (C \cap T_\delta) \setminus U$. Let q' be the closest point to q in $\overline{U}$. Then $q' \in U$, otherwise we get a contradiction to the invertibility of $\exp|_{\nu_\delta(U)}$ close to q. The minimal geodesic from q to q' then is orthogonal to U. By the choice of $\delta < \frac{1}{2}\varepsilon(p)$ the exponential map in the ball of radius δ around q' is invertible. Therefore all the unique minimal geodesics from q to q'' for q'' in some neighborhood U' of q' are transversal to U and are contained in C. The conical set $\{\exp tu \mid u \in M_q, \|u\| < \varepsilon(q),\ \exp(u) \in U',\ 0 < t < 1\}$ then is a $(k+1)$-dimensional submanifold in C which contradicts the definition of k. From the existence of T_δ and the convexity of C it follows that N is totally geodesic. For the remaining statements we refer to [CG1] and [CE]. □

Definition 3.35 *Let C be a convex subset of M. The* tangent cone *to C at a point $p \in C$ is by definition the set*

$$T_pC = \{v \in T_pM \mid \exp(t\frac{v}{\|v\|}) \in N \textit{ for some } 0 < t < r(p)\} \cup \{0\} .$$

Clearly if $p \in N = \operatorname{int}(C)$, then $T_pC = T_pN$. The following lemma contains all the technical information about T_pC we need.

Lemma 3.36 (tangent cone lemma) *Let $C \subset M$ be convex and $p \in \partial C$.*

a) Then $T_pC \setminus \{0\}$ is contained in an open halfspace of T_pM.

b) Suppose that there exists $q \in \mathit{int}\, C$ and a minimal normal geodesic $c : [0, d] \to C$ from q to p such that $|c| = \mathit{dist}(q, \partial C)$. Then

$$T_pC \setminus \{0\} = \{v \in \hat{T}_pC \mid \measuredangle(v, -\dot{c}(d)) < \frac{\pi}{2}\} ,$$

where $\hat{T}_pC$ is the subspace of T_pM spanned by T_pC.

Proof. a) T_pC is convex in T_pM since C is convex. If $T_pC \setminus \{0\}$ is not contained in an open halfspace of T_pM, then T_pC must be a linear subspace of T_pM of dimension $\dim(\operatorname{int} C))$ and hence p is an interior point of C. For b) and the details of the the analysis of convex sets we refer to [CG1]. □

The following lemma is the key for constructing the soul of M via a contraction procedure.

Lemma 3.37 (contraction lemma) *Suppose M has nonnegative sectional curvature and $C \subset M$ is a closed totally convex subset with $\partial C \neq \emptyset$. We set*

$$C^a = \{p \in C \mid dist(p, \partial C) \geq a\} \ , \qquad C^{max} = \bigcap_{C^a \neq \emptyset} C^a \, .$$

Then

a) C^a is closed and totally convex.

b) $\dim C^{max} < \dim C$.

c) If $K > 0$ then C^{max} is a point.

This is a corollary of the following more general lemma:

Lemma 3.38 *Under the assumptions of lemma 3.37, $\psi := dist_{\partial C} : M \to I\!R$ is a concave function, i.e. for any normal geodesic c which is contained in C we have*

$$\psi(c(\lambda t_1 + (1-\lambda)t_2)) \geq \lambda \psi(c(t_1)) + (1-\lambda)\psi(c(t_2)) \ . \tag{65}$$

If the sectional cuvature satisfies $K > 0$ then the strict inequality holds in (65).

Proof. It is sufficient to show that for any point $c(s_0)$ of c there is a number $\delta > 0$ such that $\psi(c(s))$ is bounded above by a linear function $h(s)$ on $]s_0-\delta, s_0+\delta[$ satisfying $h(s_0) = \psi(c(s_0)) =: d$. Let c_{s_0} be a distance minimizing normal geodesic of length d from $c(s_0)$ to ∂C and $\alpha := \sphericalangle(\dot{c}_{s_0}(0), \dot{c}(s_0))$. Then we can take

$$h(s) = d - (s - s_0)\cos\alpha \ .$$

To show $h(s) \geq \psi(c(s))$ we consider the three cases $\alpha = \frac{\pi}{2}$, $\alpha > \frac{\pi}{2}$, $\alpha < \frac{\pi}{2}$. Note that we only have to consider points $s \geq s_0$.
Case $\alpha = \frac{\pi}{2}$: Let E denote the parallel unit vector field along c_{s_0} with $E(0) = \dot{c}(s_0)$. By the second comparison theorem of Rauch, there is a number $\delta > 0$ such that the length of the curve $c_s(t) = \exp(s-s_0)E(t)$ has length $|c_s| \leq d = |c_{s_0}|$ for $0 \leq s-s_0 \leq \delta$. The geodesic $\bar{c} : s \mapsto \exp(s - s_0)E(d)$ is orthogonal to c_{s_0} at $q := c_{s_0}(d) \in \partial C$, hence $\dot{\bar{c}}(0) \notin T_qC$ by lemma 3.36, so that $\bar{c}(t) \notin \operatorname{int} C$ for $0 < t < \varepsilon(q)$. Therefore $\psi(c(s)) \leq |c_s| \leq d = d - (s - s_0)\cos\frac{\pi}{2}$.
Case $\alpha > \frac{\pi}{2}$: Let $E(0) \perp \dot{c}_{s_0}(0)$ be the unique unit vector in the convex cone spanned by $\dot{c}(s_0)$ and $\dot{c}_{s_0}(0)$ and extend it to the parallel vector field E along c_{s_0}. Define c_s as in the first case to obtain

$$|c_s| \leq d \ . \tag{66}$$

Applying the hinge version of Toponogov's theorem (or just Rauch I) to the hinge with geodesics $t \mapsto \exp tE(0)$, $0 \le t \le (s-s_0)\cos(\alpha - \frac{\pi}{2})$ and $t \mapsto c(s_0+t)$ with angle $\alpha - \frac{\pi}{2}$, one obtains

$$\operatorname{dist}\Big(c(s), \exp((s-s_0)\cos(\alpha - \frac{\pi}{2})E(0)\Big) \le -(s-s_0)\cos\alpha\ . \tag{67}$$

Combining (66) and (67), the inequality $\psi(c(s)) \le d-(s-s_0)\cos\alpha$ follows.

Case $\alpha < \frac{\pi}{2}$ **:** Choose the point $c_{s_0}(t_s)$ on c_{s_0} such that $\operatorname{dist}(c(s), c_{s_0}([0,d])) = \operatorname{dist}(c(s), c_{s_0}(t_s))$ and a normal minimal geodesic a_s from $c_{s_0}(t_s)$ to $c(s)$.
Then $\measuredangle(\dot a_s(0), \dot c_{s_0}(t_s)) = \frac{\pi}{2}$. Further E denotes the parallel vector field along $c_{s_0}|_{[t_s,d]}$ with $E(t_s) = \dot a_s(0)$. The curve $c_s(t) = \exp(|a_s|\,E(t))$, $t_s \le t \le d$, is of length $|c_s| \le (d-t_s)$ for $s - s_0 < \delta$ if δ is sufficiently small. As before $\operatorname{dist}(c(s), \partial C) \le |c_s|$, thus

$$\operatorname{dist}(c(s), \partial C) \le (d-t_s) \tag{68}$$

Applying the hinge version of Toponogov's theorem (or just Rauch I) to the hinges $(c|_{[s_0,s]},\ c_{s_0}|_{[0,t_s]},\ \alpha)$ respectively $(c_{s_0}^{-1}|_{[0,t_s]},\ a_s,\ \frac{\pi}{2})$, we obtain $|a_s|^2 \le (s-s_0)^2 + t_s^2 - 2t_s(s-s_0)\cos\alpha$ respectively $(s-s_0)^2 \le |a_s|^2 + t_s^2$, hence

$$-t_s \le -(s-s_0)\cos\alpha. \tag{69}$$

From (68) and (69) the estimate $\psi(c(s)) \le h(s)$ follows.
The discussion of the strict inequality in the case $k > 0$ is left to the reader. $\square$

Proof of the soul theorem. Let $p \in M$ and consider the filtration of M by compact totally convex sets C_t as in lemma 3.31. If $\partial C_0 = \emptyset$ let $S = C_0$. If $\partial C_0 \ne \emptyset$, application of the contraction lemma 3.37 to the compact totally convex set C_0 gives us a compact totally convex set C_0^{max} of dimension $< \dim C_0$. Repeating this procedure leads us in a finite number ($\le n$) of steps to a compact totally convex set $S \subset C_0$ with $\dim S < n$ and $\partial S = \emptyset$. In particular S is a compact totally geodesic submanifold of M.
We now show that M is diffeomorphic to the normalbundle $\nu(S)$. The diffeomorphism is constructed by means of the flow of a gradient-like vector field of dist_S. Let $q \in M \backslash S$. Then $q \in \partial C_t$ for some $t \ge 0$ or $q \in \operatorname{int} C_0$. By the contraction lemma 3.37 we have either $q \in \partial C_0^a$ for some $a \ge 0$ or $q \in \operatorname{int} C_0^{max}$. Repeating this argument a finite number of times, we find a compact totally convex set C such that $q \in \partial C$ and $S \subset \operatorname{int} C$. Any geodesic from q to S has its initial tangent vector in the tangent cone T_qC. Hence all

such initial vectors are contained in an open half space of T_qM, compare lemma 3.36. Therefore dist_S has no critical points on $M \setminus S$. Choose $\varepsilon > 0$ such that $exp|_{\nu_\varepsilon(S)}$ is a diffeomorphism onto the ε-tube around S. Here $\nu_\varepsilon(S) = \{v \in TS^\perp \mid \|v\| < \varepsilon\}$. Then $X_1 = \mathrm{grad}\ \mathrm{dist}_S$ is a gradient-like vector field on $\exp(\nu_\varepsilon(S)) \setminus S$ such that $\langle X_1|_q, \dot{c}_q(0)\rangle = -1$ for the unique minimal normal geodesic c_q from q to S. Therefore one can construct a global gradient-like vector field X on $M \setminus S$ such that $\langle X_q, \dot{c}(0)\rangle < 0$ for any distance minimizing geodesic from q to S and $X_q = X_1|_q$ for $q \in \exp(\nu_{\varepsilon/2}(S))$. Let Ψ be the flow of X. Define $F : \nu(S) \to M$ as follows: $F(v) := exp(v)$ for $\|v\| \leq \frac{\varepsilon}{4}$ and $F(tv) := \Psi(\exp(\frac{\varepsilon}{4}v), t - \frac{\varepsilon}{4})$ for $v \in \nu_1(S)$ and $t \geq \frac{\varepsilon}{4}$. Then F is a diffeomorphism as follows easily by using (64). $\square$

Remarks

1. A soul of M is not uniquely determined in general as can be seen by looking at cylinders. However any two souls of M are isometric, cf. [S] and [Y].
2. If $\mathrm{codim}(S) = 1$, then $\exp|_{\nu(S)}$ is an is an isometry between $\nu(S)$ with its standard (flat) bundle metric and M, cf. [CG1].
3. In general the normal bundle $\nu(S)$ need not to be trivial. Furthermore M is not locally isometric to a product $S \times \mathbb{R}^k$ in general. By the Toponogov splitting theorem, cf. [CG1], however any line in M splits off isometrically, so that M is isometric to $\bar{M} \times \mathbb{R}^k$, where $\mathbb{R}^k$ carries the standard flat metric and $\bar{M}$ does not contain any lines. This even holds for manifolds of nonnegative Ricci curvature, cf. [CG2], [EH]. More generally Strake [St] has shown the following: Suppose the holonomy group of $\nu(S)$ is trivial, then M is isometric to $S \times \mathbb{R}^k$ where $\mathbb{R}^k$ carries a metric of nonnegative curvature. For further results in this context we also refer to [ESS].
4. For a discussion on the structure of the fundamental group see [CG1].
5. There is no analogue of the soul theorem for complete open manifolds of positive Ricci curvature, cf. the examples in [GM2], [SY1] and [B], but compare also the result in [AG].

4 Appendix: A topological Lemma

Theorem 4.1 (Nested Coverings) *Let* $B_i^0 \subset B_i^1 \subset \ldots \subset B_i^{m+1}$ *,*$1 \leq i \leq N$*, be a family of nested open subsets in a topological space* X*, and let* $X^j := \bigcup_{i=1}^N B_i^j$ *for* $0 \leq j \leq m+1$*. Then*

$$
\begin{aligned}
& rk\Big(H_p(X^0) \to H_p(X^{p+1})\Big) \\
& \leq \sum_{k=0}^{p} \sum_{i_0<\ldots<i_k} rk\Big(H_{p-k}(B_{i_0}^{p-k} \cap \ldots \cap B_{i_k}^{p-k}) \to H_{p-k}(B_{i_0}^{p+1-k} \cap \ldots \cap B_{i_k}^{p+1-k})\Big)
\end{aligned}
$$

for $0 \leq p \leq m$ *; here* $H_p(\ldots)$ *stands for singular homology with coefficients in some arbitrary field* $\mathcal{F}$*.*

Proof. Let

$$
\mathrm{C}_{p,q}^j := \bigoplus_{i_0<\ldots<i_q} S_p\Big(B_{i_0}^j \cap \ldots \cap B_{i_q}^j\,;\mathcal{F}\Big) \qquad \text{and} \qquad \mathrm{A}_{p,0}^j := S_p^{\mathcal{U}}\Big(X^j\,;\mathcal{F}\Big) \tag{70}
$$

stand for the the groups of singular simplices which are fine w.r.t. the covering of X^j by the B_i^j . Whenever $q > 0$, homomorphisms $\delta_{p,q}^j : \mathrm{C}_{p,q}^j \to \mathrm{C}_{p,q-1}^j$ which commute with the differentials of the singular chain complexes $\mathrm{C}_{*,q}^j := \bigoplus_p \mathrm{C}_{p,q}^j$ can be defined in the manner of Cech homology: one adds up the inclusions $S_p\Big(B_{i_0}^j \cap \ldots \cap B_{i_q}^j\,;\mathcal{F}\Big) \to S_p\Big(B_{i_0}^j \cap \ldots \cap \widehat{B_{i_\mu}^j} \cap \ldots \cap B_{i_q}^j\,;\mathcal{F}\Big)$ with sign $(-1)^\mu$. Defining similarly maps $\hat{\delta}_{p,0}^j : \mathrm{C}_{p,0}^j \to \mathrm{A}_{p,0}^j$, one obtains on each level j separately a long exact sequence of chain complexes — the *generalized Mayer-Vietoris sequence* [BT, pp. 186-188] :

$$
\longrightarrow \mathrm{C}_{*,q}^j \xrightarrow{\delta_{*,q}^j} \mathrm{C}_{*,q-1}^j \longrightarrow \ldots \xrightarrow{\delta_{*,1}^j} \mathrm{C}_{*,0}^j \xrightarrow{\hat{\delta}_{*,0}^j} \mathrm{A}_{*,0}^j \longrightarrow 0 \tag{71}
$$

This sequence is natural w.r.t. the inclusion maps ($j_1 < j_2$):

$$
\gamma_{*,q}^{j_1,j_2} \;:\; \mathrm{C}_{*,q}^{j_1} \longrightarrow \mathrm{C}_{*,q}^{j_2} \tag{72}
$$

$$
\alpha_{*,0}^{j_1,j_2} \;:\; \mathrm{A}_{*,0}^{j_1} \longrightarrow \mathrm{A}_{*,0}^{j_2} \tag{73}
$$

For $q \geq 1$ we set $\mathrm{A}_{p,q}^j := \mathrm{im}(\delta_{p,q}^j)$ and define $\alpha_{*,q}^{j_1,j_2}$ as the restriction of $\gamma_{*,q-1}^{j_1,j_2}$. With this shorthand the generalized Mayer-Vietoris sequence splits naturally into short exact sequences of chain complexes:

$$
0 \longrightarrow \mathrm{A}_{*,q+1}^j \longrightarrow \mathrm{C}_{*,q}^j \longrightarrow \mathrm{A}_{*,q}^j \longrightarrow 0 \tag{74}
$$

Taking the corresponding long exact homology sequences leads in particular to *both* the commutative diagrams with exact rows:

$$\begin{array}{ccccc}
\mathrm{H}_p(\mathrm{C}^0_{*,q}) & \longrightarrow & \mathrm{H}_p(\mathrm{A}^0_{*,q}) & \longrightarrow & \mathrm{H}_{p-1}(\mathrm{A}^0_{*,q+1}) \\
\downarrow & & \downarrow{\scriptstyle \bar{\alpha}^{0,p}_{p,q}} & & \downarrow{\scriptstyle \bar{\alpha}^{0,p}_{p-1,q+1}} \\
\mathrm{H}_p(\mathrm{C}^p_{*,q}) & \longrightarrow & \mathrm{H}_p(\mathrm{A}^p_{*,q}) & \longrightarrow & \mathrm{H}_{p-1}(\mathrm{A}^p_{*,q+1}) \\
\downarrow{\scriptstyle \bar{\gamma}^{p,p+1}_{p,q}} & & \downarrow{\scriptstyle \bar{\alpha}^{p,p+1}_{p,q}} & & \downarrow \\
\mathrm{H}_p(\mathrm{C}^{p+1}_{*,q}) & \longrightarrow & \mathrm{H}_p(\mathrm{A}^{p+1}_{*,q}) & \longrightarrow & \mathrm{H}_{p-1}(\mathrm{A}^{p+1}_{*,q+1})
\end{array} \tag{75}$$

when $1 \leq p \leq m$, and

$$\begin{array}{ccccccc}
\mathrm{H}_0(\mathrm{C}^0_{*,q}) & \longrightarrow & \mathrm{H}_0(\mathrm{A}^0_{*,q}) & \longrightarrow & \mathrm{H}_{-1}(\mathrm{A}^0_{*,q+1}) & = & 0 \\
\downarrow{\scriptstyle \bar{\gamma}^{0,1}_{0,q}} & & \downarrow{\scriptstyle \bar{\alpha}^{0,1}_{0,q}} & & \downarrow & & \\
\mathrm{H}_0(\mathrm{C}^1_{*,q}) & \longrightarrow & \mathrm{H}_0(\mathrm{A}^1_{*,q}) & \longrightarrow & \mathrm{H}_{-1}(\mathrm{A}^1_{*,q+1}) & = & 0
\end{array} \tag{76}$$

else. Here the vanishing occurs already on the chain level: $\mathrm{A}^j_{-1,q+1} \subset \mathrm{C}^j_{-1,q} = 0$. When applying standard diagram chasing techniques, (75) and (76) yield the following estimates respectively:

$$\begin{aligned}
\mathrm{rk}(\bar{\alpha}^{0,p+1}_{p,q}) &= \mathrm{rk}(\bar{\alpha}^{p,p+1}_{p,q} \circ \bar{\alpha}^{0,p}_{p,q}) \\
&\leq \mathrm{rk}(\bar{\gamma}^{p,p+1}_{p,q}) + \mathrm{rk}(\bar{\alpha}^{0,p}_{p-1,q+1}) \qquad \text{for } 1 \leq p \leq m \qquad (77) \\
\mathrm{rk}\,(\bar{\alpha}^{0,1}_{0,q}) &\leq \mathrm{rk}(\bar{\gamma}^{0,1}_{0,q}) \qquad (78)
\end{aligned}$$

By induction we conclude that

$$\mathrm{rk}(\bar{\alpha}^{0,p+1}_{p,q}) \leq \sum_{k=0}^{p} \mathrm{rk}(\bar{\gamma}^{k,k+1}_{k,p+q-k}) = \sum_{k=0}^{p} \mathrm{rk}(\bar{\gamma}^{p-k,p+1-k}_{p-k,q+k}) \tag{79}$$

for $0 \leq p \leq m$. Setting q to 0, this inequality specializes — in the presence of formulae (70), (72) , and (73) — precisely to the claim in Theorem 4.1. $\square$.

Corollary 4.2 (Nested Coverings) *Let* $B^0_i \subset B^1_i \subset \ldots \subset B^{n+1}_i$ *,*$1 \leq i \leq N$*, be a family of nested open subsets in an* n*-dimensional topological manifold* M^n*. Then*

$$\begin{aligned}
&rk\Big(H_*(\bigcup_i B^0_i) \to H_*(\bigcup_i B^{n+1}_i) \Big) \\
&\leq \sum_{k=0}^{n} \sum_{i_0<\ldots<i_k} rk\Big(H_*(B^{n-k}_{i_0} \cap \ldots \cap B^{n-k}_{i_k}) \to H_*(B^{n+1-k}_{i_0} \cap \ldots \cap B^{n+1-k}_{i_k}) \Big) \qquad (80)
\end{aligned}$$

where $H_*(\ldots)$ *is again singular homology with coefficients in some arbitrary field* $\mathcal{F}$*.*

Remark: The number of terms on the r.h.s. of (80) is

$$\sum_{k=0}^{n} \binom{N}{k+1} \;<\; N \cdot \sum_{k=0}^{n} \frac{N^k}{(k+1)!} \;<\; N^{n+1} \tag{81}$$

Proof of the Corollary. Since we are dealing with open subsets in an n-dimensional manifold M^n, H_p vanishes unless $0 \leq p \leq n$. Therefore

$$\begin{aligned} &\operatorname{rk}\left(\mathrm{H}_*(\bigcup_i B_i^0) \to \mathrm{H}_*(\bigcup_i B_i^{n+1})\right) \\ &= \sum_{p=0}^{n} \operatorname{rk}\left(\mathrm{H}_p(\bigcup_i B_i^0) \to \mathrm{H}_p(\bigcup_i B_i^{n+1})\right) \\ &\leq \sum_{p=0}^{n} \operatorname{rk}\left(\mathrm{H}_p(\bigcup_i B_i^{n-p}) \to \mathrm{H}_p(\bigcup_i B_i^{n+1})\right) \end{aligned}$$

Each term on the r.h.s. can be estimated separately by applying Theorem 4.1 to the nested open sets $B_i^{n-p} \subset \ldots \subset B_i^{n+1}$, $1 \leq i \leq N$. With this shift in the indexing in mind [$m+1 = (n+1)-(n-p)$], one gets – slightly sharper than (80) – :

$$\begin{aligned} &\operatorname{rk}\left(\mathrm{H}_*(\bigcup_i B_i^0) \to \mathrm{H}_*(\bigcup_i B_i^{n+1})\right) \\ &\leq \sum_{k=0}^{n} \sum_{i_0<\ldots<i_k} \sum_{\mu=0}^{n-k} \operatorname{rk}\left(\mathrm{H}_\mu(B_{i_0}^{n-k} \cap \ldots \cap B_{i_k}^{n-k}) \to \mathrm{H}_\mu(B_{i_0}^{n+1-k} \cap \ldots \cap B_{i_k}^{n+1-k})\right) \end{aligned}$$

thus proving the Corollary. $\square$.

References

[A] Abresch, U. : Lower Curvature Bounds, Toponogov's Theorem, and Bounded Topology I/II. Ann. scient. Éc. Norm. Sup., 4^e série, t. 18, 1985, 651–670.

[AG] Abresch, U., Gromoll, D. : Open manifolds with nonnegative Ricci curvature.

[AW] Aloff, S., Wallach, N. R. : An infinite family of distinct 7-manifolds admitting positively curved Riemannian structures. Bull. Amer. Math. Soc. 81 (1975), 93-97.

[B] Bérard Bergery, L. : Quelques exemples de variétés riemanniennes complètes non compactes à courbure de Ricci positive. C. R. Acad. Sc. Paris, t. 302. Série I, n^o 4, (1986), 159-161.

[BT] Bott, R., Tu, L. W. : Differential Forms in Algebraic Topology. Graduate Texts in Mathematics 82, Springer 1982.

[CE] Cheeger, J., Ebin, D.G. : Comparison Theorems in Riemannian Geometry. American Elsevier, New York 1975.

[CG1] Cheeger, J., Gromoll, D. : On the structure of complete manifolds of nonnegative curvature. Ann. of Math. 96 (1972), 413-443.

[CG2] Cheeger, J., Gromoll, D. : The splitting theorem for manifolds of nonnegative Ricci curvature. J. Diff. Geom. 6 (1971), 119-128.

[Cg] Cheng, S. Y. : Eigenvalue comparison theorem and its geometric applications. Math. Z. 143 (1975), 289-297.

[E] Elerath, D.: An Improved Toponogov Comparison Theorem for Non-negativly Curved Manifolds. J. Diff. Geom. 15 (1980), 187-216.

[EH] Eschenburg, J., Heintze, E. : An elementary proof of the Cheeger-Gromoll splitting theorem. Ann. Glob. Analysis and Geometry 2 (1984), 141-151.

[ESS] Eschenburg, J., Schroeder, V., Strake, M. : Curvature at infinity of open nonnegatively curved manifolds. J. Diff. Geom. 30 (1989), 155-166.

[GG] Gromoll, D., Grove, K. : Rigidity of positively curved manifolds with large diameter. Seminar on Differential Geometry, Ann. of Math. Studies, Princeton University Press (1982), 203-207.

[GKM] Gromoll, D., Klingenberg, W., Meyer,W. : Riemannsche Geometrie im Grossen. Lecture Notes 55, Springer 1968.

[GM1] Gromoll, D., Meyer, W. : On complete manifolds of positive curvature. Ann. of Math. 90 (1969), 75-90.

[GM2] Gromoll, D., Meyer, W. : Examples of complete manifolds with positive Ricci curvature. J. Diff. Geom. 21 (1985), 195-211.

[G1] Gromov, M. : Almost flat manifolds. J. Diff. Geom. 13(2) (1978), 231-243.

[G2] Gromov, M. : Curvature, diameter and Betti numbers. Comment. Math. Helv. 56 (1981), 179-195.

[GS] Grove, K., Shiohama, A. : A generalized sphere theorem. Ann. of Math. 106 (1977), 201-211.

[GP] Grove, K., Petersen, P. : Bounding homotopy types by geometry. Ann. of Math. 128 (1988), 195-206.

[K] Karcher, H. : Riemannian Comparison Constructions. Preprint Bonn 1987

[SY1] Sha, Ji-Ping, Yang, Da-Gang : Examples of manifolds of positive Ricci curvature. J. Diff. Geo. (1989)

[SY2] Sha, Ji-Ping, Yang, Da-Gang : Positive Ricci Curvature on the Connected Sums of $S^n \times S^m$. Preprint.

[S] Sharafutdinov, V. A. : Convex sets in a manifold of negative curvature. Math. Zametki 26 (1979) 556-560.

[St] Strake, M. : A splitting theorem for open nonnegatively curved manifolds. Manuscripta Math. 61 (1988) 315-325.

[W] Wallach, N. L. : Compact homogeneous Riemannian manifolds with strictly positive curvature. Ann. of Math. 96 (1972) 277-295.

[Y] Yim, J. W. : Distance nondecreasing retraction on a complete open manifold of nonnegative curvature. Preprint, University of Pennsylvania, 1987.